Applied Microlocal Analysis of Deep Neural Networks for Inverse Problems

Dissertation
an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig-Maximilians-Universität München

eingereicht von

Héctor Andrade Loarca

22.11.2021

1. Gutachterin: Prof. Dr. Gitta Kutyniok

2. Gutachter: Prof. Dr. Ozan Öktem

3. Gutachterin: Prof. Dr. Carola-Bibiane Schönlieb

Berichterstatter: Prof. Dr. Markus Heydenreich

Tag der mündlichen Prüfung: 05.07.2022

Acknowledgements

The first person I would like to thank is my supervisor Prof. Gitta Kutyniok, for her constant support since the beginning of my career as a researcher, allowing me to freely choose my path but guiding it to the bleeding edge advances. I would also like to thank her for the unconditional help in the complex times we are living in now. I would like to thank Prof. Ozan Öktem for the great collaboration that took place along my Ph.D., he has been always ready to have lengthy scientific discussions, which have helped me to better understand the field. In addition, I would like to thank Philipp Petersen for also being part of this collaboration, allowing me to shape the formality of my work to the required level, and to always push to the highest standards in research. Without Gitta, Ozan, and Philipp, this thesis could not exist. In addition I thank Ozan, Prof. Carola-Bibiane Schönlieb and Prof. Lukas Gonon for accepting being part of my committee without any hesitation.

I would like to thank the BMS staff for the great help in bureaucratic and personal endeavors, in particular, to Annika Preuss for always being open to giving me advice on my constant issues with the complex german bureaucracy. In addition, I would like to thank the BMS to give me this great opportunity to make my master's and part of my Ph.D. in Berlin, as well as all the facilities provided which made my life easier. More recently, I have been very lucky with people that have helped me with the transition to Munich. I would like to thank in particular Tamara Friedriszik who has extensively helped me with all the paperwork for this transition, always with a great attitude. Besides, a special thank goes to Ron Levie, not just for being the great colleague he is but also for proofreading my thesis and helping me to formalize it. Finishing this process at the Ludwig-Maximilians-Universität in Munich is a great honor.

I also to thank all the special people that I've met along with my stay in Berlin. I would thank Julio and Alexis for the concurrent discussions on life, philosophy, and science, which have helped me to open my view of what scientific knowledge is about. I would like to thank Gary and Marco, for their friendship and for the great moments, which will continue to happen in the near and far future. I also want to thank Enrique and Carlos, since as they say, a 10+ years friendship will stay forever. I would like also to thank my great friends from the BMS, Brent, Josué, Adrián, Tatiana, and Qiao, for the great moment we shared in and out of the BMS Lounge. A special thanks go to Stephan, for being the great officemate he is, always open to talk about interesting and controversial topics.

As an important part of my career as a Ph.D. student and as a scientist, I have also explored the industrial applications of my field, and AngioWave Imaging has played the most important role in this endeavor. Therefore, I want to thank Aram and Bess for all the support through this journey. Of course, Bill is one of my great inspirations, always open to having extensive intellectual discussions, looking solely for insight, the kind of insight that fills your soul and reminds you why you are doing what you do. More recently, Dave has also been a great inspiration, his long experience on the field and deep inside has helped me to follow the correct path on problem-solving. I am looking forward to keeping working with these great scientists and collaborators, I am sure great things await us.

In addition, I would like to thank my parents, Hector and Julieta, for their great support since the beginning of my existence, they have always been there for me, teaching me how to be a good person and a hard worker, without them none of these could happen. I also want to thank, Cristina, Patricia, and Sara, to be always there, and be the best family I could ever ask for. Finally, I would like to thank specially to Natasha, for the great time we had and will have together, for her help and support, for the love and understanding, and for all the years and adventures that await us.

Affidavit

Hereby I declare that I wrote this thesis myself with the help of no more than the mentioned literature and auxiliary means.

München, 15.09.2021

Hector Andrade Loarca
. . . .
Signature

Abstract

Deep neural networks have recently shown state-of-the-art performance in different imaging tasks. As an example, EfficientNet is today the best image classifier on the ImageNet challenge [110]. They are also very powerful for image reconstruction, for example, deep learning currently yields the best methods for CT reconstruction [3, 17]. Most imaging problems, such as CT reconstruction, are ill-posed inverse problems, which hence require regularization techniques typically based on a-priori information. Also, due to the human visual system, singularities such as edge-like features are the governing structures of images. This leads to the question of how to incorporate such information into a solver of an inverse problem in imaging and how deep neural networks operate on singularities. The main research theme of this thesis is to introduce theoretically founded approaches to use deep neural networks in combination with model-based methods to solve inverse problems from imaging science. We do this by heavily exploring the singularity structure of images as a-priori information. We then develop a comprehensive analysis of how neural networks act on singularities using predominantly methods from the microlocal analysis.

For analyzing the interaction of deep neural networks with singularities, we introduce a novel technique to compute the propagation of wavefront sets through convolutional residual neural networks (conv-ResNet). This is achieved in a two-fold manner: We first study the continuous case where the neural network is defined in an infinite-dimensional continuous space. This problem is tackled by using the structure of these networks as a sequential application of continuous convolutional operators and ReLU non-linearities and applying microlocal analysis techniques to track the propagation of the wavefront set through the layers. This then leads to the so-called *microcanonical relation* that describes the propagation of the wavefront set under the action of such a neural network. Secondly, for studying real-world discrete problems, we digitize the necessary microlocal analysis methods via the digital shearlet transform. The key idea is the fact that the shearlet transform optimally represents Fourier integral operators hence such a discretization decays rapidly, allowing a finite approximation. Fourier integral operators play an important role in microlocal analysis, since it is well known that they preserve singularities on functions, and, in addition, they have a closed form microcanonical relation. Also, based on the newly developed theoretical analysis, we introduce a method that uses digital shearlet coefficients to compute the digital wavefront set of images by a convolutional neural network [7].

Our approach is then used for a similar analysis of the microlocal behavior of the learned-primal dual architecture [3], which is formed by a sequence of conv-ResNet blocks. This architecture has shown state-of-the-art performance in inverse problem regularization, in particular, computed tomography reconstruction related to the Radon

transform. Since the Radon operator is a Fourier integral operator, our microlocal techniques can be applied. Therefore, we can study with high precision the singularities propagation of this architecture.

Aiming to empirically analyze our theoretical approach, we focus on the reconstruction of X-ray tomographic data. We approach this problem by using a task-adapted reconstruction framework [1], in which we combine the task of reconstruction with the task of computing the wavefront set of the original image as a-priori information. Our numerical results show superior performance with respect to current state-of-the-art tomographic reconstruction methods; hence we anticipate our work to also be a significant contribution to the biomedical imaging community.

Zusammenfassung

Tiefe neuronale Netze haben in letzter Zeit bei verschiedenen Bildverarbeitungsaufgaben Spitzenleistungen gezeigt. Zum Beispiel ist AlexNet heute der beste Bildklassifikator bei der ImageNet-Challenge [71]. Sie sind auch sehr leistungsfähig für die Bildrekonstruktion, zum Beispiel liefert Deep Learning derzeit die besten Methoden für die CT-Rekonstruktion [3, 17]. Die meisten Bildgebungsprobleme wie die CT-Rekonstruktion sind schlecht gestellte inverse Probleme, die daher Regularisierungstechniken erfordern, die typischerweise auf vorherigen Informationen basieren. Auch aufgrund des menschlichen visuellen Systems sind Singularitäten wie kantenartige Merkmale die bestimmenden Strukturen von Bildern. Dies führt zu der Frage, wie man solche Informationen in einen Löser eines inversen Problems in der Bildverarbeitung einbeziehen kann und wie tiefe neuronale Netze mit Singularitäten arbeiten. Das Hauptforschungsthema dieser Arbeit ist die Einführung theoretisch fundierter konzeptioneller Ansätze zur Verwendung von tiefen neuronalen Netzen in Kombination mit modellbasierten Methoden zur Lösung inverser Probleme aus der Bildwissenschaft. Wir tun dies, indem wir die Singularitätsstruktur von Bildern als Vorinformation intensiv erforschen. Dazu entwickeln wir eine umfassende Analyse, wie neuronale Netze auf Singularitäten wirken, indem wir vorwiegend Methoden aus der mikrolokalen Analyse verwenden.

Um die Interaktion von tiefen neuronalen Netzen mit Singularitäten zu analysieren, führen wir eine neuartige Technik ein, um die Ausbreitung von Wellenfrontsätzen mit Hilfe von Convolutional Residual neuronalen Netzen (Conv-ResNet) zu berechnen. Dies wird auf zweierlei Weise erreicht: Zunächst untersuchen wir den kontinuierlichen Fall, bei dem das neuronale Netz in einem unendlich dimensionalen kontinuierlichen Raum definiert ist. Dieses Problem wird angegangen, indem wir die besondere Struktur dieser Netze als sequentielle Anwendung von kontinuierlichen Faltungsoperatoren und ReLU-Nichtlinearitäten nutzen und mikrolokale Analyseverfahren anwenden, um die Ausbreitung einer Wellenfrontmenge durch die Schichten zu verfolgen. Dies führt dann zu einer mikrokanonischen Beziehung, die die Ausbreitung der Wellenfrontmenge unter ihrer Wirkung beschreibt. Zweitens digitalisieren wir die notwendigen mikrolokalen Analysemethoden über die digitale Shearlet-Transformation, wobei die Digitalisierung für die Untersuchung realer Probleme notwendig ist. Die Schlüsselidee ist die Tatsache, dass die Shearlet-Transformation Fourier-Integraloperatoren optimal repräsentiert, so dass eine solche Diskretisierung schnell abklingt und eine endliche Approximation ermöglicht. Nebenbei stellen wir auch eine Methode vor, die digitale Shearlet-Koeffizienten verwendet, um den digitalen Wellenfrontsatz von Bildern durch ein Faltungsneuronales Netzwerk [7] zu berechnen.

Unser Ansatz wird dann für eine ähnliche Analyse für die gelernte primale-duale Architektur [3] verwendet, die durch eine Sequenz von conv-ResNet-Blöcken gebildet wird. Diese Architektur hat bei der Rekonstruktion inverser Probleme, insbesondere bei der Rekonstruktion der Computertomographie im Zusammenhang mit der

Radon-Transformation, Spitzenleistungen gezeigt. Da der Radon-Operator ein Fourier-Integraloperator ist, können unsere mikrolokalen Techniken angewendet werden.

Um unseren theoretischen Ansatz numerisch zu analysieren, konzentrieren wir uns auf die Rekonstruktion von Röntgentomographiedaten. Wir nähern uns diesem Problem mit Hilfe eines aufgabenangepassten Rekonstruktionsrahmens [1], in dem wir die Aufgabe der Rekonstruktion mit der Aufgabe der Berechnung der Wellenfrontmenge des Originalbildes als Vorinformation kombinieren. Unsere numerischen Ergebnisse zeigen eine überragende Leistung, daher erwarten wir, dass dies auch ein interessanter Beitrag für die biomedizinische Bildgebung sein wird.

Contents

List of Figures

List of Tables

1 Introduction

In many scientific and industrial real-world applications, some level of understanding of how model parameters (physical variables) are transformed under measurements is normally required. The model parameters are typically represented by a function in a Hilbert space, and the measurements by an operator between spaces. Since a significant portion of semantic information of the function is contained in the singular (non-smooth) part of the function, the study of how singularities are transformed under the action of operators becomes fundamental. In the case that such function represents an image, the singular part corresponds to edges, ridges, or ramps in the image.

Microlocal analysis is a mathematical theory that aims to precisely describe how the singular part of a function, or more generally a distribution, is transformed when acted upon by an operator. It was introduced in the early 1970s by Sato [103] and Hörmander [61], originally intended to study the propagation of singularities in partial differential equations, and it has been widely used in both pure and applied mathematical research since then. The main premise in microlocal analysis is that the information about the location of the singularities, also known as the singular support, needs to be complemented with specifying the directional information of such singularities, referred as "microlocal" information. The set that contains the location and directional information of the singularities is named the wavefront set. The directional or microlocal information plays a key role in elucidating the propagation of singularities by a certain class of operators, the Fourier integral operators (FIO). Such operators are frequently encountered in analysis, scientific computing, and physical sciences [61, 20]. One example is the computed tomography problem, modeled by the Radon transform operator, and it will play an important role in this thesis.

Microlocal analysis is particularly useful in inverse problems, where the goal is to reliably recover hidden model parameters (signal) from a noisy transformed version (data), both connected by a mapping known as forward operator. Moreover, we would also like to recover the wavefront set of the signal given the noisy realization of a transformed version of the signal. Such applications frequently arise when using imaging/sensing technologies where the transform is a pseudodifferential or Fourier integral operator [70].

In this thesis, we study the extraction of oriented singularities of *digital two-dimensional images* and their behavior under the action of a deep neural network feed-forward operator, as well as other forward operators related to biomedical imaging problems. We will focus on the use of microlocal analysis tools in the continuum setting and how this can be faithfully discretized, and furthermore, digitized. This work is mainly driven by the extension of its applications to inverse problems coming from biomedical imaging, but it also gives an important insight into how neural networks approximate singularities. In addition, when studying neural network architectures used for inverse problems (see

Chapter 7), it is important to understand how the singularities of the data are propagated to the output. In that sense, one is able to use this prescribed propagation to asses the reconstruction.

1.1 Microlocal analysis of inverse problems in imaging

In this thesis, we will focus on the utilization of microlocal analysis in the context of imaging sciences. Here, an image is represented by a real-valued function in two dimensions describing the interior structure of the object under study. In this context, an inverse problem aims to recover the image from noisy data, which is often not possible, either because the transformation relating the image to data is not invertible or because the data is incomplete. Next, the reconstruction of the image is just a part of the entire workflow involved in real-world inverse problems, which ultimately aims at decision-making.

A prime example is tomographic reconstruction, where a medical expert uses the reconstructed image to decide whether a patient needs certain intervention, for example, due to a tumor. In such a case, the location and shape of the tumor is often sufficient for the decision-making, whereas the exact values of the tumor density may be ignored or identified with another more specialized technique. The location and shape of the tumor can be determined from the singular part of the image, so the estimation of the wavefront set will be sufficient.

In the prime example above, the Radon transform, being the related forward operator, is a Fourier integral operator [70]. Hence, tools provided by microlocal analysis allow us to explicitly describe the relationship between the wavefront set of the functions (image) and its transformed version (tomographic data). This relation is referred to as *microcanonical relation* [61] and can also be used to identify which singularities can be recovered from data without explicitly computing the inverse Radon transform [89]. Another observation is that recovering an image from its Radon transform is less ill-posed if one knows the wavefront set a-priori [38] since as we discussed, it contains an important amount of semantic information.

As mentioned before, many real-world imaging applications, which can be seen as inverse problems, have a Fourier integral operator as their data model. This allows to use the same microlocal analysis tools as in the computed tomography application. In this thesis, we will extend this notion to neural networks, with the final aim to describe the propagation of singularities by their different layers. This will also require an extension of the microlocal analysis tools to non-smooth/non-linear operators. We are mainly motivated by the current trend of using deep neural networks to solve inverse problems. In this context, being able to describe the propagation of singularities the networks will allow us to analyze how well the networks approximate the singularities of the reconstruction. In addition, we will be able to use such singularities as a strong prior in the context of task-adapted reconstruction (see Chapter 7).

1.2 Basic notions of deep learning

In the last decade, machine learning has played an important role in imaging applications, mainly due to the exponential increase in computing power given by Moore's law and the increase in available data. Along with the distinct methods in machine learning, deep learning has been the state-of-the-art approach in most of the imaging applications, an example is the ImageNet classification challenge, being best performed by the EfficientNet architecture [110]. In inverse problems in imaging, most of the current best reconstructions are also done by deep neural networks [17, 3]. This is the main reason they will play a central role in this thesis.

Deep learning is based on the efficient, data-driven, training of neural networks. In broad terms, a neural network is a sequence of simple operations, known as *neurons*. These neurons are arranged in complex patterns but ordered in sequential evaluation, also-called *layers*. Their name, neural network, comes from the original motivation behind their introduction, based on how biological neural networks work. Historically, the very first work on neural networks done by McCulloch et al. [85] had the goal to mathematically model the human brain. The main feature transferred to neural networks from their biological counterparts is the idea that a neuron sends out a signal when a specific threshold of inputs is exceeded. For practical purposes, we will center on the so-called feed-forward neural networks.

In mathematical terms, a feed-forward neural network is defined by the parameters:

- $d \in \mathbb{N}$-input dimension.
- $L \in \mathbb{N}$-number of layers.
- $N_0, \dots, N_L \in \mathbb{N}$-number of neurons in each layer, where $N_0 = d$.
- $A_l \in \mathbb{R}^{N_\ell \times N_{\ell-1}}$, $b_l \in \mathbb{R}^{N_\ell}$- *weights* of the neural network, where $\ell = \{1, \dots, L\}$.
- $\rho : \mathbb{R} \to \mathbb{R}$-*activation function.*

Then, the neural network is given by the function $\Phi : \mathbb{R}^d \to \mathbb{R}^{N_L}$:

$$\Phi(x) = W_L(\rho(W_{L-1}(\rho(\dots, \rho(W_1(x)))))),$$

where $W_\ell(x) = A_\ell(x) + b_\ell$ is an affine transform in each layer, for $\ell = 1, \dots, L$, and ρ is applied component wise. In this context, *depth* means the number of layers, and a *deep network* means a network with many layers. We refer to a feed-forward neural network as *deep neural network* when $L \geq 2$. We also refer to Φ as a *neural network architecture*. One can also represents a neural network as the diagram depicted in Figure 1.1.

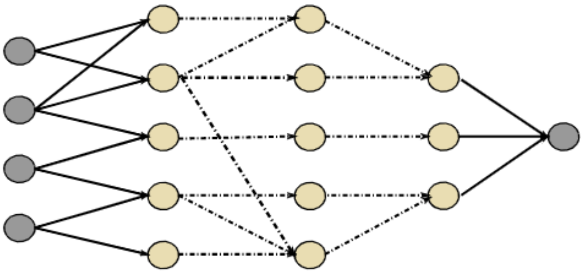

Figure 1.1: Feed-forward neural network.

In this work we will focus on the case when the matrix A_ℓ is convolutional. We first consider the case of CNNs for signals on 1D domains. Let $m_1 \in \mathbb{N}$ be the *number of feature channels*, $n_0 \in \mathbb{N}$ be the filter size, and $\{f_k\}_{k=1}^{m_1}$ be a collection of filters. Here, $f_k \in \mathbb{R}^{n_0}$. In addition, the matrix $A_\ell \in \mathbb{R}^{N_\ell \times m_1 N_\ell}$ is a convolutional matrix with the corresponding filters if $A_\ell = \{a_{ij}\}_{i,j}$ where

$$\begin{aligned}
\{\{a_{ij}\}_{i=1}^{n_0}\}_{j=1}^{m_1} &= \{f_k\}_{k=1}^{m_1}, \\
\{\{a_{ij}\}_{i=2}^{n_0+1}\}_{j=m_1+1}^{2m_1} &= \{f_k\}_{k=1}^{m_1}, \\
&\cdots \\
\{\{a_{ij}\}_{i=N_\ell}^{n_0+N_\ell}\}_{j=(N_\ell-1)m_1+1}^{N_\ell m_1} &= \{f_k\}_{k=1}^{m_1},
\end{aligned}$$

where $a_{N_\ell+t,j} := a_{tj}$ for all $j \in \{1,\dots,N_\ell m_1\}$. We can visualize the convolutional matrix A_ℓ, for all $\ell = 1,\dots,L$ by Figure 1.2.

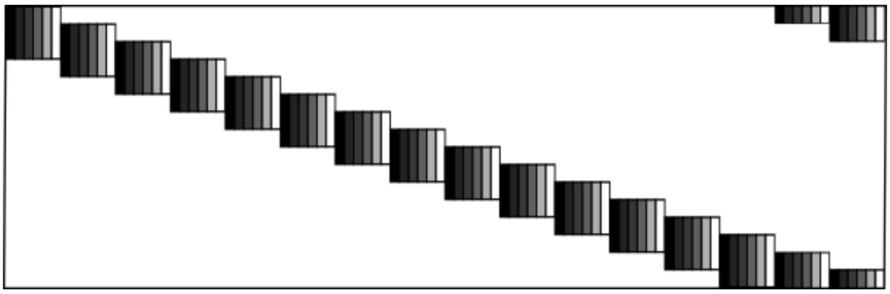

Figure 1.2: Convolutional matrix.

The neural networks generated by convolutional matrices are known as convolutional neural networks. This is due to the matrix multiplication involved that can be written as a convolution:

$$A_\ell x[i] = \sum_{s=1}^{n_0} \sum_{k=1}^{m_1} f_k[s]x[i-s] \quad \text{for } i = 1,\dots,n_0 \text{ and } \ell = 1,\dots,L.$$

Similarly, one can extend this notion to 2D (images) by simply having 2D filters $f_k \in \mathbb{R}^{n_0 \times n_1}$, obtaining the convolution formula

$$A_\ell x[i,j] = \sum_{s=1}^{n_0} \sum_{q=1}^{n_1} \sum_{k=1}^{m_1} f_k[s,q] x[i-s, j-q] \quad \text{for } i = 1, \dots n_0,\ j = 1, \dots, m_0,\ \ell = 1, \dots, L.$$

The biggest advantage of using convolutional neural networks in comparison with the classical fully connected neural networks is the efficiency of their training. This is mainly due the sparsity of their matrix representation. They also provide highly adaptive feature extraction which has been used widely in tasks as image classification [71]. Feed-forward neural networks are the backbone of modern machine learning algorithms, due to their efficiency and adaptability. By now, we have introduced the model of deep neural networks, but we have not yet discussed how to train them.

Training a neural network is the act of finding the optimal weights, i.e., $(A_\ell, b_\ell)_{\ell=1,\dots,L}$, in a way as to minimize a *loss function* on a *training set*. Let us assume that we have a number of input images with known desirable target outputs, namely, a set of input-output pairs. These pairs are elements of the training set, $(x_i, y_i)_{i=1,\dots,m} \subset \mathbb{R}^{d \times d'}$. In addition, let $\mathcal{L} : \mathbb{R}^{d' \times d'} \to \mathbb{R}_+$ be a mapping, known as *loss function*. One trains neural networks by solving the optimization problem

$$\min_{\Phi} \sum_{i=1}^{m} \mathcal{L}(\Phi(x_i), y_i) \tag{1.2.1}$$

over all neural networks Φ with a $d-$dimensional input and N_L-dimensional output under the restriction that the *architecture* of the network, namely, the parameters d, L, and $N_0, \dots, N_L$ are fixed.

In practice, the minimization problem is solved by *stochastic gradient descent*. Let $W = (\omega_k)_{k=1}^{M'} \in \mathbb{R}^{M'}$, with $M' = \sum_{\ell=1}^{L} N_\ell N_{\ell-1}$, be the weights of a neural network, e.g., $(A_\ell)_{ij=1,1}^{N_\ell, N_{\ell-1}}$ and $(b_i)_{i=1}^{N_\ell}$, for $\ell = 1, \dots, L$ for the convolutional case. a network can also be interpreted as the mapping

$$(x, \omega_1, \dots, \omega_{M'}) \mapsto \Phi_W(x), x \in \mathbb{R}^d.$$

In this interpretation, given an estimate of the optimal weights (minimize (1.2.1)) $W^t \in \mathbb{R}^{M'}$ at step t of the stochastic gradient descent algorithm, we can compute the derivative $\partial \sum_{i=1}^{M'} \mathcal{L}(\Phi_W^t(x_i), y_i)/\partial \omega_k^t$ for all $k = 1, \dots, M'$ and replace ω_k^t by

$$\omega_k^{t+1} = \omega_k^t - \lambda \partial \sum_{i=1}^{M'} \mathcal{L}(\Phi_W^t(x_i), y_i)/\partial \omega_k^t,$$

where $\lambda > 0$ is the step-size, also known as *learning rate*. Although it is clearly hard to compute all the derivatives for all weights, there is a convenient algorithm, called *backpropagation* [100], allowing very efficient computation. In short, backpropagation is an iterative application of the chain rule. We can also refer to this training process as *data-fitting*.

The elements presented on the previous pages are the basic ingredients of deep neural networks, which study is called "deep learning", in the next section we will explore a specific application of deep neural networks to solve inverse problems. Later, in Chapters 4 and 6, we will introduce the theory behind the propagation of singularities by deep neural networks, which will be used in Chapter 7 to design a tomographic reconstruction algorithm.

1.3 Deep neural network architectures for inverse problems

Before we can dive into the details of our approach of applied microlocal analysis in inverse problems, it is necessary to introduce formally, what we refer to as an inverse problem. In mathematics, an inverse problem is the task of reconstructing (estimating) a signal $f_{\text{true}} \in X$ from data $g \in Y$, having the relation

$$g = \mathcal{A}(f_{\text{true}}) + \delta g, \tag{1.3.1}$$

where X (model parameter space) and Y (data space) are inner-product spaces, and $\mathcal{A} : X \longrightarrow Y$ is the forward operator that models how the data is produced from the signal in the absence of noise. Finally, $\delta g \in Y$ is the noise, defined as a single sample of a Y-valued random variable that represents the noise component of the data. In the case that the forward operator $\mathcal{A}$ is invertible, the reconstruction of f_{true} will just require a denoising step, obtaining:

$$f_{\text{true}} = \mathcal{A}^{-1}(g^*),$$

where g^* is a denoised version of the data g. In general, the operator $\mathcal{A}$ will be more complicated, and non-invertible. In such cases, we rely on the notion of well-posedness and ill-posedness, initially introduced by Hadamard [53].

An inverse problem, defined by (1.3.1) is well-posed in terms of Hadamard, if:

1. a solution exists,
2. the solution is unique,
3. the solution depends continuously on the data.

If an inverse problem fails to hold any of these conditions, it is said to be ill-posed. Most of the interesting inverse problems in the real-world, including computed tomography, are indeed ill-posed [87].

Classically, researchers have applied optimization approaches to solve inverse problems. In the simplest situation, one can find an approximate solution of the problem by minimizing the data miss-fit with a loss function:

$$f_{\text{true}} = \operatorname*{arg\,min}_{f \in X} \mathcal{L}(\mathcal{A}(f), g). \tag{1.3.2}$$

The loss function $\mathcal{L} : Y \times Y \longrightarrow \mathbb{R}_+$ measures how well the measurements of the signal f approximate the data $g = \mathcal{A}(f)$. The loss function is typically chosen to be proportional to the negative log-likelihood [12], for example, the squared loss. In the case of ill-posed inverse problems this approach leads to overfitting, i.e., the obtained model parameters correspond too closely or exactly to a particular set of data and may therefore fail to fit additional data or predict future observations reliably.

Solutions that are unstable against data, i.e., small changes in the data lead to big changes in the solution, are known as *non-stable* or *irregular* solutions. *Regularization theory* is the area in mathematics that studies ways to find *regular* or *stable* solutions. Researchers have proposed different approaches on regularizing an inverse problem, mainly based on the introduction of a-priori information about the physics of the problem.

1.3.0.1 Model-based regularization

Classically, the most common way to introduce a-priori information to the problem is by the so-called regularization functional. This approach is known as *variational regularization.* Let $S : X \longrightarrow \mathbb{R}$ be a functional that encodes a-priori information about f_{true} and penalizes unlikely solutions, also referred to as *regularization functionals.*

In this approach, one solves a problem alternative to (1.3.2), given by

$$f_{\text{true}} = \operatorname*{arg\,min}_{f \in X} \left[\mathcal{L}(\mathcal{A}(f), g) + \lambda \mathcal{S}(f) \right], \tag{1.3.3}$$

where $\lambda \geq 0$ is a fixed *regularization parameter* controlling the influence of the a-priori knowledge provided by the regularization functional against the data miss-fit term. In the language of Bayesian estimation, a regularization functional represents a prior for the statistical estimation problem.

As an example, to show the main idea behind variational regularization. If $X = L^2(\mathbb{R}^2) \cap L^1(\mathbb{R}^2)$ and $Y = L^2(\mathbb{R}^2)$, we can define the data miss-fit as the L^2-loss

$$\mathcal{L}(\mathcal{A}(f), g) = ||\mathcal{A}(f) - g||_2^2.$$

By using as prior information the assumption that the solution f_{true} is L^1-sparse, we get the optimization problem

$$f_{\text{true}} = \operatorname*{arg\,min}_{f \in L^2(\mathbb{R}^2) \cap L^1(\mathbb{R}^2)} ||\mathcal{A}(f) - g||_2^2 + \lambda ||f||_1.$$

This kind of approach, also known as sparse regularization and is used in diverse problems in imaging, such as, denoising and inpainting [45]. Another regularization technique widely used is the so-called *Tikhonov regularization* [112], where we assume that f_{true} has a small norm, i.e.

$$f_{\text{true}} = \operatorname*{arg\,min}_{f \in L^2(\mathbb{R}^2) \cap L^1(\mathbb{R}^2)} ||\mathcal{A}(f) - g||_2^2 + \lambda ||f||_2.$$

Since one uses the L^2-norm to penalize the norm of the function, the solutions of Tikhonov regularization tend to be smooth, since L^2-norm serves as an averaging term.

In order to avoid this drawback, more recently, the well-known *total variation regularization* or TV regularization was introduced. In this case, one assumes L^1-sparsity in the gradient of the solution [98], which leads to the function $S(f) = ||\nabla f||_1$. This regularization is commonly used in denoising and computed tomography reconstruction [117]. The use of the variational regularization approach requires the actual understanding of the physics of the problem in order to define a reasonable regularization functional. Hence one assumes a-priori knowledge of the problem, and empirical data is used just to calibrate the model parameters, for example, the regularization parameter λ.

Besides variational regularization methods, there are other prime examples classically used for regularization. One example is the analytic pseudo-inverse approach (e.g. FBP) [104], where one aims to find an approximate inverse $\mathcal{A}^\dagger : Y \longrightarrow X$, such that $\mathcal{A}^\dagger(g) \approx f_{\text{true}}$ whenever $\mathcal{A}(f_{\text{true}}) = g$. One last example are the so-called iterative methods with early stopping, where one aims to solve the linear programming problem 1.3.1 with an iterative method [13], avoiding over-fitting by early stopping. We will refer to the set of approaches based on first principles, as *model-based regularization.* On the one hand, since they are based on first principles, they can be tested and validated independently, and such simple concepts aid the understanding of the results. On the the other hand, having also few parameters to calibrate, these methods will require not much data. Moreover, they require the explicit description of causal relations between data and model-parameters, which is not always possible, and it is also hard to account for uncertainty, the latter due to the lack of knowledge on the data distributions.

1.3.0.2 Data-driven regularization

An alternative conceptual approach that has been recently widely used, is the so-called *data-driven regularization.* This approach available uses real data, to learn most of the parameters of a general input-output model. In the last few years, machine learning has played an important role in mathematical modeling, in particular, its sub-field deep learning, mostly due to the increasing parallel computing power that graphical processing units have provided. As discussed in Section 1.2 deep learning is a powerful tool for non-linear function approximation. One advantage of deep learning is the requirement of weak assumptions on the input-output model, being deep neural networks, and the available training data sets.

Data-driven regularization uses a general parametric differential model, called neural network architecture, and an optimization algorithm, stochastic gradient descent, to learn the parameters from available training data. It is also widely used in imaging science and inverse problems, in particular, in medical image reconstruction [118], becoming an important tool in computer vision, as well as other areas, nowadays. As we mentioned, the main disadvantage of the model-based approach to regularization is the need of a deep understanding of the problem, for example, in the case of computed tomography, one needs to know the geometry of the sensors, and the exact physics of the phenomena which includes the reflectance of the materials involved. Also, deep learning models can be used in a wide variety of problems without much understanding of the problem. Due to the highly non-linear structure, a deep neural network when correctly trained, can capture causal relation without making any limiting assumptions on the problem [28]. In this thesis we will focus on the data-driven regularization methods involving deep neural networks. It is important to know that these are not the only data-driven regularization methods available, since the area of machine learning has plenty of other methods. We focused on these methods, since they currently represent the state-of-the-art in inverse problems.

Let us assume again that we want to solve problem (1.3.1). The main idea of the data-driven regularization approach is to define a collection of non-linear parametric mappings $\{\mathcal{A}_\theta^\dagger\}_{\theta\in\Theta}$. Each of these mappings is parametrized by a deep neural network, with weights $\theta \in \Theta$, elements of a parameter space Θ. In order to be applied to the inverse problem (1.3.1), one requires the *pseudo-inverse* property for each $\mathcal{A}_\theta^\dagger$, meaning that:

$$\mathcal{A}_\theta^\dagger(g) \approx f_{\text{true}} \text{ , whenever } \mathcal{A}(f_{\text{true}}) = g.$$

One learns an approximation solution $\mathcal{A}_{\theta^*}^\dagger(g)$, by finding the parameter θ^* which minimizes an appropriate loss functional $L : \Theta \to \mathbb{R}_+$ that quantifies the dissimilarity between $\mathcal{A}_\theta^\dagger(g)$ and f_{true}. This minimization is done over some available data $\{(f_i, g_i)\}_{i=1}^N \subset X \times Y$, where $\mathcal{A}(f_i) = g_i$ for all i. As in Section 1.2 the process of finding the optimal θ^* is known as *training* and the used data is known as *training data.* The choice of the loss function L plays an important role in the success of the training; a prime example of the loss is the empirical loss given by the mean squared distance:

$$L(\theta) = \frac{1}{N}\sum_{i=1}^N ||\mathcal{A}^\dagger(g_i) - f_i||_X^2.$$

This loss can be optimized using gradient descent algorithms, although typically the data set contains too many data points, which makes the computation of the total loss expensive. In deep learning, the standard method to solve this issue is stochastic gradient descent (see Section 1.2), which optimizes the loss over a randomly selected subset of the training data. If we rely completely on a fully data-driven approach, meaning that we learn from data the entire reconstruction operator $\mathcal{A}_{\theta^*}^\dagger$, we will encounter some limitations: it can be computationally exhaustive, having too many parameters to fit. Data-driven approaches have no standard way to incorporate a-prior knowledge and therefore will not provide any conceptual simplification. This leads to almost no understanding of the resulting reconstruction.

Both the knowledge-driven and data-driven approaches have their own benefits and shortcomings but at the same time, they are compatible. As part of the recent efforts to introduce theoretical insight into deep learning methods, researchers have proposed to combine both methods in order to make them benefit from each other's potentials. We will refer to the combined approach as hybrid methods.

1.3.0.3 Hybrid methods

As part of the general data-driven regularization, we aim to find a (non-linear) reconstruction operator $\mathcal{A}^{\dagger}_{\theta^*}$ from finite training data $\{(f_i, g_i)\}_{i=1}^{N}$. In real-world applications, such as computed tomography, many required parameters to describe the complex phenomena from scratch will make the problem computationally intractable. It will therefore need an incredible amount of data in order to fit those parameters. Also, we do understand up to some level the physics of most of such problems, mainly due to the fact that those experiments were designed with such physics in mind. Therefore, one would not like to learn the entire reconstruction operator from scratch but instead to incorporate the a-priori information of the problem to reduce the parameter space. This type of approaches is known as *hybrid methods* and will be the main approach used in this thesis.

There are different ways to combine the data-driven and model-based methods in order to regularize an inverse problem, but they can be summarized in the following three approaches:

- **Learned post-processing:** in this approach one assumes that there is a known approximate pseudo-inverse $\mathcal{A}^{\dagger} : Y \to X$, although such pseudo-inverse might produce noisy solutions. One uses a deep neural network to learn a denoiser for the reconstruction $\mathcal{A}^{\dagger}(g)$. Formally, we obtain a reconstruction operator of the form:

  $$\mathcal{A}^{\dagger}_{\theta} = \Lambda_{\theta} \circ \mathcal{A}^{\dagger},$$

  where $\Lambda_{\theta} : X \longrightarrow X$ is a learned denoiser. There are some examples in the literature of this approach, we refer to [26, 65, 67].

- **Learned regularized:** this approach is based on the use of a template of the regularization functional and learns it from data. For example, in the case of sparse regularization, one can learn a dictionary that sparsifies the solution. This is known as dictionary learning [119].

- **Learned iterative schemes:** these methods make use of classical iterative methods for optimization to solve (1.3.1), but they learn the best update in each iteration using a-priori information. Learned iterative schemes can learn different parameters of the iterations, for example, the step size in a gradient descent method. In this thesis we will explore mainly the so-called *learned primal-dual algorithm* which uses a primal-dual optimization scheme and learns the proximal operators in each iteration [3]. Due to the important role that learned iterative schemes play in this thesis, we will explore them with more detail in Section 4.3.

1.4 Task-adapted reconstruction

In real-world applications of inverse problems, one uses measured data in order to make a decision, where the reconstruction of the model parameters is a key part but not the final goal. In some sense, one recovers features from data in order to study them and make a correct decision. For example, in computed tomography, one needs to reconstruct an image from data in order to detect physical anomalies in a patient, e.g., a tumor. In that sense, the reconstruction is required due to the difficulty to understand the raw data.

We will refer to the process of making correct decisions from features, as a *task*. A typical pipeline in an inverse problem will start with the sampling of the data, a middle pre-processing step to normalize the data, a reconstruction step to extract the features, and the final task based on the extracted features. A model that delivers correct decisions over raw data is known as a *task-adapted model*, and the problem of finding such a model is known as *task-adapted reconstruction*. This approach was first introduced by Adler et al. in [1], in this thesis we will follow their approach. One can formulate mathematically the task-adapted reconstruction problem as the recovery of a true unknown decision $d^* \in D$ from data $g \in Y$ following the relation

$$g = \mathcal{A}(f^*) + \delta g \text{ and } d^* = \mathcal{T}(f^*), \tag{1.4.1}$$

where X (reconstruction space), Y (data space) and D (decision space) are Banach spaces, $\mathcal{A} : X \longrightarrow Y$ is the forward operator, and $\mathcal{T} : X \longrightarrow D$ is the task operator. Finally, δg is noise in the data space.

The final aim of task-adapted reconstruction is to find a task-adapted operator $\mathcal{B} : Y \longrightarrow D$ such that $\mathcal{B}(g) = d^*$, whenever g and d^* are related via (1.4.1). Notice that the forward operator $\mathcal{A}$ is often highly non-injective. This implies that the trivial task-adapted operator $\mathcal{T} \circ \mathcal{A}^{-1}$ is not well defined. In general, one will not be able to perform the reconstruction and the task independently, since the task itself will highly depend on the distribution of the reconstructed features. In order to find a proper task-adapted operator, we must define the two problems involved, reconstruction and task, in the same mathematical framework.

As discussed, one can see the task as a decision-making on data from features (model parameters). This type of problems forms part of the *statistical decision theory* [78]. In the abstract setting, a task is a non-randomized decision rule which minimizes the Bayes risk associated with the decision space [1, Section 6.1]. The Bayes risk measures in some sense the distance to the ground truth decision in the decision space. It is also possible to formulate the reconstruction as a non-randomized decision rule, where the decision is to make the correct reconstruction. In Chapter 7 we will explore in-depth the abstract setting of both the task and the reconstruction problem, using the framework of statistical decision theory, as well as its digital counterpart. One important requirement that we will ask from both the task and the reconstruction is to be parametrized by a deep neural network. This allows us to jointly train them using a convex combination of the corresponding loss functions.

In this thesis we will focus on the task of wavefront set extraction. Under the framework of task-adapted reconstruction, it will be possible to perform jointly a wavefront set extraction and tomographic reconstruction. Although this approach might be interesting, since it will be able to recover the fully sampled wavefront set for the low-dose tomographic data, it will not be much different than simply performing edge detection as the task.

In order to fully use the potential of the wavefront set, we must make use of the microcanonical relation. As mentioned in Section 1.1, having in addition to the position of the singularities in an image, their orientations, allows us to study the propagation of such singularities under the action of Fourier integral operators, such as the Radon transform. In this setting one can use the microcanonical relation to map the wavefront set of the low-dose data to a low-dose wavefront set of the reconstruction. In our approach the task consists in recovering the full-dose wavefront set of the reconstruction from the low-dose counterpart. By jointly training this task and the reconstruction, we can make the wavefront set of the reconstruction approximate the wavefront set of the ground truth. We will explore this scenario with high detail in Chapter 7.

Also, in Chapter 7 we will show how to perform a task-adapted tomographic reconstruction method, that jointly recovers images from low-dose tomographic data and performs wavefront set inpainting using the known wavefront set. This task-adapted method will be able not just to recover the full wavefront set of an image from its low-dose sinogram, but at the same time improve the resulted reconstruction by forcing its wavefront set to be close to the ground truth.

1.5 Microlocal analysis of digital data

All the theory that has been mentioned so far is formulated in a continuum setting. More importantly, formally, singularities can just be defined in that setting. If we want to use these tools in real-world applications, we need to formulate them in a discrete domain, and at the same time make the formulation consistent with the continuous version. Chapter 5 will contain our approach to define the digital wavefront set of a digital two-dimensional signal originally introduced in [8]. In this chapter, we will also present a data-driven approach to compute the wavefront set of digital images and its comparison with similar methods.

Our approach makes use of computational harmonic analysis, in particular, multiscale directional systems that have well-known properties of resolution of oriented singularities, to represent the data in a convenient form. We also express the problem of wavefront set extraction as a problem of semantic segmentation, where we aim to find singularity points and classify them. In this case, the class of each singularity point will correspond to its wavefront set orientation. At the same time, in order to make use of the task-adapted reconstruction, we study the propagation of the digital wavefront set under the action of a certain class of deep neural networks, the *convolutional residual neural networks* [56]. In this case, we are using a similar approach as in the case of the wavefront set extraction, where we work on the theory for the continuous domain setting, presented in Chapter 4, in order to obtain a faithful discrete version of the results.

The analysis of singularity propagation under deep neural networks is useful for task-adapted reconstruction models that involve the wavefront set, it is a powerful tool on its own. Such analysis allows us to study theoretically how the neural networks transform singularities of signals, which contain already a lot of information about the signal itself. Chapter 6 explores how this analysis is done in the discrete and digital case.

1.6 Organization of the thesis

This thesis is structured into three parts. The first part contained in Chapters 2, 3 and 4 discusses the continuum setting of wavefront set extraction and microlocal analysis of Fourier integral operators and deep neural networks. The second part formed by Chapters 5 and 6 introduces the digital counterpart. Finally, the third part consisting of Chapters 7, 8 and 9 presents real-world applications and numerical experiments, as well as a conclusion to the thesis.

More specifically, Chapter 2 introduces the basic notions of microlocal analysis. Sections 2.1 and 2.2 present the notion of distributions and the definition of wavefront sets as well as some examples. Sections 2.3 to 2.4 discuss the propagation of wavefront sets by pseudodifferential and Fourier integral operators, while Section 2.5 presents the results for the Radon transform. Chapter 3 introduces the general resolution of wavefront sets by harmonic analysis techniques, all in the continuum setting. In particular, Sections 3.3 and 3.4 present the continuous shearlet transform, as well as the wavefront set resolution provided by it. Later, Chapter 4 introduces a novel approach for analyzing the propagation of singularities by residual convolutional neural networks. This is done in a non-standard way since we need to express neural networks as operators acting on continuous spaces. In Section 4.3 we use the above-mentioned theory in a particular architecture, the learned primal-dual architecture. This architecture will play a central role in this thesis since it will be used for tomographic reconstruction.

In order to digitize the theory presented in the previous chapters, Chapter 5 presents a deep learning approach for wavefront set extraction. This approach is based on the digital shearlet transform and convolutional neural networks. Sections 5.2 and 5.4 present a novel architecture that extracts wavefront sets of digital images by classifying local patches of their shearlet coefficients, where each class represents a particular direction on the wavefront set. Section 5.3 extends this notion to general semantic edge detection, where we show that shearlets serve as a good feature extractor for edge detection and classification. Chapter 6 introduces the notion of digital microlocal analysis of convolutional neural networks. In Section 6.1 we introduce a discretization technique for Fourier integral operators based on discrete shearlets. This technique profits from the fact that shearlets sparsely represent such operators, which leads to an efficient discretization. Section 6.2 analyzes the approximation rates of the above discretization and their ability to faithfully digitize the Fourier integral operators. This digitization allows us to have a digital microcanonical relation, a mapping that describes the propagation of wavefront sets in the digital realm. In addition, Section 6.3 applies these principles to residual convolutional neural networks and the digital Radon transform. This allows us to describe the microlocal behavior of the learned primal-dual architecture.

In Chapter 7 we present the final contributions of the thesis, an application for task-adapted tomographic reconstruction. This application uses the wavefront set of tomographic data to improve state-of-the-art reconstructions algorithms, by incorporating it in the training procedure. Sections 7.2, 7.3 and 7.3 present the basic notions of statistical decision theory as a framework that merges the reconstruction and task.

Section 7.5.1 presents the concept of task-adapted reconstruction and Section 7.5.1 introduces our application. In this application, we jointly train the learned primal-dual to perform tomographic reconstruction with two different tasks, wavefront set extraction and wavefront set inpainting, the latter improving the state-of-the-art methods. Finally, Chapter 8 presents the numerical experiments that support the theory in the previous chapter, and Chapter 9 concludes the work and discusses future challenges.

2 Microlocal analysis

The study of singularities of signals has an important role in different scientific areas, mainly due to the significant amount of information contained in them. When working with one-dimensional signals, the singularities are points in the domain in which the signals are non-smooth. Being points, the sole description of their location is sufficient to fully describe them. Next, 2D signals contain anisotropic features, resulting in oriented singularities. In this case, the location of the singularities cannot fully describe them. In addition, one also needs to describe their orientations.

Microlocal analysis was introduced as a tool to describe oriented singularities and their behavior under the action of a certain class of operators. It is also a theory defined on continuous spaces, depending strongly on asymptotic analysis when taking infinitely small scales.

This chapter is intended to present the main concepts and results in microlocal analysis, which will allow us to further extend the theory into a digital form, in order to apply it in the analysis of deep learning models. For that, we need to first introduce the main concepts of distribution theory, since it is where the concept of wavefront set originates from.

2.1 Distribution theory

Before we start, we would like to mention that this section is based on [62] and [86]. The word "distribution" appears in physics whenever one needs to describe a "function-like" physical concept that does not rigorously follow the definition of a classical function. The classical example is the Dirac δ-distribution, introduced to describe the density of a point mass.

Example 2.1.1 (Dirac δ-distribution)**.** *The definition of the Dirac $\delta-$distribution is motivated by the need to define a function that full-fills*

$$\int_{\mathbb{R}} \delta(x-a) f(x) dx = f(a), \quad \textit{for all } f \in C^{\infty}(\mathbb{R}) \textit{ and } a \in \mathbb{R}.$$

In particular, we also aim the function to be non-negative, non-zero only in a single point, and follows

$$\int_{\mathbb{R}} \delta(x) dx = 1.$$

Since there is no function that full-fills those properties, one needs to introduce the notion of distributions.

Notice from Example 2.1.1 that no classical function has these properties. Distributions are mappings that have similar properties. They arise naturally in the theory of partial differential equations (PDEs) [39]. Fundamental solutions of PDEs are usually singular distributions, the behavior of their singularities encode the behavior of the solutions. It is also well known that distributions are extensively used in quantum mechanics to describe the wave function. The study and the behavior of singularities of distributions can also be applied to classical function spaces. In the following, we outline the basic parts of distribution theory that are necessary for the understanding of microlocal analysis. Let us first start with the notion of some relevant spaces of continuous functions.

Definition 2.1.2 ([86]). *Let $\Omega \subset \mathbb{R}^n$ be an open set, and $\mathcal{E}(\Omega) := C^\infty(\Omega)$ be the vector space of real (or complex) valued smooth functions on Ω. In addition, the support of a function $f \in \mathcal{E}(\Omega)$ is defined by:*

$$\mathrm{supp}(f) = \overline{\{x \in \Omega | f(x) \neq 0\}}.$$

Therefore, one can define $\mathcal{D}(\Omega) := C_0^\infty(\Omega)$ as the space of smooth functions compactly supported in Ω, i.e., functions whose support is a compact subset of Ω.

When studying a differential equation coming from physics, one sometimes assumes that the solutions are in $\mathcal{E}(\Omega)$ or $\mathcal{D}(\Omega)$. Unfortunately, not every function is differentiable and the concept of distribution is the solution for this flaw, by taking an extension of the space of continuous functions where differentiation is always defined. In order to define this notion formally, we need to introduce the notion of topological space and topological dual:

Definition 2.1.3 (Topological space,[86]). *Let X be a set. A set $\mathcal{T}_X \subset 2^X$ is a* topology *of X if*

(i) $\varnothing \in \mathcal{T}_X$.

(ii) $X \in \mathcal{T}_X$.

(iii) The arbitrary union of elements of $\mathcal{T}_X$ is also an element of $\mathcal{T}_X$.

(iv) The finite intersection of elements of $\mathcal{T}_X$ is also an element of $\mathcal{T}_X$.

We refer to the tuple $(X, \mathcal{T}_X)$ as a topological space*. We call elements in the topology $\mathcal{T}_X$* open sets*.*

There are different ways to define a topology on $\mathcal{E}(\Omega)$ and $\mathcal{D}(\Omega)$, in this work we will work with the so-called *Whitney topologies* (see [86, Definition 4.4]). These topologies allow to define the notion of topological dual on the topological spaces $(\mathcal{E}(\Omega), \mathcal{T}_{\mathcal{E}(\Omega)})$ and $(\mathcal{D}(\Omega), \mathcal{T}_{\mathcal{D}(\Omega)})$. For simplicity in notation, from now on we will just write $\mathcal{E}(\Omega)$ and $\mathcal{D}(\Omega)$ when we are referring to their corresponding topological spaces.

Definition 2.1.4 (Topological dual)**.** *Let $(X, \mathcal{T}_X)$ be a topological space. In addition let $(\mathbb{R}, \mathcal{T}_{\mathbb{R}})$ be the topological space generated by the norm topology $\mathcal{T}_{\mathbb{R}}$ given by union of elements of the basis:*

$$\tau_{\mathbb{R}} := \{B_{\epsilon, ||\cdot||_2}(x) : x \in \mathbb{R}, \epsilon > 0\},$$

where $B_{\epsilon, ||\cdot||_2}(x)$ is given by:

$$B_{\epsilon, ||\cdot||_2}(x) = \{y \in \mathbb{R} : ||x - y||_2 < \epsilon\},$$

where $|| \cdot ||_2$ is the ℓ_2-norm.

We say that a function $f : X \to \mathbb{R}$ is continuous functional with respect to the topologies *$\mathcal{T}_X$ and $\mathcal{T}_{\mathbb{R}}$, if for every $V \in \mathcal{T}_{\mathbb{R}}$ the inverse image $f^{-1}(V) \subset X$ is an element of $\mathcal{T}_X$. The* topological dual *of $(X, \mathcal{T}_X)$, namely $(X', \mathcal{T}_{X'})$ is the topological space of continuous functionals $f : X \to \mathbb{R}$ with respect to the topologies $\mathcal{T}_X$ and $\mathcal{T}_{\mathbb{R}}$, i.e.*

$$X' := \{f : X \to \mathbb{R} : f \text{ is a continuous functional}\}.$$

In this case, the dual topology *$\mathcal{T}_{X'}$ is generated by union of elements in the set*

$$\tau_{X'} := \{B_{\epsilon, ||\cdot||_{\sup}}(f) : f \in X', \epsilon > 0\},$$

where $B_{\epsilon, ||\cdot||_{\sup}}$ is defined as

$$B_{\epsilon, ||\cdot||_{\sup}}(f) := \{g \in X' : \sup_{x \in X}(|f(x) - g(x)|) < \epsilon\}.$$

From now on, we will refer to the topological dual space $(X', \mathcal{T}_{X'})$ as just X'. We are now ready to define the space of distributions.

Definition 2.1.5 ([62])**.** *Let $\Omega \subset \mathbb{R}^n$ be an open set. The set of distributions on Ω, namely $\mathcal{D}'(\Omega)$, is the* topological dual *of $\mathcal{D}(\Omega)$ (see Definition 2.1.4). Similarly, one can define the topological dual of $\mathcal{E}(\Omega)$, namely, $\mathcal{E}'(\Omega)$ whose elements are also known as* compactly supported distributions *on Ω.*

Notice that in Definition 2.1.5, the space $\mathcal{D}'(\Omega)$ is equipped with the dual topology introduced in Definition 2.1.4, although this is not explicitly mentioned. In order to define the support of a distribution we first observe that if, $\Omega_0 \subset \Omega$ is an open subset, then $\mathcal{D}(\Omega_0)$ is a closed subspace of $\mathcal{D}(\Omega)$. Furthermore, there is a natural restriction map $\mathcal{D}'(\Omega) \to \mathcal{D}'(\Omega_0)$ for any open subset. Indeed, if $u \in D'(\Omega)$, then the restriction of u to Ω_0, namely $u\big|_{\Omega_0}$ is given by

$$u\big|_{\Omega_0}(\psi) := \langle u, \psi \rangle \quad \text{for every } \psi \in \mathcal{D}(\Omega_0).$$

This ensures that the next definition is well-defined.

Definition 2.1.6 ([62])**.** *Let $u \in \mathcal{D}'(\Omega)$. The* support *of u,* $\operatorname{supp} u$, *is the smallest closed set K such that the restriction of u to $\Omega \setminus K$ is 0.*

Frequently distributions are considered "generalized functions". The next example shows the intuition behind this. Let $u \in L^1_{\text{loc}}(\Omega)$ be a locally integrable function, i.e.,

$$\int_K |u(x)|dx < +\infty \quad \text{for every compact set } K \subset \Omega.$$

Then u is a distribution with the standard definition

$$u(\psi) := \int_\Omega u(x)\psi(x)dx, \quad \text{for every } \psi \in \mathcal{D}(\Omega). \tag{2.1.1}$$

From now on, we will call the smooth functions with compact support $\psi \in \mathcal{D}(\Omega)$ used in (2.1.1), *test functions*. The map $L^1_{\text{loc}}(\Omega) \to \mathcal{D}'(\Omega)$ defined by (2.1.1) is injective, which means that u is almost everywhere determined by the distribution. In particular, every smooth function defines a distribution, and the support of a function as a distribution coincides with its support as a function. We obtain the following inclusions

$$\mathcal{D}(\Omega) \subset \mathcal{E}'(\Omega) \subset \mathcal{D}'(\Omega) \quad \text{and} \quad \mathcal{E}(\Omega) \subset \mathcal{D}'(\Omega).$$

In general one can define the partial derivatives of a distribution as follows:

Definition 2.1.7 ([62])**.** *Let $u \in \mathcal{D}'(\Omega)$ be a distribution on Ω. The partial derivative of u of order α is defined by*

$$(\partial^\alpha u)(\psi) := (-1)^{|\alpha|} u(\partial^\alpha \psi) \quad \textit{for } \psi \in \mathcal{D}(\Omega), \tag{2.1.2}$$

where $\partial^\alpha \psi = (-i)^{\alpha_1 + \ldots + \alpha_n} \frac{\partial^{\alpha_1}\psi}{\partial x^{\alpha_1}} \cdots \frac{\partial^{\alpha_n}\psi}{\partial x^{\alpha_n}}$, for the multi-index $\alpha = (\alpha_1, \ldots, \alpha_n)$. In addition, $|\alpha| = \sum_{i=1}^n |\alpha_i|$.

Notice now that, if $\mathcal{P} : \mathcal{D}(\Omega) \to \mathcal{D}(\Omega)$ is a differential operator, then for a distribution $u \in \mathcal{D}'(u)$ one can define $\mathcal{P}u \in \mathcal{D}'(u)$ given by

$$(\mathcal{P}u)(\psi) := u(\mathcal{P}\psi), \quad \text{for every } \psi \in \mathcal{D}(\Omega).$$

In this sense, the concept of distribution naturally arises from the extension of differential operators to non-smooth functions. Therefore, one can think of distribution theory as the completion of differential calculus, the same we see Lebesgue integration theory as completion of integral calculus. Another important notion that can be naturally extended to the distribution realm is the Fourier transform. For this, we need to introduce a third space of distributions, the space of *tempered distributions*. For that, we need to first introduce the notion of a Schwartz function.

Definition 2.1.8 (Schwartz functions)**.** *Let $\psi : \mathbb{R}^n \to \mathbb{R}$ for $n \in \mathbb{N}$ be a* Schwartz function *if*

$$d_{\alpha,\beta}(\psi) := \sup_{x \in \mathbb{R}^n} |x^\alpha \partial^\beta \psi(x)| < \infty \quad \textit{for any multi-indices } \alpha, \beta \in \mathbb{N}_0^n. \tag{2.1.3}$$

The Schwartz space $\mathcal{S}(\mathbb{R}^n)$ *is the set of all Schwartz functions in $\mathbb{R}^n$.*

The notion of Schwartz function gives rise to the *tempered distributions*. In order to formally define them, we need to first introduce the notion of *semi-norm topology*.

Definition 2.1.9 (Semi-norm)**.** *Let X be a vector space over the real numbers $\mathbb{R}$. A real-valued function $d : X \to \mathbb{R}$ is called a* semi-norm *if it satisfies the following two conditions:*

1. ***Triangle inequality:*** *$d(f+g) \leq d(f) + d(g)$ for all $f, g \in X$.*
2. ***Absolute homogeneity:*** *$d(sf) = |s|d(f)$ for all $f \in X$ and $s \in \mathbb{R}$.*

Notice that the function $d_{\alpha,\beta} : \mathcal{S}(\mathbb{R}^n) \to \mathbb{R}$ defined in (2.1.3) is a semi-norm.

Definition 2.1.10 (Semi-norm topology)**.** *Let X be a vector space and $D = \{d_i\}_{i\in\mathcal{I}}$ a countable family of semi-norms (see Definition 2.1.9) on X. The* semi-norm topology *$\mathcal{T}_X$ is then given by the union of elements of the set*

$$\tau_X : = \{B_{i,\epsilon}(f) : f \in X, i \in \mathcal{I}, 0 < \epsilon\},$$

where

$$B_{i,\epsilon}(f) : = \{g \in X : d_i(f-g) < \epsilon\}.$$

We are now ready to introduce the notion of *tempered distributions*.

Definition 2.1.11 (Tempered distributions)**.** *Let $X : = (\mathcal{S}(\mathbb{R}^n), \mathcal{T}_{\mathcal{S}(\mathbb{R}^n)})$ be the topological space where the topology $\mathcal{T}_{\mathcal{S}(\mathbb{R}^n)}$ is generated by the semi-norm topology with semi-norms* (2.1.3)*. The space of* tempered distributions *$S'(\mathbb{R}^n)$ is the topological dual of X.*

Now, since $\mathcal{S}(\mathbb{R}^n) \subset L^2(\mathbb{R}^n)$ we can define the Fourier transform as

$$\hat{\psi}(\xi) := \int_{\mathbb{R}^n} \psi(x) e^{-2\pi i \langle x, \xi \rangle} dx \quad \text{for } \xi \in \mathbb{R}^n,$$

meaning that the Fourier transform of a tempered distribution $u \in \mathcal{S}'(\mathbb{R}^n)$ be define as

$$\hat{u}(\psi) := u(\hat{\psi}) \quad \text{for every } \psi \in \mathcal{S}(\mathbb{R}^n). \tag{2.1.4}$$

Using Plancharel's formula, one can show that the Fourier transform with the above definition extends to a weak-* continuous linear map from $\mathcal{S}'(\mathbb{R}^n)$ to $\mathcal{S}'(\mathbb{R}^n)$, see [99]. Since this class of distributions will play a central role in this thesis, we would like to also introduce the notion of Schwartz functions and tempered distributions for open sub-domains of $\mathbb{R}^n$.

Definition 2.1.12. *Let $\Omega \subset \mathbb{R}^n$ be an open domain of $\mathbb{R}^n$. The space of Schwartz functions over Ω, $\mathcal{S}(\Omega)$, is defined as the set of functions on Ω such that their extension by* 0 *to all of $\mathbb{R}^n$ is a Schwartz function. Furthermore, the space of tempered distributions over Ω, $\mathcal{S}'(\Omega)$ is defined as the topological dual of the space $(\mathcal{S}(\Omega), \mathcal{T}_{\mathcal{S}(\Omega)})$. Here the topology $\mathcal{T}_{\mathcal{S}(\Omega)}$ is semi-norm topology given by the semi-norms*

$$d_{\alpha,\beta}(\psi) : = \sup_{x\in\Omega} |x^\alpha \partial^\beta \psi(x)| < \infty.$$

Remark 2.1.13. *Following Definition 2.1.12, we can now think of the Fourier transform as mapping from $\mathcal{S}(\Omega)$ to $\mathcal{S}(\mathbb{R}^n)$. Let $\psi \in \mathcal{S}(\Omega)$, then by Definition 2.1.12 there is a extension by* 0 *to all* $\mathbb{R}^n$, *namely $\widetilde{\psi} \in \mathcal{S}(\mathbb{R}^n)$. The Fourier transform of ψ, $\hat{\psi} \in \mathcal{S}(\mathbb{R}^n)$, is given by*

$$\psi(\xi) := \int_{\mathbb{R}^n} \widehat{\widetilde{\psi}}(x) e^{-2\pi i \langle x, \xi \rangle} dx = \int_{\Omega} \psi(x) e^{-2\pi i \langle x, \xi \rangle} dx \quad \textit{for all } \xi \in \mathbb{R}^n.$$

Since later in Section 2.3 and 2.4 we are going to associate the frequencies $\xi \in \mathbb{R}^n$ to orientations, the Fourier transform applied to Schwartz functions is considered from now on as a mapping from $\mathcal{S}(\Omega) \to \mathcal{S}(\mathbb{R}^n \setminus \{0\})$. Similarly as in (2.1.4) we can also define the Fourier transform as a mapping from $\mathcal{S}'(\Omega)$ to $\mathcal{S}(\mathbb{R}^n \setminus \{0\})$ given by

$$\hat{u}(\psi) := u(\hat{\psi}) \quad \text{for every } \psi \in \mathcal{S}(\Omega).$$

We will use this type of extension, based on duality, extensively in the rest of the thesis, mainly to define operators in tempered distributions.

Finally, the Fourier transform in $\mathcal{S}(\Omega)$ relates the regularity of a function ψ at point x to the asymptotic decay rate of its Fourier transform locally. More specifically, if $C^{|\alpha|}(\Omega)$ is the space of $|\alpha|-$times differentiable functions, and $\psi \in \mathcal{S}(\Omega) \cap C^{|\alpha|}(\Omega)$, then

$$\widehat{\partial^\alpha \psi}(\xi) = \xi^\alpha \hat{\psi}(\xi).$$

This means that if $\hat{\psi}$ decays as $\mathcal{O}(|\xi|^{-\alpha})$ then ψ is α-differentiable. Microlocal analysis extends this notion to define singularities in distributions. The next section will explore this direction in detail.

2.2 The wavefront set

The main motivation to study singularities of distributions has its origins in the history of modern physics, namely, in the formalization of quantum mechanics. Feynman propagators are distributions, and Stueckelberg realized very early that renormalization was essentially the problem of defining a product of distributions [109]. Renormalization can be understood as the change of metric function that makes important physical quantities finite, and it plays a fundamental role in the formal construction of quantum field theory.

In [76], it was shown that distributions cannot in general be multiplied. The first reason is that, while distributions generalize the concept of functions, there is no way to define the evaluation of general distribution at points in Ω. For example, we cannot evaluate the $\delta-$function at 0. Hence, the multiplication of distributions cannot be defined value-wise.

One can illustrate the fact above by studying the family of characteristic functions $\chi_\epsilon : \mathbb{R} \to \mathbb{R}$ defined by

$$\chi_\epsilon(x) := \begin{cases} \frac{1}{\epsilon} & \text{if} \quad |x| \leq \epsilon/2, \\ 0 & \text{otherwise.} \end{cases}$$

For any $\psi \in \mathcal{D}(\mathbb{R})$, we have

$$\int_{\mathbb{R}} \chi_\epsilon(x)\psi(x)dx = \epsilon^{-1}\int_{-\epsilon/2}^{\epsilon/2}\psi(x)dx = \epsilon^{-1}(\epsilon\psi(0) + \mathcal{O}(\epsilon^3)),$$

which leads to the limit convergence $\lim_{\epsilon\to 0}\chi_\epsilon = \delta$. However, the square of χ_ϵ does not converge to a distribution, since

$$\int_{\mathbb{R}} \chi_\epsilon^2(x)\psi(x)dx = \epsilon^{-2}\int_{\epsilon/2}^{\epsilon/2} f(x)dx = \epsilon^{-2}(\epsilon f(0) + \mathcal{O}(\epsilon^3))$$

diverges for $\epsilon \to 0$.

It is natural to ask ourselves, in which cases we can multiply distributions. The simplest case is when one of the two distributions is a smooth function. Indeed, consider the distribution $u \in \mathcal{D}'(\mathbb{R}^n)$ and a smooth function $\phi \in \mathcal{E}(\mathbb{R}^n)$. Then, for all test function $\psi \in \mathcal{D}(\mathbb{R}^n)$ we can define the product $u\phi$ by $\langle u\phi, \psi\rangle = \langle u, \phi\psi\rangle$. This idea can be extended even further. For this, we need to understand the notion of singular support.

Notice that, on the one hand, from the definition of the support of a function (Definition 2.1.2), a function $\psi \in \mathcal{D}(\mathbb{R}^n)$ can vanish at isolated points of its support. On the other hand, one cannot define the support of a distribution $u \in \mathcal{D}'(\mathbb{R}^n)$ in the same fashion, since the value of a distribution at a point is generally not defined. In distribution theory, one needs to make use of the notion of duality to define the support. In that sense, if $u \in \mathcal{D}'(\mathbb{R}^n)$, a point $x \in \mathbb{R}^n$ is on $\operatorname{supp}(u)$ if and only if there is no open neighborhood $U \ni x$ such that $u = 0$ on U. More accurately, there is no neighborhood $U \ni x$ such that

$$\langle u, \psi\rangle = 0 \quad \text{for all } \psi \in \mathcal{D}(\mathbb{R}^n) \text{ with } \operatorname{supp}(\psi) \subseteq U.$$

Similarly, one can define the *singular support*, as the set where the distribution is not smooth.

Definition 2.2.1 (Singular support). *Let $u \in \mathcal{D}'(\mathbb{R}^n)$ be a distribution, then the* singular support *of u is the set*

$$\operatorname{sing\,supp}(u) := \{x \in \mathbb{R}^n : \textit{there is a neighborhood } U \textit{ of } x\textit{, s.t. } u|_U \textit{ is smooth}\}^c;$$

In other words, $x \in \operatorname{sing\,supp}(u)$ if and only if for every neighborhood $U \ni x$ there is no smooth function $\phi \in \mathcal{E}(U)$ such that

$$\langle u, \psi\rangle = \langle \phi, \psi\rangle = \int_U \phi(x)\psi(x)dx \quad \textit{for all } \psi \in \mathcal{D}(\mathbb{R}^n) \textit{ with } \operatorname{supp}(\psi) \subseteq U.$$

Having defined the notion of the support and singular support of a distribution, we are able to define the product of distributions with weaker assumptions tha smoothness.

Theorem 2.2.2 ([16, Section 2.1]). *If u and v are two distributions in $\mathcal{D}'(\mathbb{R}^n)$ such that $\operatorname{sing\,supp}(u) \cap \operatorname{sing\,supp}(v) = \varnothing$, then the product uv, given by*

$$uv(\psi) := \int_{\mathbb{R}^n} u(x)v(x)\psi(x)dx \textit{for every } \psi \in \mathcal{D}(\mathbb{R}^n) \tag{2.2.1}$$

is well-defined.

Generally, just knowing whether the product of two distributions exists is not sufficient, since we would like to compute it. For this purpose we can use of the properties of the Fourier transform of a product in order to define uv, in particular

$$\langle \hat{u}, v\rangle = \langle u, \hat{v}\rangle, \text{ and } \widehat{\psi\phi} = \hat{\psi} * \hat{\phi},$$

where $*$ represents the convolution defined by

$$\hat{\psi} * \hat{\phi}(\xi) = \int_{\mathbb{R}^n} \hat{\psi}(\eta)\hat{\phi}(\xi-\eta)d\eta \quad \psi, \phi \in \mathcal{D}'(\mathbb{R}^n).$$

Definition 2.2.3 ([16, Section 2.2])**.** *Let $u, v \in \mathcal{D}'(\mathbb{R}^n)$ be distributions. We say that $w \in \mathcal{D}'(\mathbb{R}^n)$ is the* product *of u and v if and only if, for each $x \in \mathbb{R}^n$, there exists some test function $\psi \in \mathcal{D}(\mathbb{R}^n)$, with $\psi = 1$ in a neighborhood of x, so that for each $\omega \in \mathbb{R}^n$ the integral*

$$\widehat{\psi^2 w}(\omega) = (\widehat{\psi u} * \widehat{\psi v})(\omega) := \int_{\mathbb{R}^n} \widehat{\psi u}(\xi)\widehat{\psi v}(\omega-\xi)d\xi \tag{2.2.2}$$

converges absolutely. In addition, under these conditions, we can define w by the left-hand side term in (2.2.1)*.*

Notice that the right-hand side in (2.2.2) is well-defined. This is due to the fact that ψu and ψv are compactly supported distributions. Therefore, due to the Paley-Wiener-Schwartz theorem (see [61]) their Fourier transform is a smooth function in a classical way. Also notice that Theorem 2.2.2 presents a sufficient condition for the well-defined notion of the product of two distributions, with assumptions on the singular support. The singular support of the distribution contains the location of its singularities. But this is not enough information to define the product in general cases. Indeed, according to Definition 2.2.3, one also needs to study the Fourier transform of each distribution, when multiplied by a test function.

The wavefront set was introduced by Hörmander [62] to find a sufficient condition by which the product of two distributions is well-defined. In that sense, we would like to find conditions, so that the integral (2.2.2) is absolutely convergent. Now, if $\psi \in \mathcal{D}(\mathbb{R}^n)$ is a test function, then ψv is compactly supported. Thus, there is a constant $C > 0$ and an integer $m \in \mathbb{Z}$ such that

$$|\widehat{\psi v}(\xi-\omega)] \leq C(1+|\xi-\omega|)^m \quad \text{for all } \xi, \omega \in \mathbb{R}^n.$$

The smallest integer m for which this is satisfied is called the *order* of the distribution ψv. The integral in (2.2.2) would be absolutely convergent if there is a constant $C' > 0$ such that

$$|\widehat{\psi u}(\xi)| \leq C'(1+|\xi|)^{-m-n-1} \quad \text{for some } n \in \mathbb{N}.$$

Next, we would also like the product of distributions to follow the Leibniz rule under derivation

$$\partial(uv) = (\partial u)v + u(\partial v).$$

But, since the derivative of order n decreases the order of regularity of u by n, what we really need is that

$$|\widehat{\psi u}(\xi)| \leq C'(1+|\xi|)^{-N} \quad \text{for every } N \in \mathbb{N}.$$

This required condition motivates the introduction of the wavefront set as a tool to determine when the product of distributions is well-defined.

As we did with the singular support, we are going to define the wavefront set by its complement, which is the set of points where the estimates above are attained, and therefore, where the multiplication of the distributions is defined. The definition of the wavefront set of a distribution $u \in \mathcal{D}'(\mathbb{R}^n)$ involves two localization processes. The first is done in the spatial domain, by the multiplication with a smooth test function ψ supported on a neighborhood of a point. The second is done in the Fourier domain in order to analyze the decay rate of the function $\widehat{\psi u}$. This localization is done in the directional sense, meaning, one restricts the frequency directions $\xi \in \mathbb{R}^n \setminus \{0\}$ to a neighborhood of directions. The directional localization is referred to as microlocalization. This is the reason why the study of oriented singularities is known as *microlocal analysis*. In order to faithfully localize the directions, we need to define the notion of fast decreasing in a conical neighborhood.

Definition 2.2.4 (Conical neighborhood)**.** *A* conical neighborhood *of $\xi \in \mathbb{R}^n$ is a set $V \subset \mathbb{R}^n$ such that for any $\nu \in V$ and $\alpha > 0$, $\alpha\nu \in V$ and V contains the ball*

$$B(\xi,\epsilon)\colon=\{\omega \in \mathbb{R}^n : |\omega - \xi| < \epsilon\},$$

where $\epsilon > 0$ is the radius of $B(\xi,\epsilon)$.

Definition 2.2.5 ([16, Section 3.1])**.** *A smooth function $\phi \in \mathcal{D}(\mathbb{R}^n)$ is said to be* fast decreasing on a conical neighborhood $V \subset \mathbb{R}^n$ *if, for any integer N, there is a constant $C_N > 0$ such that*

$$|\phi(\xi)| \leq C_N(1+|\xi|)^{-N} \quad \textit{for all } \xi \in V.$$

Spatial-frequency points holding the property of fast decreasing in the conical neighborhood are known as *directed smooth points.* Points that are not directed smooth are part of the wavefront set. This notion is defined in the following

Definition 2.2.6 (Wavefront set, [62, Section 8.1])**.** *Let $u \in \mathcal{D}'(\mathbb{R}^n)$ a distribution and $N \in \mathbb{N}$. A point $(x;\xi) \in \mathbb{R}^n \times \mathbb{R}^n \setminus \{0\}$ is an N-*regular directed point *of u if there exist an open neighborhood of x, namely U_x, and a conical neighborhood of ξ, namely V_ξ, and a* smooth cut-off function *$\psi \in D(\mathbb{R}^n)$ with* $\operatorname{supp}\psi \subset U_x$ *and $\psi(x) = 1$ such that*

$$|\widehat{\psi u}(\xi)| \leq C_N(1+|\xi|)^{-N} \qquad \textit{for all } \xi \in \mathbb{R}^n \textit{ with } \xi/|\xi| \in V_\xi$$

*holds for some $C_N > 0$. The N-*wavefront set $\mathrm{WF}_N(u)$ *is the set given by*

$$\mathrm{WF}_N(u)\colon=\{(x;\xi) \in \mathbb{R}^n \times \mathbb{R}^n \setminus \{0\} : (x;\xi) \textit{ is an } N\textit{−regular directed point}\}^c$$

Finally, the wavefront set $\mathrm{WF}(u)$ *is defined as*

$$\mathrm{WF}(u) := \bigcup_{N\in\mathbb{N}} \mathrm{WF}_N(u).$$

Notice that the *singular support* of u can be characterized in terms of the wavefront set as

$$\operatorname{sing\,supp}(u) : = \{x \in \mathbb{R}^n : (x;\xi) \in \mathrm{WF}(u) \text{ for some } \xi \in \mathbb{R}^n\}.$$

In other words, for each point in $\operatorname{sing\,supp}(u)$, the wavefront set of u, $\mathrm{WF}(u)$ is composed of the directions where the Fourier transform of ψu is not fast decreasing, with ψ a cut-off function with sufficiently small support.

Definition 2.2.6 assumes that $u \in \mathcal{D}'(\mathbb{R}^n)$ and $WF(u) \in \mathbb{R}^n \setminus \{0\}$. This holds when $u \in \mathcal{D}'(X)$ with X an open subset of $\mathbb{R}^n$ which is the case for many of the distributions considered in this thesis. However, this is not the case when X is a C^∞ manifold. For instance, the distributions on the Radon transform data (Definition 2.3.1) are distributions on manifolds. In order to introduce the general definition of wavefront set we need to first present the notion of a C^∞ structure on a manifold X.

Definition 2.2.7 ([61, Definition 6.3.1])**.** *Let X be an $n-$dimentional manifold. A C^∞ structure on a manifold X is a family $\mathcal{F}$ of homeomorphisms κ of open sets $X_\kappa \subset X$ on open sets $\tilde{X}_\kappa \subset \mathbb{R}^n$, called* local coordinate systems*, such that it holds*

(i) If $\kappa, \kappa' \in \mathcal{F}$, then the map

$$\kappa'\kappa^{-1} : \kappa(X_\kappa \cap X_{\kappa'}) \to \kappa'(X_\kappa \cap X_{\kappa'}) \tag{2.2.3}$$

(between open sets in $\mathbb{R}^n$) is infinitely differentiable. The same holds for the inverse map.

(ii) $\bigcup X_\kappa = X$.

(iii) If κ_0 is a homeomorphism of an open set $X_0 \subset X$ on an open set in $\mathbb{R}^n$ and the map

$$\kappa\kappa_0^{-1} : \kappa_0(X_0 \cap X_\kappa) \to \kappa(X_0 \cap X_\kappa)$$

as well as its inverse is infinitely differentiable for every $\kappa \in \mathcal{F}$, it follows that $\kappa_0 \in \mathcal{F}$.

A manifold with C^∞ structure is called a C^∞ *manifold.* The sets X_κ are called *coordinates patches* and the cartesian coordinates of $\kappa(x)$, $x \in X_\kappa$, are called *local coordinates* in X_κ. Since the wavefront set is contained in the co-tangent bundle of the space X we will introduce some preliminary notions before the general definition of the wavefront set.

Definition 2.2.8 (Topological vector bundle)**.** *Let $(X, \mathcal{T})$ be a topological space. Then a* topological vector bundle *over X consists of:*

1. *A topological space E.*

2. *A continuous function* $\pi : E \to X$.
3. *For each* $x \in X$*, the structure of a finite-dimensional* k*-vector space on the pre-image*

$$E_x := \pi^{-1}(\{x\}) \subset E$$

such that there exists:

- *An open cover* $\{U_i \subset X\}_{i \in I}$.
- *For each* $i \in I$ *and* $n_i \in \mathbb{N}$ *and a homeomorphism*

$$\phi_i : U_i \times k^{n_i} \to \pi^{-1}(U_i) \subset E$$

such that $\phi_i(\{x\} \times k^{n_i}) \subset \pi^{-1}(\{x\})$ *and* ϕ_i *is a linear map in each* fiber *with*

$$\phi_i(x) : k^{n_i} \to E_x = \pi^{-1}(\{x\}) \quad \textit{for all } x \in U_i.$$

Definition 2.2.9 (Tangent bundle and cotangent bundle)**.** *The* tangent bundle *over a space* X *is a* topological vector bundle *whose fiber over a point* $x \in X$ *is the* tangent space *at that point. The* cotangent bundle $T^*(X)$ *is the dual vector bundle of* $T(X)$*, that is, each fiber is the dual vector space of the corresponding fiber on* $T(X)$*.*

Definition 2.2.10 (Wavefront set in manifolds)**.** *Let* X *be a* C^∞ *manifold and* $u \in \mathcal{D}'(X)$ *a distribution on* X*. Moreover, let* $T^*(X)$ *be the cotangent bundle of* X*. The wavefront set* $WF(u) \subset T^*(X) \setminus \{0\}$ *is defined so that the restriction to a coordinate patch of* X*,* X_κ *is equal to the pullback* $\kappa^* \operatorname{WF}(u \circ \kappa^{-1})$*.*

With this definition $\operatorname{WF}(u)$ is a closed subset of $T^*(X) \setminus \{0\}$ which is conic in the sense that the intersection with the vector space given by the fiber $T^*_x(X)$ is a cone for every $x \in X$.

The name *wavefront set* is inspired by the fact that the singularities of the solutions of the wave equation move within it (see [62]), meaning that the wavefront set describes the evolution of the wavefront. We are now ready to introduce a sufficient condition to have a well-defined product between distributions, i.e., the distribution $w \in \mathcal{D}'(\mathbb{R}^n)$ from Definition 2.2.3 exists and is unique. This condition is also known as *the Hörmander condition*, since it was first introduced by Hörmander.

Theorem 2.2.11 (Product theorem/Hörmander condition, [16, Section 3.2])**.** *Let* u *and* v *be distributions in* $\mathcal{D}'(U)$*. Suppose that there is no point* $(x;\xi) \in \operatorname{WF}(u)$*, such that* $(x;-\xi) \in \operatorname{WF}(v)$*, then the product* uv *is uniquely and well-defined by* w *of Definition 2.2.3 (left-hand side term in* (2.2.1)*). Moreover, in this case*

$$\operatorname{WF}(uv) \subset S_+ \cup S_u \cup S_v,$$

where $S_+ := \{(x;\xi+\omega) : (x;\xi) \in \operatorname{WF}(u)$ *and* $(x;\omega) \in \operatorname{WF}(v)\}$*,* $S_u := \{(x;\xi) : (x;\xi) \in \operatorname{WF}(u)$ *and* $x \in \operatorname{supp}(v)\}$ *and* $S_v := \{(x;\xi) : (x;\xi) \in \operatorname{WF}(v)$ *and* $x \in \operatorname{supp}(u)\}$*.*

This theorem provides a criterion, from which we can prove that a product of distributions exists. This criterion holds even if we cannot compute their Fourier transforms $\widehat{\psi u}$ and $\widehat{\psi v}$, and do not know the explicit form of the distributions. The product theorem has been widely applied in quantum field theory, for example in the definition of the Feynman propagator. In addition, it is also fundamental for the theory of renormalization in curved space-times.

In the following remark we extend the notion of wavefront set and the product theorem to $\mathcal{S}'(\Omega)$.

Remark 2.2.12. *Let us first notice that the notion of* singular support *(Definition 2.2.1) and* wavefront set *(Definition 2.2.6) is easily extended to the space $\mathcal{D}'(\Omega)$. In addition, since $\mathcal{S}'(\Omega) \subset \mathcal{D}'(\Omega)$, both notions can also be extended to the space of tempered distributions.*

Now, the wavefront set can play an important role in inverse problems. One, for example, could aim to recover the wavefront set of the signal under study which in some cases is easier than recovering the signal itself. The interest behind recovering the wavefront set comes from the importance of directed singularities in mathematical image processing, where they carry most of the information of the image to which they belong to. Microlocal analysis provides the tools to analyze the transformation of the wavefront set under a certain class of operators, namely, the Fourier integral operators, by the so-called micro-canonical relations. Fourier integral operators arise naturally in many real-world applications, for example, computed tomography (CT) reconstruction [116] and partial differential equations [18]. The micro-canonical relation allows us to have partial access to the wavefront set of an image, from its X-ray measurements, without the need for any reconstruction. It is also used to analyze the singularities propagation in the time evolution of a PDE. We will explore this concept in more depth on Section 2.3 and extend this notion to certain kind of neural network architectures in Section 4.

In order to bring a bit of light in Definition 2.2.6, we would like to introduce in the following some examples, in which we compute the wavefront set of certain distributions.

Example 2.2.13. *The simplest example is the Dirac delta distribution $\delta \in \mathcal{D}'(\mathbb{R}^n)$ (see Definition 2.1.1). The singular support of $\delta(x)$ is $\{0\}$ and for $\psi \in \mathcal{D}'(\mathbb{R}^n)$ and $\xi \in \mathbb{R}^n$, we have that*

$$\widehat{\psi\delta}(\xi) = \psi(0) \quad \textit{for all } \xi \in \mathbb{R}^n,$$

which is not fast decreasing if $\psi(0) \neq 0$. This shows that $\{(0;\xi) : \xi \in \mathbb{R}^n\} \subset \mathrm{WF}(\delta)$. Now, if one considers any directed point $(0;\xi) \in \mathrm{WF}(\delta)$, then $(0;-\xi) \in \mathrm{WF}(\delta)$. This means that the Hörmander condition (Theorem 2.2.11) is not satisfied when trying to define δ^2 and thus the powers of δ cannot be defined.

We would also like to compute the wavefront set for a more general class of distributions. In imaging science, distributions representing characteristic functions of general domains are fundamental. The characteristic function of a domain $\Omega \subset \mathbb{R}^n$ is the function:

$$\chi_\Omega := \begin{cases} 1 & \text{if } x \in \Omega, \\ 0 & \text{otherwise.} \end{cases} \tag{2.2.4}$$

The characteristic function corresponds to the distribution u_Ω defined by $\langle \chi_\Omega, f \rangle = \int_\Omega f(x)dx$. In the case where the domain $\Omega \subset \mathbb{R}^n$ is bounded by a smooth surface $\delta\Omega$, we can compute $\mathrm{WF}(u_\Omega)$ with the following proposition.

Proposition 2.2.14. *Let $\Omega \subset \mathbb{R}^n$ be a region with smooth boundary $\partial\Omega$, and let χ_Ω be the characteristic distribution of Ω. Then*

$$\mathrm{WF}(u_\Omega) = \{(x;\xi) : x \in \partial\Omega, \textit{ and } \xi \textit{ normal to } \partial\Omega \textit{ in } x\}.$$

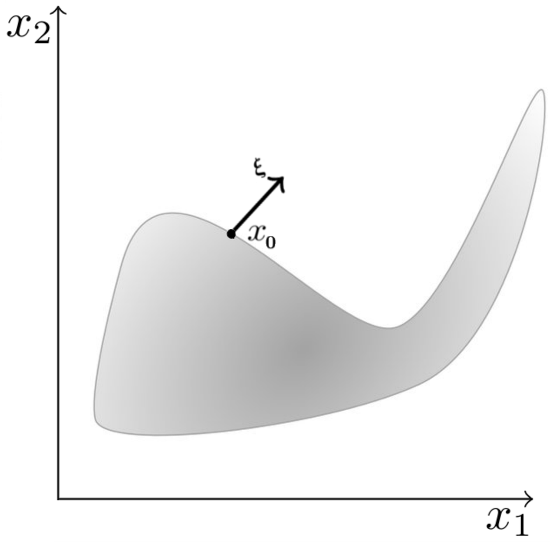

Figure 2.1: Domain Ω with smooth boundary, and normal vector to the boundary

The proof of the above proposition can be found in detail in [111, p. 129]. The set of vectors ξ which are normal to all tangent vectors to $\partial\Omega$ is known as the co-normal bundle, defined as follows.

Definition 2.2.15. *Let $\Omega \subset \mathbb{R}^n$ be an open domain in $\mathbb{R}^n$. The* co-normal bundle *of the boundary $\partial(\Omega)$ is defined as*

$$C := \{(x;\xi) \in \partial\Omega \times \mathbb{R}^n \setminus \{0\} : \xi \in C_x\},$$

where the set C_x, known as the co-normal fiber *in x, is the set*

$$C_x := \{\xi \in \mathbb{R}^n \setminus \{0\} : \xi \textit{ is normal to all tangent vectors to } \partial\Omega \textit{ at } x\}.$$

Figure 2.1 depicts an example of an element of the fiber C_x at x. Proposition 2.2.14 states that the wavefront set of u_Ω is the conormal bundle of $\partial\Omega$.

Besides some simple examples, the wavefront set is difficult to compute in practice. This is mainly due to the asymptotic criteria involved in its definition, which means computing the wavefront set requires computing the "full" Fourier transform at every point. Continuous transforms associated with certain directional multiscale systems offer a convenient remedy. We describe this situation in Chapter 3.

2.3 Pseudodifferential operators

So far we know that the singularities in a distribution contain an important amount of information. In addition, the knowledge of the orientations of the singularities becomes important in certain cases, in which one would like to understand how the singularities are transformed under the application of certain operators.

One would like to characterize the singularities that are preserved by these operators. The most common examples of operators that preserve singularities are the so-called *pseudodifferential operators* (also written by ΨDO). One can prove that if $\mathcal{P}$ is a pseudodifferential operator, any singularity in $\mathcal{P}(f)$ corresponds to a singularity in $f \in \mathcal{S}(\mathbb{R}^2)$. In some sense, the action $\mathcal{P}$ will not introduce new singularities to the function. In general, there is an explicit transformation rule from $\mathrm{WF}(f)$ to $\mathrm{WF}(\mathcal{P}(f))$.

In the case the forward operator of an inverse problem is modeled as a ΨDO $\mathcal{P}$, having at hand a way to find the singularities of a function f, by knowing the singularities of its measured data $\mathcal{P}(f)$, becomes handy in real-world applications, modeled by an inverse problem. We will motivate the notion of a pseudodifferential operator with the Radon transform. The Radon transform is the forward operator that is commonly used in the problems of parallel X-ray tomography. We can define it as follows.

Definition 2.3.1 (Radon transform)**.** *Let $\Omega \subset \mathbb{R}^2$ be an open domain and $f \in \mathcal{S}(\Omega)$ be a Schwartz function. The Radon transform of a planar function $f : \mathbb{R}^2 \to \mathbb{R}$ is the linear operator given by*

$$\mathcal{R}f(s,\varphi) := \int_{x\in L(s,\varphi)\cap\Omega} f(x)dx = \int_{-\infty}^{\infty} f(s\omega(\varphi)+t\omega^{\perp}(\varphi))\chi_{\Omega}(s\omega(\varphi)+t\omega^{\perp}(\varphi))dt \quad \text{for all } (s,\varphi)\in\mathbb{R}\times(0,\pi), \tag{2.3.1}$$

where $\omega(\varphi) := (\cos\varphi, \sin\varphi)$ is the unitary vector with orientation described by the angle φ with respect to the x_1-axis, $\omega^{\perp}(\varphi) := (-\sin\varphi, \cos\varphi)$ and

$$L(s,\varphi) := \{x \in \mathbb{R}^2 : x \cdot \omega(\varphi) = s\}$$

is the line with distance s to the origin and normal vector $\omega(\varphi)$. Finally, $\chi_\Omega : \mathbb{R}^2 \to \{0,1\}$ is the characteristic function of Ω (see (2.2.4)).

The Radon transform $\mathcal{R}f(s,\varphi)$ will measure a signal $f : \mathbb{R}^2 \to \mathbb{R}$ by its line integrals. These measurement samples are referred to as the *sinogram*. In the physical model, the line integrals model the absorption rate of the measured body with respect to the X-ray [90]. Another relevant notion in computed tomography is the adjoint of the Radon transform, also known as *back-projection*

Definition 2.3.2 (Back-projection)**.** *Let $\Omega \subset \mathbb{R}^2$ be an open domain. The* back-projection *of a function $g : \mathbb{R}\times(0,\pi) \to \Omega$ is given by*

$$\mathcal{R}^*(g)(x) := \int_0^{\pi} g(x\cdot\omega(\theta),\theta)\,d\theta \quad \text{for } x \in \Omega. \tag{2.3.2}$$

The back-projection maps a function $g : \mathbb{R} \times (0, \pi) \to \Omega$ defined on lines in Ω to a function $f : \Omega \to \mathbb{R}$ defined on points in $x \in \Omega$ by simply averaging g over all lines that go through x.

Remark 2.3.3. *In this thesis we will consider the Radon transform of Schwartz functions $\mathcal{S}(\mathbb{R}^2)$ (Definition 2.1.8). Following [59, Theorem 2.4]), we have that the Radon transform $\mathcal{R}$ maps the space $\mathcal{S}(\mathbb{R}^2)$ onto $\mathcal{S}(\mathbb{R} \times (0, \pi))$, i.e., $\mathcal{R} : \mathcal{S}(\mathbb{R}^2) \to \mathcal{S}(\mathbb{R} \times (0, \pi))$. Following the same result we also have that the back-projection maps the space $\mathcal{S}(\mathbb{R} \times (0, \pi))$ onto $\mathcal{S}(\mathbb{R}^2)$.*

A simple calculation shows that the back-projection is the dual to the Radon transform [82, Theorem 2.75], i.e.,

$$\big\langle \mathcal{R}(f), g \big\rangle = \big\langle f, \mathcal{R}^*(g) \big\rangle \quad \text{for all } f \in \mathcal{S}(\Omega) \text{ and } g \in \mathcal{S}(\mathbb{R} \times (0, \pi)). \tag{2.3.3}$$

The inner product in the right-hand side above refers to the natural inner product on $L^1(\mathbb{R}^2)$, whereas the inner product in the left-hand side is the natural inner product on $L^1\big(\mathbb{R} \times (0, \pi)\big)$.

Similarly, as we did with the Fourier transform we can extend the notion of the Radon transform to tempered distributions $u \in \mathcal{S}'(\Omega)$. This can be done using the duality in (2.3.3).

Remark 2.3.4. *The duality in* (2.3.3) *can be used to extend the Radon transform to various classes of distributions, like compactly supported distributions [59] and tempered distributions [44, 80]. One can define the Radon transform on a tempered distribution $f \in \mathcal{S}'(\Omega)$ as*

$$\mathcal{R}(f)(\phi) := f\big(\mathcal{R}^*(\phi)\big) \quad \textit{for all } \phi \in \mathcal{S}\big(\mathbb{R} \times (0, \pi)\big).$$

Similarly, for a tempered distribution $g \in \mathcal{S}'(\mathbb{R} \times (0, \pi))$ we have that

$$\mathcal{R}^*(g)(\psi) := g(\mathcal{R}(\psi)) \quad \textit{for all } \psi \in \mathcal{S}(\Omega).$$

The extension of $\mathcal{R}^*$ to $\mathcal{S}\big(\mathbb{R} \times (0, \pi)\big)$ is defined analogously and one can in addition show that [82, Sections 2.9.3.2 and 4.3.1],

$$\begin{aligned} \mathcal{R}\colon \mathcal{S}'(\mathbb{R}^2) &\to \mathcal{S}'\big(\mathbb{R} \times (0, \pi)\big) && \text{is a topological isomorphism,} \\ \mathcal{R}^*\colon \mathcal{S}'\big(\mathbb{R} \times (0, \pi)\big) &\to \mathcal{S}'(\mathbb{R}^2) && \text{is surjective.} \end{aligned} \tag{2.3.4}$$

Next, in limited-angle tomography, the data is given on lines contained in some open set $\Xi \subset \mathbb{R} \times (0, \pi)$. The *partial* Radon transform can be defined as follows.

Definition 2.3.5 (Partial Radon transform)**.** *Let $\Omega \subset \mathbb{R}^2$ and $\Xi \subset (0, \pi)$ be open domains. The* partial *Radon transform $\mathcal{R}_\Xi : \mathcal{S}'(\Omega) \to \mathcal{S}'(\Xi)$ is then defined as*

$$\mathcal{R}_\Xi := P_{\mathcal{S}'(\Xi)} \circ \mathcal{R},$$

where $P_{\mathcal{S}'(\Xi)}$ is the restriction of $\mathcal{S}'(\mathbb{R}^2)$ to $\mathcal{S}'(\Xi)$.

Note that the restriction of a tempered distribution $\mathcal{S}'(\Omega)$ to $\mathcal{S}(\Xi)$ where $\Xi \subset \mathbb{R}^2$ is open is well defined and it corresponds to a tempered distribution in $\mathcal{S}'(\Xi)$. The corresponding *partial* back-projection is simply defined as in (2.3.2) but by setting g to 0 on lines not contained in Ξ.

Definition 2.3.6 (Partial back-projection)**.** *Let $\Omega \subset \mathbb{R}^2$ and $\Xi \subset (0,\pi)$ be open domains. In particular, let us assume $\Xi := \mathbb{R} \times I \subset \mathbb{R} \times (0,\pi)$ for some open interval $I \subset (0,\pi)$, then the* partial *back-projection operator $\mathcal{R}^*_\Xi : \mathcal{S}(\Xi) \to \mathcal{S}(\Omega)$ is given by*

$$\mathcal{R}^*_\Xi(g)(x) = \int_I g(x \cdot \omega(\theta), \theta)\, d\theta \quad \textit{for } x \in \Omega \textit{ and } g \in \mathcal{S}(\Xi).$$

This can be extended to $\mathcal{S}'(\Xi)$ in a similar fashion as we did for the full back-projection $\mathcal{R}^*$.

From now on we will simply write $\mathcal{R}$ and $\mathcal{R}^*$ to refer to the *partial* operators, specifying explicitly the domains Ω and Ξ.

The problem of estimating f from incomplete measurements of $\mathcal{R}_\Xi f$, known as *limited-angle tomographic reconstruction*, is severely ill-posed. Specifying the wavefront set is a strong prior for the reconstruction [38]. For this reason, one would like to understand in detail how the wavefront set of a function is transformed under the Radon transform. Let us first introduce some results to motivate such analysis.

Theorem 2.3.7 ([70, Section 1.4.2])**.** *Let $f \in \mathcal{S}(\Omega)$ and $\mathcal{R}$ be the Radon transform given by (2.3.1). Then*

$$\mathcal{R}^*\mathcal{R}f(x) = \int_\Omega e^{ix\cdot\xi} \frac{2}{||\xi||} \hat{f}(\xi) d\xi = \frac{1}{\pi} \int_{\mathbb{R}^2\setminus\{0\}} \int_\Omega e^{i(x-y)\cdot\xi} \frac{1}{||\xi||} f(y) dy d\xi. \tag{2.3.5}$$

This theorem introduces a way to express the operator $\mathcal{R}^*\mathcal{R}$ with the integral representation

$$\mathcal{P}f(x) := \mathcal{R}^*\mathcal{R}f(x) = \frac{1}{4\pi^2} \int_{\mathbb{R}^2} \int_{\mathbb{R}^2} e^{i(x-y)\cdot\xi} p(x,y,\xi) u(y) dy d\xi, \quad x \in \mathbb{R}^2. \tag{2.3.6}$$

The representation of an operator given by (2.3.6), in addition to certain estimates on the amplitude function p allows one to prove that no new singularities will be introduced under the action of $\mathcal{P}$, in other words

$$\mathrm{WF}(\mathcal{P}f) \subset \mathrm{WF}(f). \tag{2.3.7}$$

Operators of the form (2.3.6) are known as *pseudodifferential operators* also referred to as ΨDOs. The precise definition of ΨDOs will follow. Before we study (2.3.7) in detail, let us first study the simplest case, a linear partial differential operator. Consider a linear partial differential operator $\mathcal{P}$ given by

$$\mathcal{P}(x,D)=\sum_{|\nu|\leq m} a_\nu(x)D_x^\nu \quad \text{for } x\in\mathbb{R}^2 \text{ and } D:\mathcal{S}(\mathbb{R}^2)\to\mathcal{S}(\mathbb{R}^2),$$

where $\nu=(\nu_1,\ldots,\nu_n)$ is a multi-index and

$$D_x^\nu=(-i)^{(\nu_1+\ldots+\nu_n)}\frac{\partial^{\nu_1}}{\partial x_1^{\nu_1}}\cdots\frac{\partial^{\nu_n}}{\partial x_n^{\nu_n}}.$$

Applying the Fourier transform, we have

$$\widehat{D_x^\nu f}(\xi)=\frac{1}{2\pi}\int_{\mathbb{R}^2}e^{-ix\cdot\xi}D_x^\nu f(x)dx,\quad \xi\in\mathbb{R}^2\setminus\{0\}.$$

Using this form we obtain

$$\mathcal{P}(x,D)f(x)=\frac{1}{4\pi^2}\int_{\mathbb{R}^2}e^{i(x-y)\cdot\xi}p(x,\xi)f(y)dyd\xi, \tag{2.3.8}$$

where the corresponding amplitude is given by

$$p(x,\xi)=\sum_\nu a_\nu(x)\xi^\nu,\quad x\in\mathbb{R}^2 \text{ and } \xi\in\mathbb{R}^2\setminus\{0\}.$$

Notice that the function p satisfies the following property: Let α,β be any multi-indices, and $K\subset\mathbb{R}^2$ a compact set, there is a constant C such that

$$|\partial_\xi^\alpha\partial_x^\beta p(x,\xi)|\leq C(1+||\xi||)^{m-|\alpha|} \quad \text{for all } x\in\mathbb{R}^2 \text{ and } \xi\in\mathbb{R}^2\setminus\{0\}. \tag{2.3.9}$$

In other words, the differentiating with respect to ξ raises the order of decay of p. We can extend this notion to operators other than differential operators. An operator that has the form (2.3.8) where p holds the estimate (2.3.9) is *pseudodifferential operator*. We present the precise definition in the following.

Definition 2.3.8 (Amplitude function, [70, Section 1.4.2])**.** *Let $\Omega,\Xi\subset\mathbb{R}^2$ be open domains. An* amplitude of order m *is a function $p:\Xi\times\Omega\times\mathbb{R}^2\setminus\{0\}\to\mathbb{R}$ that satisfies the following properties*

1. *$p(y,x,\xi)\in C^\infty(\Xi\times\Omega\times\mathbb{R}^2\setminus\{0\})$,*
2. *For every compact $K\subset\Omega$ and for multi-indices α,β,γ,*
   a) *there is a constant $C=C(K,\alpha,\beta,\gamma)$ such that*

   $$|D_\xi^\alpha D_x^\beta D_y^\gamma p(y,x,\xi)|\leq C(1+||\xi||)^{m-|\alpha|},$$

   for $x,y\in K$ and $||\xi||>1$, and

   b) *$p(y,x,\xi)$ is locally integrable for x and y in K and $||\xi||\leq 1$.*

The amplitude is also referred to as symbol. *The term with the highest degree of a symbol p is known as the* principal symbol.

Notice that in Definition 2.3.8, the amplitude p does not necessarily need to be a polynomial in ξ, it just needs to follow the same estimates as a polynomial.

Definition 2.3.9 (Definition 5, [70, Section 1.4.2])**.** *Let $\Omega \subset \mathbb{R}^2$ be an open domain. In addition, let $\mathcal{P} : \mathcal{S}(\Omega) \to \mathcal{S}(\Omega)$ be an operator. Then, $\mathcal{P}$ is a* pseudodifferential operator *(ΨDO) of order m if for all $f \in \mathcal{S}(\Omega)$*

$$\mathcal{P}f(x) = \frac{1}{4\pi^2} \int_{\mathbb{R}^2} \int_{\mathbb{R}^2} e^{i(x-y)\cdot\xi} p(y, x, \xi) f(y) dy d\xi,$$

where p is an amplitude of order m. If in addition, for each compact set $K \subset \Omega$, there is a constant $C_K > 0$ such that,

$$|p(y, x, \xi)| \geq C_k (1 + ||\xi||)^m, \tag{2.3.10}$$

we say that $\mathcal{P}$ is an elliptic *ΨDO of order m.*

Again, by duality, we can think of the operator $\mathcal{P}$ as an operator acting in $\mathcal{S}'(\Omega)$. The estimates imposed on the amplitude function p in Definition 2.3.8 give pseudodifferential operators the property of not introducing new singularities to the functions it acts upon. This is also known as the *pseudolocal property.*

Theorem 2.3.10 (Pseudolocal property, [114, Chapter 6])**.** *If $\mathcal{P} : \mathcal{S}'(\Omega) \to \mathcal{S}'(\Omega)$ is a pseudodifferential operator, then $\mathcal{P}$ satisfies the* pseudolocal property*. Namely, for all $f \in \mathcal{S}'(\Omega)$,*

$$\operatorname{sing\,supp}(\mathcal{P}f) \subset \operatorname{sing\,supp}(f) \text{ and } \mathrm{WF}(\mathcal{P}f) \subset \mathrm{WF}(f).$$

If in addition, $\mathcal{P}$ is elliptic, then

$$\operatorname{sing\,supp}(\mathcal{P}f) = \operatorname{sing\,supp}(f) \text{ and } \mathrm{WF}(\mathcal{P}f) \subset \mathrm{WF}(f).$$

The proof of Theorem 2.3.10 can be found in [114]. It is worth to notice that this theorem implies that, although a pseudodifferential operator can spread out the support of a function, it does not spread out the singular support of the function. In the case of the operator $\mathcal{P} = \mathcal{R}^*\mathcal{R}$, we have

$$\mathcal{R}^*\mathcal{R}f(x) = \frac{1}{\pi} \int_{\mathbb{R}^n} e^{i(x-y)\cdot\xi} \frac{1}{||\xi||} f(y) dy d\xi.$$

Since $p(x, y, \xi) = \frac{4\pi}{||\xi||}$ satisfies the estimates in Definition 2.3.8 for the order $m = -1$, and satisfies the estimate of Equation (2.3.10), $\mathcal{R}^*\mathcal{R}$ is an elliptic pseudodifferential operator of order $m = -1$. Applying Theorem 2.3.10, we have that $\mathrm{WF}(\mathcal{R}^*\mathcal{R}f) = \mathrm{WF}(f)$, this becomes very useful when performing tomographic reconstruction on incomplete data, where one is able to use the pseudolocal property to find the singularities of the signal that are preserved in the data. We will later explore this approach in more detail in

Chapter 8. Notice that a pseudodifferential operator, by definition, shares the same domain and codomain. This is not the case for the Radon transform $\mathcal{R}$. The need to have a similar theory for these kind of more general operators gives rise to the concept of a *Fourier integral operator.*

2.4 Fourier integral operators

One useful property of the Radon transform is the so-called projection slice theorem. This will help us to later relate the Fourier transform of a function with its Radon transform.

Theorem 2.4.1 (Projection slice theorem, [70, Theorem 1])**.** *Let $\Omega, \Xi \subset \mathbb{R}^2$ be open domains. In addition, let $f \in L^1(\mathbb{R}^2)$, $h \in L^\infty(\mathbb{R})$ and $\omega : [0, 2\pi) \to \mathbb{R}$, given by*

$$\omega(\varphi) :=(\cos\varphi, \sin\varphi) \quad \textit{for all } \varphi \in [0, 2\pi].$$

Then the Radon transform of f, $\mathcal{R}f(s,\varphi)$ (Definition 2.3.1) follows

$$\int_{x\in\mathbb{R}^2} f(x)h(x\cdot\omega(\varphi))dx = \int_{s=-\infty}^{\infty} \mathcal{R}f(s,\varphi)h(s)ds.$$

Proof. Let $\varphi \in (0, \pi)$. First, note that the function $x \mapsto f(x)h(x\cdot\omega(\varphi))$ is in $L^1(\Omega)$ since h is bounded and measurable. For the same reason, the function

$$(s,t) \mapsto f(s\omega(\varphi) + t\omega^\perp(\varphi))h(s)$$

is in $L^1(\Omega)$. We have that

$$\begin{aligned}\int_{x\in\mathbb{R}^2} f(x)h(x\cdot\omega(\varphi)) &= \int_{s=-\infty}^{\infty}\int_{t=-\infty}^{\infty} f(s\omega(\varphi) + t\omega^\perp(\varphi))h(s)dtds \\ &= \int_{s=-\infty}^{\infty} \mathcal{R}f(s,\varphi)h(s)ds,\end{aligned}$$

where the first equality holds by rotation invariance of the Lebesgue integral [99], Fubini's theorem and the fact that $s = \omega(\varphi)\cdot(s\omega(\varphi) + t\omega^\perp(\varphi))$. The second equality holds by definition of $\mathcal{R}$. □

For simplicity, let us next introduce the *Fourier slice theorem* of the Radon transform, a special case of Theorem 2.4.1:

Theorem 2.4.2 (Fourier slice theorem, [70, Theorem 2])**.** *Let $\Omega \subset \mathbb{R}^2$ and $\Xi \subset \mathbb{R}\times(0,\pi)$ be open domains, and $f \in L^1(\mathbb{R}^2)$. Then, for $(\xi,\varphi) \in \Xi$, we have that*

$$\hat{f}(\xi\omega(\varphi)) = \frac{1}{(2\pi)^{1/2}}\mathcal{F}_s\mathcal{R}f(\xi,\varphi),$$

where $\omega(\varphi) = (\cos\varphi, \sin\varphi) \in \mathbb{S}^1$, $\xi \in \mathbb{R}\setminus\{0\}$, and $\mathcal{F}_s\mathcal{R}f$ is the Fourier transform of $\mathcal{R}f(s,\varphi)$ in the s coordinate (Definition 2.3.1).

Proof. Let $f \in \mathcal{S}(\Omega)$. By the projection slice theorem (Theorem 2.4.1), we have that

$$\int_{x\in\mathbb{R}^2} f(x)h(x\cdot\omega(\varphi))dx = \int_{\mathbb{R}} \mathcal{R}f(s,\varphi)h(s)ds,$$

where $\omega(\varphi) = (\cos\varphi, \sin\varphi) \in \mathbb{S}^1$ and $h \in L^\infty(\mathbb{R})$. By taking $h(s) = e^{-is\xi}$, we finally obtain

$$\hat{f}(\xi\omega(\varphi)) = \int_{x\in\mathbb{R}^2} f(x)e^{-i(x\cdot\omega(\varphi))\xi}dx = \int_{\mathbb{R}} \mathcal{R}f(s,\varphi)e^{-is\xi}ds = \frac{1}{(2\pi)^{1/2}}\mathcal{F}_s\mathcal{R}f(\xi,\varphi).$$

□

By making use of Theorem 2.4.2, we get a special Fourier representation of the Radon transform given by

$$\begin{aligned}\mathcal{R}f(s,\varphi) &= \frac{1}{(2\pi)^{1/2}}\int_{\xi\in\mathbb{R}} e^{is\xi}\mathcal{F}_s(\mathcal{R}f)(\xi,\varphi)d\xi \\ &= \int_{\mathbb{R}} e^{is\xi}\hat{f}(\xi\omega(\varphi))d\xi \qquad (2.4.1)\\ &= \frac{1}{2\pi}\int_{\xi\in\mathbb{R}}\int_{x\in\mathbb{R}^2} e^{i(s-(x\cdot\omega(\varphi)))\xi}f(x)dxd\xi,\end{aligned}$$

where $\omega(\varphi) = (\cos\varphi, \sin\varphi)$. The expression in Equation (2.4.1) looks like a pseudodifferential operator, except that the exponential term has the phase function $\phi((s,\varphi),x,\xi) := (s-(x\cdot\omega(\varphi)))\xi$ as argument. In addition, f and $\mathcal{R}f$ are defined over different domains. Similar to the case of a pseudodifferential operator, if we impose certain estimates on the phase function ϕ, we are able to describe the way the operator acts on the singularities of f. These operators are known as *Fourier integral operators*, and we have the necessary concepts to introduce them.

Definition 2.4.3 (Phase function, [70, Definition 6]). *Let $\Omega, \Xi \subset \mathbb{R}^2$ be open subsets. A real valued functions $\phi \in C^\infty(\Xi\times\Omega\times\mathbb{R}^2\setminus\{0\})$ is called a* phase function *if*

1. *ϕ is positive homogeneous of degree 1 in ξ; that is, $\phi(y,x,r\xi) = r\phi(y,x,\xi)$ for all $r > 0$.*
2. *$(\partial_y\phi, \partial_\xi\phi)$ and $(\partial_x\phi, \partial_\xi\phi)$ do not vanish for all $(y,x,\xi) \in \Xi\times\Omega\times\mathbb{R}^2\setminus\{0\}$.*

Definition 2.4.4 (Fourier Integral Operator, [70, Definition 7]). *Let $\Omega, \Xi \subset \mathbb{R}^2$ be open subsets. A* Fourier integral operator *(FIO) $\mathcal{P} : \mathcal{S}(\Omega) \longrightarrow \mathcal{S}(\Xi)$ is an operator of the form*

$$\mathcal{P}f(y) = \int_{\xi\in\mathbb{R}^2\setminus\{0\}}\int_{x\in X} e^{i\phi(y,x,\xi)}p(y,x,\xi)f(x)dxd\xi \quad \textit{for all } y\in\Xi \textit{ and } f\in\mathcal{S}(\Omega),$$

where ϕ is a phase function that follows the estimates of Definition 2.4.3. In addition, $p \in C^\infty(\Xi\times\Omega\times\mathbb{R}^2\setminus\{0\})$ is an amplitude function. Following that, for all compact

$K \subset \Xi \times \Omega$ and for all multi-index α, β, γ, there is a constant $C = C(K, \alpha, \beta, \gamma)$ such that

$$|D_\xi^\alpha D_x^\beta D_y^\gamma p(y, x, \xi)| \leq C(1 + ||\xi||)^{m-|\alpha|} \textit{ for all } x, y \in K \textit{ and } \xi \in \mathbb{R}^2 \setminus \{0\},$$

for some $m \in \mathbb{R}$, the degree of the operator.

We will also define the operator $\mathcal{P} : \mathcal{S}'(\Omega) \to \mathcal{S}'(\Xi)$ using the extension by duality, that is, for $u \in \mathcal{S}'(\Omega)$

$$\mathcal{P}(u)(g) := u(\mathcal{P}^* g) \quad \text{for all } g \in \mathcal{S}(\Xi),$$

where P^* is the adjoint of $\mathcal{P}$. In the case of a Fourier integral operator, since the domain and co-domain are different, the transformation formula for the wavefront set does not have the trivial form as for the pseudodifferential operator. In this case, the phase function spreads out the singularities. We will define the *microcanonical relation* that considers the spreading of singularities by ϕ.

Definition 2.4.5 (Microcanonical relation)**.** *Let $\mathcal{P} : \mathcal{S}'(\Omega) \to \mathcal{S}'(\Xi)$ be a Fourier integral operator with a phase $\phi \in C^\infty(\Xi \times \Omega \times \mathbb{R}^2 \setminus \{0\})$. Let us define*

$$\Sigma_\phi := \{(y, x, \xi) \in \Xi \times \Omega \times \mathbb{R}^2 \setminus \{0\} | \partial_\xi \phi(y, x, \xi) = 0\}, \tag{2.4.2}$$

and

$$C_\phi := \{\big((y; \partial_y \phi(y, x, \xi)\big), \big(x; -\partial_x \phi(y, x, \xi))\big) | (y, x, \xi) \in \Sigma_\phi\}. \tag{2.4.3}$$

We call the set C_ϕ the microcanonical relation *of ϕ.*

The microcanonical relation helps us to find the singularities of $\mathcal{P}f$ from the singularities of f. We will present this result in the next theorem, originally introduced by Hörmander [63].

Theorem 2.4.6 ([63])**.** *Let $\mathcal{P} : \mathcal{S}'(\Omega) \to \mathcal{S}'(\Xi)$ be a Fourier integral operator with an associated microcanonical relation C_ϕ. Then for $f \in \mathcal{S}'(\Omega)$ we have that*

$$WF(\mathcal{P}f) \subset C_\phi \circ WF(f), \tag{2.4.4}$$

where

$$C_\phi \circ WF(f) := \{(y; \mu) | \exists (x; \lambda) \in WF(f) \textit{ with } ((y; \mu), (x; \lambda)) \in C_\phi\}.$$

The microcanonical relations allow us to find the wavefront set of a function f from the measurements $\mathcal{P}f$. Such a subset of the wavefront set is contained, according to (2.4.4), in $C_\phi \circ \mathrm{WF}(f)$, where C_ϕ is the microcanonical relation of $\mathcal{P}$ (Definition 2.4.5). In many cases, the forward operator modelling real-world inverse problems in imaging are in fact Fourier integral operators. This fact makes it possible to use the microcanonical relation in order to compute a subset of the wavefront set of the target function without the need of a reconstruction. Being the main example to motivate the definition of pseudodifferential and Fourier integral operator, the Radon transform is generally the standard example for the use of the microcanonical relation.

Researchers use the microcanonical relation to find singularities in images that can be faithfully reconstructed from their Radon transform. We will explore this in more detail in the next section.

2.5 Microlocal analysis of the Radon transform

In this section we will explore in depth the microlocal properties of the Radon transform. In Section 2.4 we showed that the Radon transform can be represented in the integral form

$$\mathcal{R}f(s,\varphi) = \frac{1}{2\pi}\int_{\xi\in\mathbb{R}}\int_{x\in\mathbb{R}^2} e^{i(s-(x\cdot\omega(\varphi)))\xi} f(x)dxd\xi. \tag{2.5.1}$$

This implies that $\mathcal{R}$ is a Fourier integral operator with phase function $\phi((s,\varphi),x,\xi) = (s-(x\cdot\omega(\varphi)))\xi$, with $\omega(\varphi) = (\cos\varphi, \sin\varphi)$ and amplitude $p(y,x,\xi) = \frac{1}{2\pi}$. Using Definition 2.4.5 we can find the microcanonical relation of $\mathcal{R}$ given by

$$\mathrm{WF}(\mathcal{R}f) \subset C_\phi \circ \mathrm{WF}(f),$$

where

$$\Sigma_\phi := \{(y,x,\xi) \in Y\times X\times\mathbb{R}^2\setminus\{0\} | \partial_\xi\phi(y,x,\xi) = 0\},$$

and

$$C_\phi := \{((y;\partial_y\phi(y,x,\xi)),(x;-\partial_x\phi(y,x,\xi))) : (y,x,\xi)\in\Sigma_\phi\}.$$

In the case of $\mathcal{R}$, we have $\phi((s,\varphi),x,\xi) = \xi(s - x\cdot\omega(\varphi))$. Let us compute the exact form of C_ϕ by computing the derivatives of ϕ:

$$\begin{aligned}\partial_x\phi((s,\varphi),x,\xi) &= -\xi\omega(\varphi),\\ \partial_{(s,\varphi)}\phi((s,\varphi),x,\xi) &= \xi(1,-x\cdot\omega^\perp(\varphi)),\\ \partial_\xi\phi((s,\varphi),x,\xi) &= (p - x\cdot\omega(\varphi),0),\end{aligned} \tag{2.5.2}$$

where $\cdot : \mathbb{R}^2\times\mathbb{R}^2\to\mathbb{R}$ is the dot product. Notice that $\partial_x\phi$ and $\partial_{(s,\varphi)}\phi$ are not zero for $\xi\neq 0$, which means that the phase function is non-degenerate. This confirms that $\mathcal{R}$ is a Fourier integral operator. Since the amplitude function of $\mathcal{R}$, $p((s,\varphi),x,\xi) = 1/2\pi$, is homogeneous of degree zero, and has order $m = -1/2$, this also implies that $\mathcal{R}$ is elliptic. Using the derivatives (2.5.2), Σ_ϕ is given by:

$$\Sigma_\phi = \{((s,\varphi),x,\xi)\in(\mathcal{R}\times[0,2\pi))\times\mathbb{R}^2\times\mathbb{R}\setminus\{0\} | s - x\cdot\omega(\varphi) = 0\}.$$

Therefore, the microcanonical relation can be represented by the mapping:

$$\begin{aligned}\big((s,\varphi),x,\xi\big) &\mapsto \big((((x\cdot\omega(\varphi),\varphi);\partial_{(s,\varphi)\phi}),(x;-\partial_x\varphi)\big)\\ &= \big((((x\cdot\omega(\varphi),\varphi);\xi(1,-x\cdot\omega^\perp(\varphi))),(x;\xi\omega(\varphi))\big).\end{aligned} \tag{2.5.3}$$

We can then characterize the propagation of singularities of $\mathcal{R}$.

Theorem 2.5.1 (Propagation of singularities of $\mathcal{R}$, [91, Theorem A.6])**.** *Let $f \in \mathcal{S}(\Omega)$, then*

a) Let $(x_0; \lambda_0) \in \Omega \times \mathbb{S}^1$, and let $\varphi_0 \in (0, \pi)$ such that $\lambda_0 = \omega(\varphi_0) = (\cos\varphi_0, \sin\varphi_0)$. If $(x_0; \lambda_0) \in \mathrm{WF}(f)$, then $((x_0 \cdot \lambda_0, \varphi_0); (1, -x_0 \cdot \lambda_0^{\perp})) \in \mathrm{WF}(\mathcal{R}f)$.

b) Let $(s_0, \varphi_0) \in \mathbb{R}\times(0, \pi)$ and assume $((s_0, \varphi_0); (1, -A)) \in \mathrm{WF}(\mathcal{R}f)$. Then, $(s_0\omega(\varphi_0)+ A\omega^{\perp}(\varphi_0); \omega(\varphi_0)) \in \mathrm{WF}(f)$.

Proof. Since $\mathcal{R}$ is an elliptic Fourier integral operator, we have

$$\mathrm{WF}(\mathcal{R}f) = C_\phi \circ \mathrm{WF}(f), \tag{2.5.4}$$

where C_ϕ is given by Definition 2.4.5. Therefore

$$\mathrm{WF}(\mathcal{R}f) = \{((s, \varphi); \mu)|\exists(x; \lambda) \in WF(f) \text{ with } (((s, \varphi); \mu), (x, \xi)) \in C_\phi\}$$

Using Equation (2.5.3) we finally get the results in part a) and b). □

Part a) in Theorem 2.5.1 implies that $\mathcal{R}$ detects singularities perpendicular to the line of integration, since by definition $\lambda_0 = \omega(\varphi_0)$ is perpendicular to the line $L(x_0 \cdot \omega(\varphi_0), \omega(\varphi_0))$. Part b) implies that if $(s_0, \varphi_0) \in \operatorname{sing supp}(\mathcal{R}f)$ it must come from a singular directed point of f in the line $L(s_0, \omega(\varphi_0))$ and in a direction perpendicular to such line.

For a more convenient representation, the microcanonical relation for the Radon transform in Definition 2.3.1 at tempered distribution $f \in \mathcal{S}'(\mathbb{R}^2)$ is a precise relationship between $\mathrm{WF}(\mathcal{R}(f))$ and $\mathrm{WF}(f)$. If $\mathcal{P}$ denotes taking the power set, then this can be expressed as a map

$$K\colon \mathcal{P}\Big(\big(\mathbb{R} \times (0, \pi)\big) \times \mathbb{S}^1\Big) \to \mathcal{P}(\mathbb{R}^2 \times \mathbb{S}^1), \tag{2.5.5}$$

where $K\,\mathrm{WF}((\mathcal{R}(f))) = \mathrm{WF}(f)$ for $f \in \mathcal{S}'(\mathbb{R}^2)$. In limited-angle tomography we only have access to the Radon transform on an open subset $\Xi \subset \mathbb{R} \times (0, \pi)$. The microcanonical relation then holds for the so-called *visible* wavefront set of a function/distribution f that is given by

$$\mathrm{WF}^{\mathrm{vis}}(f) := \mathrm{WF}(f) \cap K(\Xi). \tag{2.5.6}$$

Following [70], we next provide a more precise characterization of the microcanonical relation in terms of a mapping between wavefront sets in image and sinogram, respectively.

Theorem 2.5.2. *The microcanonical relation for the Radon transform $\mathcal{R}\colon \mathcal{S}'(\mathbb{R}^2) \to \mathcal{S}'\big(\mathbb{R} \times (0, \pi)\big)$ at $f \in \mathcal{S}'(\mathbb{R}^2)$ can be represented by the mapping*

$$\mathrm{Can}_{\mathcal{R}(f)}\colon\ \mathrm{WF}(f) \to \mathrm{WF}\big(\mathcal{R}(f)\big)$$

defined as

$$\mathrm{Can}_{\mathcal{R}(f)}\big(x; \omega(\theta)\big) := \Big(\big(x\cdot\omega^{\perp}(\theta), \theta+\pi/2\big); \omega\Big(\arctan\big(-x\cdot\omega(\theta)\big)\Big)\Big) \quad \textit{for } \big(x; \omega(\theta)\big) \in \mathrm{WF}(f), \tag{2.5.7}$$

with $\omega(\theta) := (\cos\theta, \sin\theta)$ *and* $\omega(\theta)^{\perp} := (-\sin\theta, \cos\theta)$. *This means,*

$$\big(x;\omega(\theta)\big) \in \mathrm{WF}(f) \iff \mathrm{Can}_{\mathcal{R}(f)}\big(x;\omega(\theta)\big) \in \mathrm{WF}\big(\mathcal{R}(f)\big).$$

Proof. The Radon transform in (2.3.1) is a Fourier integral operator with phase function $\phi((s,\theta),x,\xi) := \big(s - \big(x\cdot\omega(\theta)\big)\big)\xi$ and amplitude $p(y,x,\xi) := 1/(2\pi)$. Thus, the microcanonical relation of $\mathcal{R} : \mathcal{S}'(\mathbb{R}^2) \to \mathcal{S}'(\mathbb{R}\times(0,\pi))$ is given by (2.5.4).

We then derive the exact form of C_ϕ by computing the derivatives of ϕ, which are

$$\begin{aligned} \partial_x\phi\big((s,\theta),x,\xi\big) &= -\xi\omega(\theta), \\ \partial_{(s,\theta)}\phi\big((s,\theta),x,\xi\big) &= \xi\big(1, -x\cdot\omega^{\perp}(\theta)\big), \\ \partial_\xi\phi\big((s,\theta),x,\xi\big) &= \big(s - x\cdot\omega(\theta), 0\big). \end{aligned} \tag{2.5.8}$$

Notice that $\partial_x\phi$ and $\partial_{(s,\theta)}\phi$ are not zero for $\xi \neq 0$, which means that the phase function is non-degenerate. This confirms that $\mathcal{R}$ is a Fourier integral operator.

Next, $\mathcal{R}$ is of order $m = -1/2$ with an amplitude function $p((s,\theta),x,\xi) = 1/(2\pi)$ that is homogeneous of degree zero, implying that $\mathcal{R}$ is elliptic. Using the derivatives (2.5.8), Σ_ϕ is given by

$$\Sigma_\phi = \Big\{\big((s,\theta),x,\xi\big) \in \big(\mathbb{R}\times[0,2\pi)\big)\times\mathbb{R}^2\times\mathbb{R}\setminus\{0\} \colon s - x\cdot\omega(\theta) = 0\Big\}.$$

Therefore, the microcanonical relation can be represented by the coordinate mapping

$$\begin{aligned} \big((s,\theta),x,\xi\big) &\mapsto \Big(\Big(\big(x\cdot\omega(\theta),\theta\big);\partial_{(s,\theta)\phi}\Big),\big(x,-\partial_x\theta\big)\Big) \\ &= \Big(\Big(\big(x\cdot\omega(\theta),\theta\big);\xi\big(1,-x\cdot\omega(\theta)^{\perp}\big)\Big),\big(x,\xi\omega(\theta)\big)\Big). \end{aligned} \tag{2.5.9}$$

Now, let $\big(x;\omega(\theta)\big) \in \mathrm{WF}(f)$ be an oriented singular point of f. By (2.5.9), we obtain that

$$\big(x;\omega(\theta)\big) \in \mathrm{WF}(f) \Longrightarrow \Big(\big(x\cdot\omega(\theta)^{\perp},\theta+\pi/2\big);\omega\Big(\arctan\big(-x\cdot\omega(\theta)\big)\Big)\Big) \in \mathrm{WF}\big(\mathcal{R}(f)\big).$$

Finally, [91, Theorem 6.3] gives

$$\big(x;\omega(\theta)\big) \in \mathrm{WF}(f) \iff \Big(\big(x\cdot\omega(\theta)^{\perp},\theta+\pi/2\big);\omega\Big(\arctan\big(-x\cdot\omega(\theta)\big)\Big)\Big) \in \mathrm{WF}\big(\mathcal{R}(f)\big). \tag{2.5.10}$$

This concludes the proof. □

We next focus on the propagation of singularities performed by the adjoint of the Fréchet derivative of the Radon transform, which is the back-projections operator $\mathcal{R}^*$ in Definition 2.3.2. In the next proposition, we use (2.3.5) to introduce the corresponding mapping associated with the microcanonical relation for $\mathcal{R}^*$ in a similar fashion to Theorem 2.5.2.

Proposition 2.5.3. *The microcanonical relation for the back-projection operator* $\mathcal{R}^*\colon \mathcal{S}'\big(\mathbb{R}\times(0,\pi)\big)\to\mathcal{S}'(\mathbb{R}^2)$ *in* (2.3.2) *at* $g\in\mathcal{S}'\big(\mathbb{R}\times(0,\pi)\big)$ *can be represented by the mapping*

$$\mathrm{Can}_{\mathcal{R}^*(g)}\colon\ \mathrm{WF}(g)\to\mathrm{WF}\big(\mathcal{R}^*(g)\big),$$

which is defined at $\big((s,\theta);\omega(\vartheta)\big)\in\mathrm{WF}(g)$ *as*

$$\mathrm{Can}_{\mathcal{R}^*(g)}\big((s,\theta);\omega(\vartheta)\big):=\big((s\cos\theta-\tan\vartheta\sin\theta, s\sin\theta+\tan\vartheta\cos\theta);\theta-\pi/2\big),\quad (2.5.11)$$

where $\omega(\theta):=(\cos\theta,\sin\theta)$*. This means,*

$$\big((s,\theta);\omega(\vartheta)\big)\in\mathrm{WF}(g)\Leftrightarrow\mathrm{Can}_{\mathcal{R}^*(g)}\big((s,\theta);\omega(\vartheta)\big)\in\mathrm{WF}\big(\mathcal{R}(f)\big).$$

Proof. Note first that (2.3.5) implies that the operator $\mathcal{R}^*\mathcal{R}\colon\mathcal{S}(\mathbb{R}^2)\to\mathcal{S}(\mathbb{R}^2)$ is an elliptic pseudodifferential operator with amplitude function $p(y,x,\xi):=1/\|\xi\|$. By duality we can extend this to a mapping $\mathcal{R}^*\mathcal{R}\colon\mathcal{S}'(\mathbb{R}^2)\to\mathcal{S}'(\mathbb{R}^2)$. In addition, the pseudolocal property of pseudodifferential operators (see Theorem 2.3.10) implies that $\mathcal{R}^*\mathcal{R}$ will preserve the wavefront set of functions in $\mathcal{S}(\mathbb{R}^2)$, i.e., $\mathrm{WF}(\mathcal{R}^*\mathcal{R}(f))=\mathrm{WF}(f)$. This allows us to represent the microcanonical relation for the inverse mapping in terms of the microcanonical relation mapping for $\mathcal{R}$. Finally, by inverting the mapping implicit in (2.5.10), for $g\in\mathcal{S}'\big(\mathbb{R}\times(0,\pi)\big)$, we obtain that

$$\big((s,\theta);\omega(\vartheta)\big)\in\mathrm{WF}(g)\Leftrightarrow\big((s\cos\theta-\tan\vartheta\sin\theta, s\sin\theta+\tan\vartheta\cos\theta);\theta-\frac{\pi}{2}\big)\in\mathrm{WF}\big(\mathcal{R}^*(g)\big).\quad (2.5.12)$$

□

Notice that every ΨDO is also a FIO with the phase function $\phi(y,x,\xi):=(y-x)\cdot\xi$. These types of operators appear in various imaging applications in the natural sciences, in particular, computed tomography (CT), magnetic resonance imaging (MRI), and electroencephalography (EEG). This suggests a great potential for microlocal analysis in such fields. The first thing that stops people to directly apply microlocal analysis theory to real-world problems, is the continuous nature of the theory. It is important to understand that singularities in signals formally exist just in the continuous setting. Indeed, the wavefront sets are defined based on asymptotic analysis of the Fourier transform.

In reality, we have just access to a finite number of Fourier coefficients, therefore, such asymptotic analysis will be technically impossible to perform on the computer. Another challenge that one will encounter trying to design an algorithm that resolves the wavefront set is related to the localization procedures presented in Definition 2.2.6. In short, one needs to localize the singular directed points in space and orientation, but such localization is unspecified by the definition. One way to alleviate the last challenge is by the introduction of multiscale directional systems. These systems, coming from applied harmonic analysis, are able to resolve the wavefront set by analyzing the asymptotic behavior of the corresponding coefficients for specific localization and orientation. Chapter 3 will introduce in detail the notion of multiscale directional systems, and in particular, shearlet systems. These systems allow us to resolve the wavefront set reliably. In addition, the faithful discretization of the shearlet transform allows us to implement a digital wavefront set extractor. This extractor is based on the digital shearlet transform and a deep convolutional neural network classifier. We will present this algorithm in Chapter 5.

3 Continuous wavefront set resolution via harmonic analysis

The area of harmonic analysis was originally introduced as the study of representations of functions or signals as the linear combination of basic waves. In some sense, it generalizes the notion of Fourier analysis. As part of this aim, in the last decades, applied harmonic analysis has developed multiscale systems for efficient representations as well as the analysis of regularity and detection of features such as singularities. This approach is motivated by a paradigm shift, recently observed in distinct scientific disciplines, namely, sparse approximation. The novel paradigm of sparse approximations has enabled the highly efficient encoding of signals, and various of new image processing methodologies, such as missing data recovery and morphological separation, see [45]. We will discuss these methodologies in Sections 3.2, 3.3 and 3.4, but first we will motivate the notion of wavefront set extraction in the context of real-world applications.

3.1 Motivation

In many scientific and industrial real-world applications, one requires a precise understanding how model parameters, are transformed under some measuring process that is described by an operator. This analysis is generally very challenging, and one attempt to simplify it consists of treating the singular (non-smooth) and smooth parts of the function separately. In fact, as already discussed in Chapter 2, a significant portion of the useful information is often contained in the singular part. In our particular case, for images, the singular part can be understood as edges, ridges, or ramps in the image. In this context, microlocal analysis is very useful, since it aims to precisely describe how the singular part of a function, or more generally a distribution, is transformed under the action of an operator. The most important observation in microlocal analysis is that information about the location of the singularities (singular support) needs to be complemented by directional information, known as microlocalization.

In Section 2.4, we discussed how this extra ("microlocal") information is key in understanding the propagation of singularities by Fourier integral operators, including differential and pseudodifferential operators, as well as many integral operators arising in integral geometry. Such operators are frequently encountered in scientific computing, physical science, and many biomedical imaging problems such as X-ray tomography [61, 20]. Furthermore, microlocal analysis is also particularly useful in inverse problems, where the goal is to reliably recover a hidden model parameter (signal) from a noisy transformed version. The goal here is to recover the wavefront set of the signal (image) given the noisy realization of a transformed version of it. Such applications frequently arise when using imaging/sensing technologies, where the transform is a Fourier integral operator [70].

The idea that edges are often sufficient to semantically understand an image is also based on neurophysiology. As presented by Field et al. in [42], the complex cells in the human visual cortex respond primarily to oriented edges and gratings, making humans very sensitive to oriented edge-like structures. More specifically, it has been discussed in [84] that the human visual cortex performs multiple operations of image processing, the first of which is rough sketching involving edge detection (for more information on the role of singularities in the visual cortex we refer to [113, 14, 15]). For the aforementioned and additional reasons, there exists a rich set of applications of microlocal analysis to tomographic imaging. In these applications, the transformation relating the image to data is the Radon transform (Definition 2.3.1) which can be regarded as a Fourier integral operator. We discussed Section 2.5 the microlocal analysis of this operator. Similar principles hold for transforms integrating along with other types of curves, for example, ellipses with foci on the x-axis and geodesics [115]. Another observation is that recovering a signal from its ray transform is less ill-posed if one knows the wavefront set a priori. This was demonstrated in [38], where the severely ill-posed reconstruction problem in limited angle tomography becomes mildly ill-posed if the wavefront set of the solution is provided as prior information. We refer to [91] for an application of this principle to cryo-electron tomography.

3.2 Multiscale systems

Multiscale systems are formed by functions resulting from transformations applied to a set of generating functions in a specific space. These transformations need to include a scaling action, leading to different resolutions. The scaling of a generating function enables the system to extract features of different sizes, allowing it to detect persistent properties along scales, including singularities, or edges, in multiple dimensions. The first celebrated multiscale system was the *wavelet system* [81]. The wavelet system is the best-known example of an isotropic multiscale system, and we will discuss it in detail in the next section.

Although we study first multiscale systems defined on $L^2(\mathbb{R}^2)$, later we present the main results on wavefront set resolution for tempered distributions $\mathcal{S}'(\Omega)$.

3.2.1 Isotropic multiscale systems

Let us first introduce the definition of the 2D wavelet transform [30].

Definition 3.2.1 (Two-dimensional wavelet transform). *Let $\psi \in L^2(\mathbb{R})$ be a function, we refer to ψ as a* wavelet function *if it follows*

$$C_\psi := \int_{\xi \in \mathbb{R}^2} \frac{|\hat{\psi}(\xi)|^2}{|\xi|} d\xi < \infty,$$

where $\hat{\psi}$ is the Fourier transform of ψ (see Section 2.1) . Then, the associate two-dimensional wavelet transform *of $f \in L^2(\mathbb{R}^2)$ is given by*

$$\mathcal{W}_\psi f(a,b) := \int_{\mathbb{R}^2} f(x_1,x_2)\psi_{a,b}(x_1,x_2)dx,$$

where $a > 0$ and $b \in \mathbb{R}^2$, and $\psi_{a,b}$ is defined as

$$\psi_{a,b}(x_1,x_2) := a^{-1/2}\psi(ax_1 - b_1, ax_2 - b_2).$$

The function $\psi_{a,b}$ is also refered to as wavelets.

The main reasons for the success of wavelets is their ability to optimally approximate 1D signals [81]. They are able to represent singularities much more efficiently than Fourier methods. the wavelet transform also have fast algorithmic implementations which precisely digitize the continuous domain transform. In addition, the wavelet system has a rich mathematical structure, which allows one to design families of wavelets with various desirable properties expressed in terms of regularity, decay, and vanishing moments. These properties set off a revolution in image and signal processing, including the introduction of the algorithm JPEG2000, a commonly used algorithm for image compression.

Although wavelets have shown to be optimal for extraction of point-wise singularities in 1D functions, in more than one dimension, representing functions solely by scaling and translation is not enough to efficiently represent singularities along curves. In the case of 2D, one needs to encode not just the position of the singularities, but also the orientation, i.e., the wavefront set. Because of this, wavelets are not able to sparsely represent 2D functions, like images. In fact, wavelets are deficient in describing edge singularities due to their isotropic nature. This was formally proven in 2004 by Candès and Donoho [22]. At the same time, they introduced a new multiscale system called *curvelets* to overcome such limitations. The curvelet system uses parabolic scaling and rotation to efficiently represent curvilinear singularities. More generally, systems that make use of transformations to change the orientation of a function are known as geometric multiscale systems.

3.2.2 Multiscale directional systems

As discussed above, the curvelet system was introduced by Candès and Donoho for overcoming the lack of directional representation of the already existing multiresolution systems, such as wavelets. The curvelet system can be regarded as a breakthrough in the *optimal representation* of such *curvilinear singularities*, formalized with the concept of *cartoon-like functions.*

The class of *cartoon-like functions* is of special interest in imaging sciences. It was introduced by Donoho in [34] to provide a simplified model of natural images, which emphasizes anisotropic features, such as edges, and is consistent with many models of the human visual system. A natural image basically consists of smooth regions separated by edges. Based on this, one could consider piecewise regular functions. An example of this is depicted in Figure 3.1. This concept is formalized by Defintion 3.2.2

Definition 3.2.2 (Cartoon-like images class [73, Definition 1])**.** *The class $\mathcal{E}^2(\mathbb{R}^2)$ of* cartoon-like images *is the set of functions $f : \mathbb{R}^2 \to \mathbb{C}$ of the form*

$$f = f_0 + f_1\chi_B,$$

where $B \subset [0,1]^2$ is a set with boundary ∂B a closed C^2-curve with bounded curvatures and $f_i \in C^2(\mathbb{R}^2)$ (two times continuously differentiable) are functions with $\operatorname{supp}(f_i) \subset [0,1]^2$ *and $||f_i||_{C^2} \leq 1$ for each $i = 0, 1$. In this context the C^2-norm $||\cdot||_2$ is given by*

$$||f||_{C^2} = \max_x(|f(x)| + |f'(x)| + |f''(x)|)$$

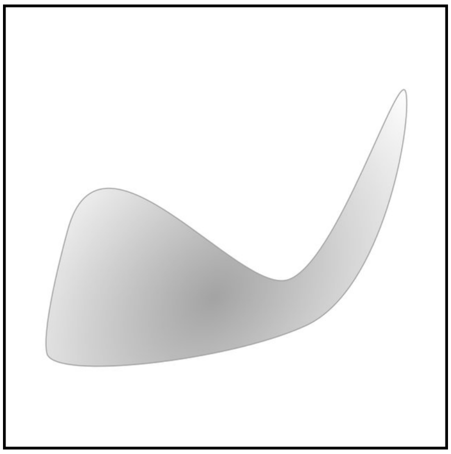

Figure 3.1: Visual representation of a cartoon-like function.

Before we can talk about the approximation power of wavelet and curvelet systems, it is important to discuss the notion of a frame. Frame theory was introduced due to the need for redundant systems in functional analysis. It was originally developed by Duffin and Schaeffer in [35], with the intention of having stability even when the decomposition is nonunique. A general definition is presented below.

Definition 3.2.3. *Let $\mathcal{H}$ be a Hilbert space, a sequence $\{\psi_i\}_{i\in I} \subset \mathcal{H}$ is called a* frame *for $\mathcal{H}$, if there exist constants $0 < A \leq B < \infty$ such that*

$$A||f||^2 \leq \sum_{i\in I} |\langle f, \psi_i\rangle|^2 \leq B||f||^2 \quad \textit{for all } f \in \mathcal{H}.$$

The constants A and B are called lower *and* upper frame bounds*. Furthermore, if A and B can be chosen to be equal, we call the frame a $(A-)$* tight frame*. If the choice $A = B = 1$ is possible, $\{\psi_i\}_{i\in I}$ is called a* Parseval frame*.*

An important consequence of the above definition is that a frame in a Hilbert space $\mathcal{H}$ spans $\mathcal{H}$, i.e. $\overline{\operatorname{span}\{\psi_i : i \in I\}} = \mathcal{H}$. According to this, one can characterize the approximation power of a frame and even a general collection of vectors over the Hilbert space with the concept of *best $N-$term approximation.*

Definition 3.2.4. *Let $\mathcal{H}$ be a Hilbert space and $\Psi = \{\psi_i\}_{i\in I} \subset \mathcal{H}$ be a collection of vectors, e.g., a frame. The set*

$$\Sigma_N(\Psi) := \left\{ \sum_{i\in I} c_i\psi_i : \#\{i : c_i \neq 0\} \leq N \right\} \subseteq \mathcal{H}$$

is called the non-linear N-term approximation manifold. *Then, the* best N-term approximation error *of $f \in \mathcal{H}$ is defined as*

$$\sigma_N(f, \Psi) := \inf_{g\in\Sigma_N(\Psi)} ||f - g||.$$

In the realm of sparse representation, the optimal N-term approximation in the sense that minimizes $\sigma_N(f, \Psi)$ for $f \in \mathcal{H}$ plays an important role. In this case, optimality means that $\sigma_N(f, \Psi)$ is as small as possible. For cartoon-like functions $\mathcal{E}^2(\mathbb{R}^2)$, one is able to find the explicit optimal approximation rate, i.e., the rate of decay of $\sigma_N(f, \Psi)$ with respect to N, as expressed in the next theorem.

Theorem 3.2.5 ([34])**.** *Let $\{\psi_i\}_{i\in I} \subset L^2(\mathbb{R}^2)$ be a frame for $L^2(\mathbb{R}^2)$. Then the optimal best $N-$term approximation rate for any $f \in \mathcal{E}^2(\mathbb{R}^2)$ is*

$$\sigma_N(f, \{\psi_i\}_i) = O(N^{-1}), \quad \textit{as } N \to \infty.$$

The proof of this result is quite long and technical. To not distract our main goal, we will refer to [34] for the reader with an interest in a detailed explanation.

We can now present a heuristic argument, which highlights the limitations of traditional wavelet systems with respect to more sophisticated multiscale directional systems when aiming to sparsely approximate cartoon-like images optimally. For this, let $f \in \mathcal{E}^2(\mathbb{R}^2)$ be a cartoon-like image containing a singularity along a smooth curve, and $\{\psi_{a_j,b}\} \subset L^2(\mathbb{R}^2)$ the two dimensional wavelet system (Definition 3.2.1). For scale $a_j = 2^{-j}$ ($j \in \mathbb{N}$) sufficiently small, the only significant wavelet coefficients $\langle f, \psi_{a_j,b}\rangle$ are those located near the singularity. Since at scale $a_j = 2^{-j}$, each wavelet $\psi_{a_j,b}$ is supported, or essentially supported, inside a box of size $2^{-j} \times 2^{-j}$, there exist about 2^j elements of the wavelet basis overlapping the singularity curve. The associated wavelet coefficients can be controlled by

$$|\langle f, \psi_{a_j,b}\rangle| \leq ||f||_\infty ||\psi_{a_j,b}||_{L^1} \leq C2^{-j},$$

where $||f||_\infty = \max_{x\in\mathbb{R}^2} |f(x)|$. It follows that the N^{th} largest wavelet coefficients in magnitude, which we denote by $\langle f, \psi_{j,b}\rangle_{(N)}$, is bounded by $O(N^{-1})$. Thus, if f is approximated by its best $N-$term approximation f_N, the L^2 error obeys

$$||f - f_N||^2_{L^2} \leq \sum_{\ell\geq N} |\langle f, \psi_{a_j,b}\rangle_{(\ell)}|^2 \leq CN^{-1}.$$

Therefore,

$$\psi_N(f, \{\psi_{a_j,b}\}) = O(N^{-1/2}), \quad \text{as } N \to \infty. \tag{3.2.1}$$

This approximation rate is clearly slower than the optimal approximation rate of the cartoon-like functions given by Theorem 3.2.5.

In order to gain more intuition on the reason for the poor representation power of wavelet systems in two dimensions, we will introduce an alternative form of Theorem 3.2.5.

Theorem 3.2.6 ([73]). *Let $f \in \mathcal{E}^2(\mathbb{R}^2)$ be a cartoon-like image. There exists a constant C such that, for any N, a triangulation of $[0,1]^2$ with N triangles can be constructed so that the piecewise linear interpolation -according to the triangulation- f_N of these triangles satisfies*

$$||f - f_N||_{L^2} \leq CN^{-1}, \quad N \to \infty.$$

The proof of this theorem, given by Donoho [34], makes extensive use of adapted triangulations. This suggests that analyzing elements with elongated and oriented supports are useful to achieve optimally sparse approximations of piece-wise smooth functions in 2D. Indeed, the isotropic scaling of wavelets is the main reason for its suboptimal approximation rate on cartoon-like images, being suboptimal for representing elongated, oriented singularities. In addition, one also needs to be able to modify the orientation of the support. Figure 3.2 shows this fact.

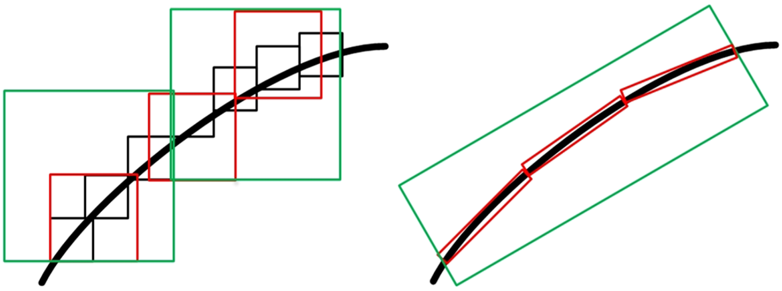

Figure 3.2: Elongated, and oriented singularities representations. Left: Using isotropic scaling. Right: Using anisotropic scaling and orientation.

The need for anisotropic scaling and orientation sensitivity is at the core of the construction of curvelets. This system is explicitly defined in the following.

Definition 3.2.7 (Continuous curvelet transform, [21]). *Let us denote by $x = (x_1, x_2)^\intercal$ the spatial variable, and $\xi = (\xi_1, \xi_2)^\intercal$ the variable in frequency domain. Further, let $r = \sqrt{\xi_1^2 + \xi_2^2}$, $\omega = \arctan(\xi_1/\xi_2)$ be the polar coordinates in frequency domain.*

Moreover, let $V, W \in C_c^\infty(\mathbb{R}^2)$ be smooth window functions with compact support *satisfying the admissibility conditions*

$$\begin{aligned} \sum_{l=-\infty}^{\infty} V^2(t-l) = 1, \quad t \in \mathbb{R}, \\ \sum_{j=-\infty}^{\infty} W^2(2^j r) = 1, \quad r > 0. \end{aligned} \tag{3.2.2}$$

In addition, let $(a, b, \theta) \in (0, 1] \times \mathbb{R}^2 \times [0, 2\pi)$ *be the set of parameters, representing scale, location, and orientation, respectively. Using the polar coordinates* (r, ω) *in frequency domain, we now define the* $a-$*scaled window*

$$U_a(r, \omega) := a^{3/4} W(ar) V\left(\frac{\omega}{\sqrt{a}}\right).$$

The window U_a *is then applied for building curvelet functions as follows. Let* $\varphi_{a,0,0} \in L^2(\mathbb{R}^2)$ *be defined by its Fourier transform*

$$\hat{\psi}_{a,0,0}(\xi) := U_a(\xi).$$

Finally, the curvelet system *is generated by translation and rotation of the basic element* $\psi_{a,0,0}$ *as*

$$\psi_{a,b,\theta} := \psi_{a,0,0}(R_\theta(x-b)), \tag{3.2.3}$$

with translation $b \in \mathbb{R}^2$*, and the* 2×2 *rotation matrix* R_θ *with orientation angle* $\theta \in [0, 2\pi)$ *is given by*

$$R_\theta = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}. \tag{3.2.4}$$

The curvelet transform *of a function* $f \in L^2(\mathbb{R}^2)$ *is then defined by*

$$\mathcal{CL}_\psi f(a, b, \theta) = \int_{\mathbb{R}^2} f(x_1, x_2) \psi_{a,b,\theta}(x_1, x_2) dx. \tag{3.2.5}$$

The curvelets in Equation (3.2.5) attain wedge-shaped oriented support on the Fourier domain depicted in Figure 3.3.

It was shown in [21] that curvelets attain the optimal best $N-$term approximation rate of Theorem 3.2.5 up to some log factor, confirming that the parabolic scaling and orientation sensibility was the key to sparsely represent cartoon-like images.

This type of sparse representation typically aimed to be used in real-world applications, such as denoising or inpainting, but curvelets have no unified continuous-to-digital theory. In general terms, the rotation matrix is hard to faithfully digitize [19]. It is in fact easy to see that the discrete grid is not rotation invariant, compare Figure 3.4. This flaw implies that the implementation would not be consistent with the theory for the continuous setting.

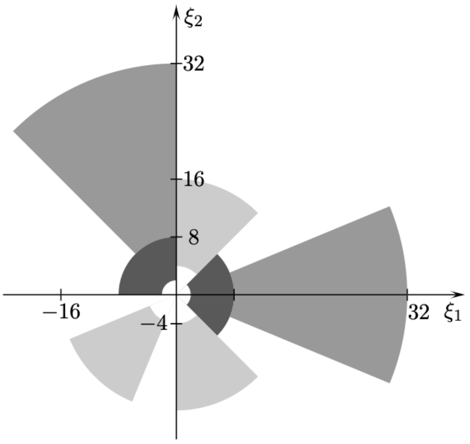

Figure 3.3: Support of different curvelets in the Fourier domain.

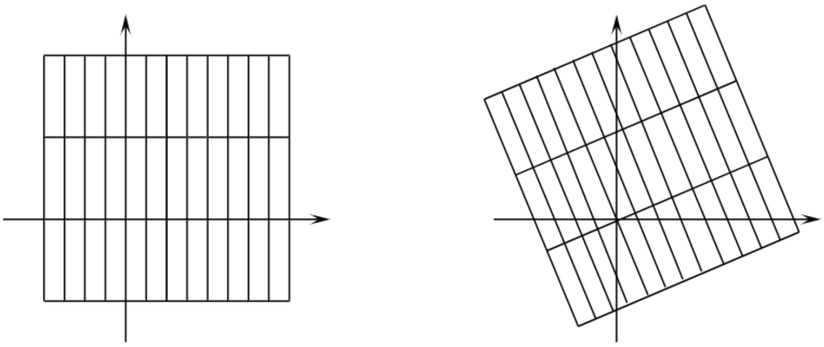

Figure 3.4: Rotation of the discrete grid.

To address the issue of faithful digitization, different new multiscale systems were introduced. One example are the *contourlets* proposed by Do and Vetterli in 2005 [32], which is basically a filter bank approach to the curvelet transform. This system has the main advantage of having faithful digitalization but still has some disadvantages with respect to wavelets. Both curvelets and contourlets do not have a unified treatment of the continuum and digital situation, and a theory for compactly supported systems to guarantee high spatial localization. This makes it very hard to compute approximation bounds in practical applications. Finally, as an attempt to overcome the main limitations of wavelets as well as curvelets, the *shearlet system* was developed by Kutyniok, Labate, and Guo in 2006 [50]. This will be introduced in the next section.

3.3 The continuous shearlet transform

The flaws of the wavelet system in representing 2D oriented curve-like singularities, as seen in the last section, lead to the slow decay rate of the best $N-$term approximation in Equation (3.2.1). The main problem of wavelets is that there are too many relevant wavelet coefficients, which worsen the decay rate significantly. One possibility to circumvent this problem is the use of parabolic scaling (see Figure 3.5), defined by

$$A_a := \begin{pmatrix} a & 0 \\ a & \sqrt{a} \end{pmatrix}.$$

In the particular case where $a = a_j := 2^{-j}$ $(j \in \mathbb{N})$, with $A_j := A_{a_j}$, one obtains elements of the type $\tilde{\psi}_{a_j,b}(x) = 2^{\frac{3j}{4}}\psi(A_j x - b)$. In Figure 3.6, it is visually clear that the use of such scaling reduces the number of relevant coefficients by a significant amount. Indeed, for each scale a_j, there exist only $O(2^{j/2})$ many coefficients intersecting the line singularity, leading to

$$|\langle f, \psi_{a_j,b}\rangle| \leq ||f||_\infty ||\psi_{a_j,b}||_{L^1} \lesssim 2^{\frac{-3j}{4}}.$$

On the one hand, this result implies that parabolic scaling allows more efficient approximation than its uniform counterpart. On the other hand, as one can see in Figure 3.5, parabolic scaling also produces anisotropic functions, in the sense that the shape of their (essential) supports are highly directional.

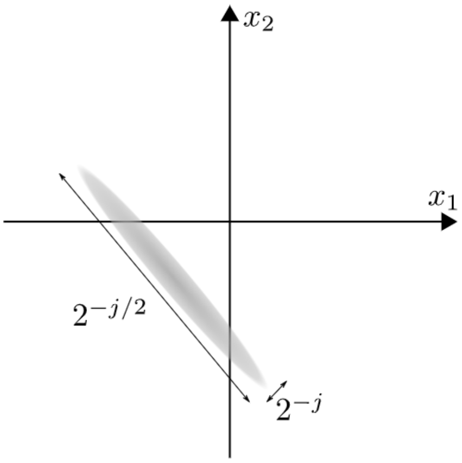

Figure 3.5: Parabolic scaling for $a = 2^j$

As we saw with curvelets that have highly anisotropic elements, for general directional systems, we require an operation that changes their orientation. We also know that the orientation changing operator used by curvelets, namely the rotation, cannot be faithfully digitized, creating a problem with the numerical implementation. In 2005, Kutyniok, Labate, and Guo [50] noticed that a much better choice is the shearing matrix.

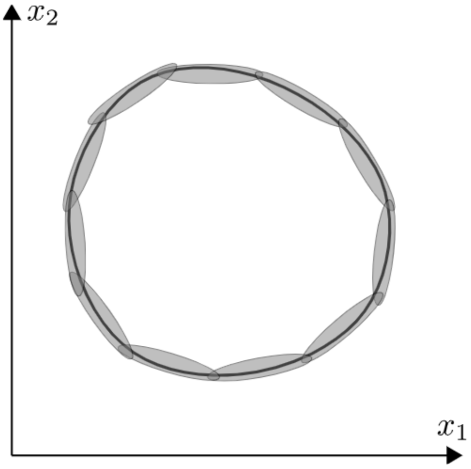

Figure 3.6: Efficient covering of a curved line singularity by anisotropic functions.

Unlike the rotation operator, the shear operator maintains consistent properties with the digital lattice. This makes the continuum and digital settings able to be treated uniformly, resulting in a faithful discretization, where most of the theoretical properties of the continuous shearlet transform are maintained in the discrete transform. The notion of shearing allows us to introduce the continuous shearlet system [50].

Definition 3.3.1 ([73])**.** *For $a > 0$, let*

$$A_a := \begin{pmatrix} a & 0 \\ 0 & \sqrt{a} \end{pmatrix}$$

be the parabolic scaling matrix*, and for $s \in \mathbb{R}$*

$$S_s := \begin{pmatrix} 1 & s \\ 0 & 1 \end{pmatrix}$$

be the shearing matrix*. Given $\psi \in L^2(\mathbb{R}^2)$, the* shearlet system *associated with ψ is defined as*

$$\mathcal{SH}_\psi := \{a^{\frac{3}{4}} \psi(S_s A_a x - b) : a > 0, s \in \mathbb{R}, b \in \mathbb{R}^2\}.$$

As in the case of wavelets and curvelets, the shearlet parameters have a direct geometric interpretation. The scale a represents the size of the elements to capture, the shearing s represents the orientation, and the location b represents the position of the elements.

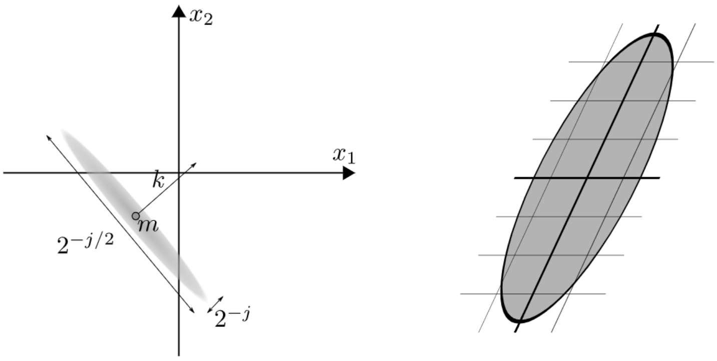

Figure 3.7: Shearlet system. Left: Geometry of shearlets with $a = 2^{-j}$, Right: The effect of shearing.

One can notice that in Definition 3.3.1 the function ψ referred to as *generating function*, is not explicitly defined. On the one hand, the choice of the specific generating function depends strongly on the properties the system shall have. On the other hand, there are weak assumptions on ψ that allow us to obtain a frame of $L^2(\mathbb{R}^2)$. In the following, we will introduce different forms to define the shearlet generating function with their own properties.

3.3.1 Classical shearlets

In general, for transformation associated to a shearlet system to be well-defined we require it to be an isometry up to isometric embeddings to $L^2(\mathbb{R}_+ \times \mathbb{R} \times \mathbb{R}^2)$, similar to the admissibility condition of wavelets (compare Definition 3.2.1).

Definition 3.3.2 (Admissible shearlets [48, Section 1])**.** *A function $\psi \in L^2(\mathbb{R}^2)$ is called an* admissible shearlet*, if*

$$C_\psi := \int_{\mathbb{R}^2} \frac{|\hat{\psi}(\xi)|}{|\xi_1|^2} d\xi = \int_{\mathbb{R}} \int_{\mathbb{R}_+} |\psi\widehat{(S_s A_a}\cdot)(\xi)|^2 a^{-3/2} dads < \infty. \tag{3.3.1}$$

One obtains equality of the second and third term in Equation (3.3.1) by using the substitution $\omega(a, s) = S_s A_a \xi$. The reason why the integral (3.3.1) converges is the structure of the shearlet group and the fact that the Haar measure is given by $a^{-3} dadsdb$. An extension of this notion of convergence is the concept of vanishing gradients:

Definition 3.3.3. *Let $f \in L^2(\mathbb{R}^2)$, we say that f has $n-$*vanishing moments in x_1-direction *if*

$$\int_{\mathbb{R}^2} \frac{|\hat{\psi}(\xi)|^2}{|\xi_1|} d\xi < \infty. \tag{3.3.2}$$

The term *vanishing moments* comes from the fact that if we assume sufficient spatial decay of ψ, Condition (3.3.2) is equivalent to

$$\int_{\mathbb{R}} x_1^k \psi(x_1, x_2) dx_1 = 0 \quad \text{for all} \quad x_2 \in \mathbb{R}, k < n.$$

Finally, Fourier analysis allows to prove that the shearlet system presents stable representation; in other words, it preserves the norms, analogous to the Plancherel's identity.

Theorem 3.3.4 ([48])**.** *If $\psi \in L^2(\mathbb{R}^2)$ is an admissible shearlet, then for all $f \in L^2(\mathbb{R}^2)$ we have the orthogonality relation*

$$||f||_2^2 = \int_{(a,s,b)\in\mathbb{R}_+\times\mathbb{R}\times\mathbb{R}^2} |\mathcal{SH}_\psi(f)(a,s,b)|^2 a^{-3} dadsdb,$$

where $\mathcal{SH}_\psi(f)(a,s,b) = \int_{\mathbb{R}^2} a^{-3/4}\psi(S_s A_a x - b) f(x) dx$.

Now we know shearlets can be regarded as stable representation systems. The first question one might ask is whether it is possible to sample a frame from this continuous system. The first shearlet generator introduced in [50] is the so-called classical shearlet, composed of a wavelet and a bump function. These functions are compactly supported in the Fourier domain, also known as band-limited functions. Choosing the system to be band-limited or compactly supported depends on the need to have high resolution in the spatial domain or in the Fourier domain. The uncertainty principle does not allow us to have both.

Definition 3.3.5 (Classical shearlets [50])**.** *Let $\psi \in L^2(\mathbb{R}^2)$ be defined by*

$$\hat{\psi}(\xi_1, \xi_2) = \hat{\psi}_1(\xi_1)\hat{\psi}_2(\xi_2/\xi_1),$$

where $\psi_1, \psi_2 \in L^2(\mathbb{R})$ satisfy the following properties:

(a) $\sum_{j\in\mathbb{Z}} |\hat{\psi}_1(2^{-j}\xi)|^2 = 1$ *for almost every $\xi \in \mathbb{R}$ ("wavelet-like").*

(b) $\operatorname{supp}(\hat{\psi}_1) \subseteq \left[-\frac{1}{2}, -\frac{1}{16}\right] \cup \left[\frac{1}{16}, \frac{1}{2}\right]$.

(c) $\hat{\psi}_1 \in C^\infty(\mathbb{R})$.

(d) $\sum_{k=-1,0,1} |\hat{\psi}_2(\xi + k)|^2 = 1$ *for almost every $\xi \in [-1, 1]$ ("bump-like").*

(e) $\operatorname{supp}(\hat{\psi}_2) \subseteq [-1, 1]$.

(f) $\hat{\psi}_1 \in C^\infty(\mathbb{R})$.

Then, we call ψ a classical shearlet*.*

The classical shearlets lead to a Parseval frame, as it is shown next.

Theorem 3.3.6 ([50])**.** *Let ψ be a classical shearlet. Then the shearlet system $\mathcal{SH}_\psi$ forms a Parseval frame for $L^2(\mathbb{R}^2)$.*

Proof. Let $a_j := 2^{-j}$ and $A_j := A_{a_j}$. By the properties of ψ_1 and ψ_2 from Definition 3.3.5, we have

$$\begin{aligned}\sum_{j\in\mathbb{Z}}\sum_{k\in\mathbb{Z}}|\hat{\psi}(S_{-k}^{\mathsf{T}}A_{-j}\xi)|^2 &= \sum_{k\in\mathbb{Z}}|\hat{\psi}_1(2^{-j}\xi_1)|^2\sum_{k\in\mathbb{Z}}|\hat{\psi}_2(2^{j/2}\xi_2/\xi_1-k)|^2\\ &= \sum_{j\in\mathbb{Z}}|\hat{\psi}_1(2^{-j}\xi_1)|^2 = 1 \quad \text{for almost every } \xi\in\mathbb{R}^2.\end{aligned}$$

Using Plancherel's and Parseval's identity [99], we conclude that $\mathcal{SH}_\psi$ indeed forms a Parseval frame for $L^2(\mathbb{R}^2)$. □

We refer as *frequency tiling* to the set of Fourier essential supports of each function in a representation system. The frequency tiling that is produced by classical shearlets (see Figure 3.8) is biased towards the ξ_2-axis. This leads to problems when analyzing singularities that are aligned with the x_1-axis, with the need to perform an infinite number of shearings. Of course, this is not possible in the digital realm and we need to address it.

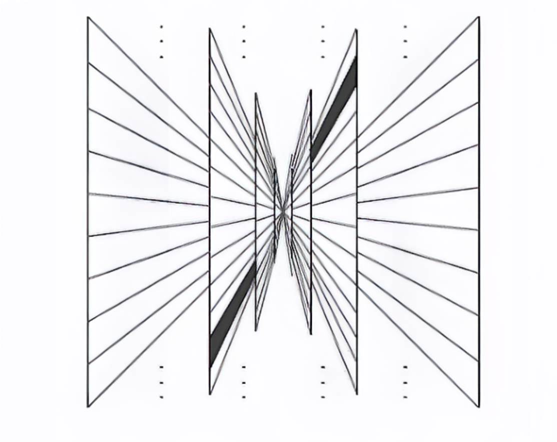

Figure 3.8: Frequency tiling of the classical shearlets system.

3.3.2 Cone-adapted shearlets

The standard approach for fixing the bias problem of classical shearlets are the so-called, *cone-adapted shearlets* [73]. This approach splits the Fourier domain into *cones*, and allows us to have a significantly more balanced frequency tiling, covering the frequency domain optimally with wedge-shaped supports. Figure 3.9 shows the shearlet cones and tiling of frequency domain.

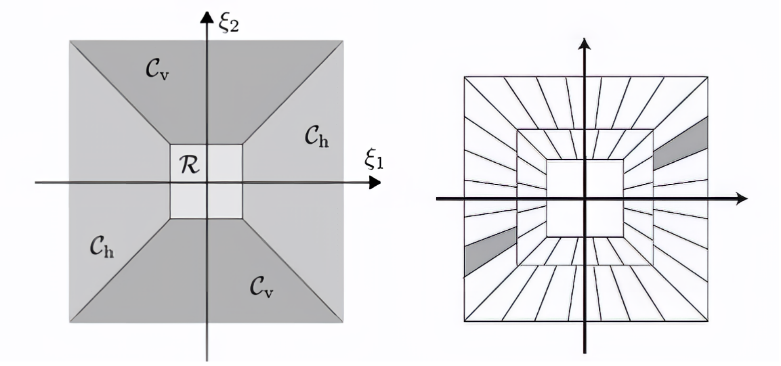

Figure 3.9: Frequency tiling of the cone-adapted shearlet system. Left: Cones, Right: Frequency tiling.

Formally, the cones of the frequency plane divisions, shown at Figure 3.9 are given by

$$\begin{aligned}\mathcal{C}_h :=& \{\xi : |\xi_1| \geq 1, |\xi_2/\xi_1| \leq 1\},\\ \mathcal{C}_v :=& \{\xi : |\xi_2| \geq 1, |\xi_1/\xi_2| \leq 1\},\\ \mathcal{R} :=& [-1,1]^2.\end{aligned}$$

In addition, one also needs to modify the scaling procedure by introducing a new scaling matix. The *cone-adapted* shearlet system makes use of the following three matrices:

$$A_a := \begin{pmatrix} a & 0 \\ 0 & \sqrt{a} \end{pmatrix}, \quad \widetilde{A}_a := \begin{pmatrix} \sqrt{a} & 0 \\ 0 & a \end{pmatrix}, \quad \text{and} \quad S_s := \begin{pmatrix} 1 & s \\ 0 & 1 \end{pmatrix} \quad \text{for } a > 0 \text{ and } s \in \mathbb{R}.$$

Next, given $(a, s, t) \in \mathbb{R}_+ \times \mathbb{R} \times \mathbb{R}^2$, $\psi, \tilde{\psi}, \varphi \in L^2(\mathbb{R}^2)$, and $x \in \mathbb{R}^2$, we define

$$\begin{aligned}\psi_{a,s,b}(x) :=& a^{\frac{3}{4}} \psi\left(S_s A_a (x-b)\right),\\ \tilde{\psi}_{a,s,b}(x) :=& a^{\frac{3}{4}} \widetilde{\psi}\left(S_s^{\intercal} \widetilde{A}_a (x-t)\right),\\ \varphi_b(x) :=& \varphi(x-b).\end{aligned} \tag{3.3.3}$$

Following [48], we then define the cone-adapted continuous shearlet transform [73] as follows.

Definition 3.3.7 (Cone-adapted continuous shearlet transform, [73])**.** *Let $\psi, \tilde{\psi}, \varphi \in L^2(\mathbb{R}^2)$. The* cone-adapted shearlet system *associated with $\psi, \tilde{\psi}, \varphi$ is defined by*

$$\mathcal{SH}_{\psi,\tilde{\psi},\varphi} := \mathcal{P}_{\mathcal{C}_h}\Psi(\psi) \cup \mathcal{P}_{\mathcal{C}_v}\tilde{\Psi}(\tilde{\psi}) \cup \mathcal{P}_{\mathcal{R}}\Phi(\varphi),$$

where $\mathcal{P}_{\mathcal{C}_h}, \mathcal{P}_{\mathcal{C}_v}, \mathcal{P}_{\mathcal{R}}$ are the projection operators and

$$\begin{aligned}\Psi(\psi) :=& \{\psi_{a,s,b} : a \in \mathbb{R}_+, s \in \mathbb{R}, b \in \mathbb{R}^2\},\\ \tilde{\Psi}(\tilde{\psi}) :=& \{\tilde{\psi}_{a,s,b} : a \in \mathbb{R}_+, s \in \mathbb{R}, b \in \mathbb{R}^2\},\\ \Phi(\varphi) :=& \{\varphi_b : b \in \mathbb{R}^2\},\end{aligned}$$

with $\psi_{a,s,b}, \tilde{\psi}_{a,s,b}$ *and* φ_b *are given by Equations* (3.3.3).

In the above definition, the functions $\psi, \tilde{\psi}, \varphi \in L^2(\mathbb{R}^2)$ are not specified. A first approach would be to use classical shearlets. Fortunately, cone-adapted shearlets form also a Parseval frame as classical shearlets (see Theorem 3.3.6).

Theorem 3.3.8 ([73]). *Let* $\psi \in L^2(\mathbb{R}^2)$ *be a classical shearlet and* $A_j = A_{a_j}$ *where* $a_j = 2^{-j}$*. Then*

$$\Psi(\psi) := \{a_j^{-\frac{3}{4}}\psi(S_k A_j(x-t)) : j \in \mathbb{Z}_+, k \in \mathbb{Z}, t \in \mathbb{Z}^2\}$$

forms a Parseval frame for

$$\{f \in L^2(\mathbb{R}^2) : \operatorname{supp} \hat{f} \subset \{\xi \in \mathbb{R}^2 : |\xi_1| \geq 1, |\xi_2/\xi_1| \leq 1\}\}.$$

The proof of Theorem 3.3.8 is done using the same arguments as in Theorem 3.3.6. Fortunately, we can extend this result to the entire space $L^2(\mathbb{R}^2)$ as follows. Let us consider

$$P_{\mathcal{C}_h}\Psi(\psi) \cup P_{\mathcal{C}_v} \cup P_{\mathcal{R}}\Phi(\varphi),$$

where $P_{\mathcal{C}_h}, P_{\mathcal{C}_v}$ and $P_{\mathcal{R}}$ are the projections in the Fourier domain, see Figure 3.9. This form changes the Plancherel formula given by Theorem 3.3.4, obtaining the cone-adapted version

$$\begin{aligned} ||f||_2^2 = &\int_{b\in\mathbb{R}^2} |\langle \mathcal{P}_{\mathcal{R}} f, T_b\varphi\rangle|^2 db + \int_{b\in\mathbb{R}^2}\int_{-1}^{1}\int_0^1 |\mathcal{SH}_\psi(\mathcal{P}_{\mathcal{C}_h} f)(a,s,b)|^2 a^{-3} dadsdb \\ &+ \int_{b\in\mathbb{R}^2}\int_{-1}^{1}\int_0^1 |\mathcal{SH}_{\tilde{\psi}}(\mathcal{P}_{\mathcal{C}_v} f)(a,s,b)|^2 a^{-3} dadsdb, \\ &\text{for all} \quad f \in L^2(\mathbb{R}^2), \end{aligned} \tag{3.3.4}$$

where $P_{\mathcal{R}}, P_{\mathcal{C}_h}$ and $P_{\mathcal{C}_v}$ are the projection operators on the Fourier domain.

In this thesis, we are interested in spatial localization, but as discussed before, a band-limited function cannot have compact support due to the uncertainty principle. To also allow compactly supported functions, we need to give up on the Parseval frame property. The good news is that we can still control the frame bounds in this situation. The following result shows this fact.

Theorem 3.3.9 ([73]). *For* $\alpha > \gamma > 3$, $q > q' > 0$ *and* $r > 0$*, let*

$$|\hat{\psi}(\xi_1, \xi_2)| \leq C_1(\alpha, \gamma, q, q', r)\min\{1, |q\xi_1|^\alpha\}\min\{1, |q'\xi_1|^{-\gamma}\}\min\{1, |r\xi_2|^{-\gamma}\}, \tag{3.3.5}$$

and assume that

$$\sum_{j,k\in Z} |\hat{\psi}(S_{-k}^{\intercal} A_{2^j}\xi)|^2 \geq C_2(\alpha, \gamma, q, q', r) > 0 \quad \textit{for almost every} \quad \xi \in \mathbb{R}^2.$$

For $\tilde{\psi}$*, we assume the same conditions. Then, for a scaling function* φ *(see [73]), the cone-adapted shearlet system* $\mathcal{SH}_{\psi,\tilde{\psi},\varphi}$ *forms a frame for* $L^2(\mathbb{R}^2)$ *with frame bounds following the estimate*

$$C_1(\alpha, \gamma, q, q', r) \leq A \leq B \leq C_2(\alpha, \gamma, q, q', r).$$

Proof. See [73]. □

It is also important to study the approximation properties of shearlets. Similarly to curvelets, shearlet systems, under certain assumptions, meet the optimal $N-$term approximation for $\mathcal{E}^2(\mathbb{R}^2)$ rate up to a log-factor, which is regarded as negligible. The next theorem, states the exact conditions.

Theorem 3.3.10 ([73])**.** *Let $\psi \in L^2(\mathbb{R}^2)$ be compactly supported. For $\alpha > 5$, $\gamma \geq 4$, and some $h \in L^1(\mathbb{R})$, assume that Equation (3.3.2) is satisfied as well as*

$$\left| \frac{\partial}{\partial \xi_2} \hat{\psi}(\xi) \right| \leq |h(\xi_1)|(1 + |\xi_2/\xi_1|)^{-\gamma}$$

and the same conditions for $\tilde{\psi}$. If $\mathcal{SH}_{\psi,\tilde{\psi},\varphi}$ forms a frame for $L^2(\mathbb{R}^2)$, there there exists as constant $C > 0$ such that for all $f \in \mathcal{E}(\mathbb{R}^2)$, we have

$$\sigma_N(f, \mathcal{SH}_{\psi,\tilde{\psi},\varphi}) \leq CN^{-1}(\log N)^{3/2} \quad as \quad N \to \infty.$$

The approximation properties of the shearlet system makes them capable to optimally approximate oriented singularities in two dimensions. By using this principle, we will later be able to use shearlets to resolve the wavefront set of a function. This notion was first introduced by Kutyniok and Labate [72] for the band limited case, and later extended to the compactly supported case by Grohs [48]. In the next section, we will introduce the main assumptions that make this possible. From now on, we will refer to the systems that attain an approximation rate given in Theorem 3.2.6 (possibly up to log factors) as having *optimal representation*, such as shearlets and curvelets.

3.4 Continuous shearlet system for wavefront set extraction

Before we are able to discuss the capabilities of shearlet systems to resolve the wavefront set, we need to introduce the notion of Sobolev spaces. From the shearlet admissibility condition (Definition 3.3.2) and the notion of vanishing moments (Definition 3.3.3), one can see that a lot of functions are suitable to be shearlets. In particular, all they need to satisfy is to have one vanishing moment in the x_1-direction. This is equivalent to be a partial derivative in x_1-direction of a square-integrable function. The need to define derivatives in functional spaces introduces the idea of Sobolev spaces.

We denote by $H_{(n_1,n_2)}$ the Sobolev space defined by

$$H_{(n_1,n_2)}(\mathbb{R}^2) \colon= \left\{ f \in L^2(\mathbb{R}^2) : \left(\frac{\partial}{\partial x_1}\right)^{n_1} \left(\frac{\partial}{\partial x_2}\right)^{n_2} f \in L^2(\mathbb{R}^2) \right\}.$$

The norm of $H_{(n_1,n_2)}(\mathbb{R}^2)$, $||\cdot||_{H_{(n_1,n_2)}(\mathbb{R}^2)}$ is defined as:

$$||f||_{H_{(n_1,n_2)}(\mathbb{R}^2)} \colon= \left|\left| \left(\frac{\partial}{\partial x_1}\right)^{n_1} \left(\frac{\partial}{\partial x_2}\right)^{n_2} f \right|\right|_{L^2(\mathbb{R}^2)}$$

This definition allows us to write the vanishing moments in terms of the Sobolev derivatives.

Theorem 3.4.1. *Let $\varphi \in H_{(n,0)}(\mathbb{R}^2)$ with $\hat{\varphi}(0) \neq 0$. Then the function*

$$\psi(x) = (-1)^n \left(\frac{\partial}{\partial x_1}\right)^n \varphi(x) \tag{3.4.1}$$

is a continuous shearlet with n vanishing directional moments in x_1-direction. Conversely, if ψ is a continuous shearlet with n vanishing moments in x_1-direction. Then ψ can be written in the form given by Equation (3.4.1) *with a function $\varphi \in H_{(n,0)}(\mathbb{R}^2)$.*

Now, if $f \in L^2(\mathbb{R}^2)$, since $\mathcal{R} \cup \mathcal{C}_h \cup \mathcal{C}_v$ covers the Fourier domain, we can decompose $f = \mathcal{P}_\mathcal{R} f + \mathcal{P}_{\mathcal{C}_h} \cup \mathcal{P}_{\mathcal{C}_v} f$, where $\mathcal{P}_\mathcal{R}$ is the projection upon $\mathcal{R}$, similarly $\mathcal{P}_{\mathcal{C}_h}$ and $\mathcal{P}_{\mathcal{C}_v}$. A classical result from microlocal analysis states that directional singularities with slope $\geq$ 1 manifest themselves as slow decay in $\mathcal{C}_h$ and singularities with slope ≤ 1 as slow decay in $\mathcal{C}_v$, see [72]. Therefore, $\mathcal{P}_{\mathcal{C}_h} f$ can be seen as the part of f containing singularities with slope ≥ 1. Also, $\mathcal{P}_{\mathcal{C}_v} f$ is therefore the part of f containing singularities with slope ≤ 1 and $\mathcal{P}_\mathcal{R} f$ as a smooth low-pass approximation of f.

Using the aforementioned notion of microlocal analysis jointly with the cone-adapted construction of the shearlet system, Kutyniok and Labate have shown in [72] that the shearlet coefficients in the norm equation (3.3.4) characterize the wavefront set $\mathrm{WF}(f)$ of a tempered distribution f. Before we introduce such a result, in order to be consistent with the notation of the singular directed points of Section 2.2, we need to parametrize the wavefront set orientations given by a directional vector $\xi \in \mathbb{R}^2$ with the shearing parameter s. Since normal vectors fully describe their orientations, we can further assume that $\xi \in \mathbb{S}^1$.

Now, in the horizontal cone, shearing changes the orientation of shearlets by an angle $\theta := \theta(s) = \arctan(s)$. On the vertical cone, we have $\theta := \theta(s) = \arctan(1/s)$. Therefore, we can parametrize the directional vector ξ with the shearing parameter

$$\begin{aligned} \xi_h(s) =& (\cos(\arctan(s)), \sin(\arctan(s))), \\ \xi_v(s) =& (\cos(\arctan(1/s)), \sin(\arctan(1/s))). \end{aligned} \tag{3.4.2}$$

Finally, the result reads as follows.

Theorem 3.4.2 (Resolution of the Wavefront Set I,[72])**.** *Let $\psi, \tilde{\psi} \in L^2(\mathbb{R}^2)$ be classical shearlet generators, f be a tempered distribution and let us assume that orientations $\xi \in \mathbb{S}^1$ are parametrized by the shearing parameter s according to* (3.4.2)*. Let us further define the shearlet transforms corresponding to the horizontal and vertical cones:*

$$\begin{aligned} \mathcal{SH}^h_\psi(f) &= \mathcal{SH}_\psi(\mathcal{P}_{\mathcal{C}_h} f), \\ \mathcal{SH}^v_{\tilde{\psi}}(f) &= \mathcal{SH}_{\tilde{\psi}}(\mathcal{P}_{\mathcal{C}_v} f). \end{aligned}$$

(i) Let $\mathcal{F} \subset \mathbb{R}^2$ *be defined by*

$$\begin{aligned}\mathcal{F} := \{ & b_0 \in \mathbb{R}^2 : \textit{there exists a neighborhood } U \textit{ of } b_0, \textit{ such that for every } b \in U, \\ & |\mathcal{SH}^h_\psi f(a,s,b)| = O(a^{-k}) \textit{ and } |\mathcal{SH}^v_{\tilde{\psi}} f(a,s,b)| = O(a^{-k}) \textit{ as } a \to 0, \textit{ for all } k \in \mathbb{N}, \\ & \textit{with the } O(\cdot)\textit{-terms uniform over } (b,s) \in U \times [-1,1]\}.\end{aligned}$$

Then

$$\operatorname{sing\,supp}(f)^c = \mathcal{F}.$$

(ii) Let $\mathcal{D} := \mathcal{D}_1 \cup \mathcal{D}_2$*, where*

$$\begin{aligned}\mathcal{D}_1 := \{ & (b_0, \xi(s_0)) : \textit{there exists a neighborhood } U \textit{ of } (b_0, s_0) \in \mathbb{R}^2 \times [-1,1], \textit{ such that,} \\ & \textit{for every } (b,s) \in U,\ |\mathcal{SH}^v_{\tilde{\psi}} f(a,s,b)| = O(a^{-k}) \textit{ as } a \to 0, \textit{ for all } k \in \mathbb{N}, \\ & \textit{with the } O(\cdot)\textit{-terms uniform over } (b,s) \in U\},\end{aligned}$$

and

$$\begin{aligned}\mathcal{D}_2 := \{ & (b_0, \xi(s_0)) : \textit{there exists a neighborhood } U \textit{ of } (b_0, s_0) \in \mathbb{R}^2 \times [1,\infty], \textit{ such that,} \\ & \textit{for every } (b,1/s) \in U,\ |\mathcal{SH}^h_\psi f(a,s,b)| = O(a^{-k}) \textit{ as } a \to 0, \textit{ for all } k \in \mathbb{N}, \\ & \textit{with the } O(\cdot)\textit{-terms uniform over } (b,1/s) \in U\}.\end{aligned}$$

Then the wavefront set $\mathrm{WF}(f)$ *is given by*

$$\mathrm{WF}(f)^c = \mathcal{D}.$$

On the one hand, the statement (ii) of Theorem 3.4.2 shows that the continuous shearlet transform on the horizontal cone $\mathcal{SH}^h_\psi f(a,s,b)$ identifies the wavefront set for directions $\xi(s)$ such that $|s| = |\frac{\xi_2}{\xi_1}| \leq 1$ (in the frequency domain). On the other hand, the continuous shearlet $\mathcal{SH}^v_{\tilde{\psi}} f(a,s,b)$ identifies the wavefront set for directions $\xi(s)$ such that $|s| = |\frac{\xi_1}{\xi_2}| \leq 1$, i.e. $|\frac{\xi_2}{\xi_1}| \geq 1$. This allows us to separate the wavefront set in different orientations.

Similar to the original work [72], the proof of Theorem 3.4.2 requires the introduction of some lemmata, originally inspired by the work [23]; a similar result was shown for curvelets. In this work, we will include some of these propositions while referring to [72] for the rest.

Lemma 3.4.3 ([72])**.** *Let* $f \in L^2(\mathbb{R}^2)$ *with* $||f||_\infty < \infty$ *and* $\psi, \tilde{\psi} \in L^2(\mathbb{R}^2)$ *classical shearlet generators and* $\varphi \in L^2(\mathbb{R}^2)$ *a scaling function. If* $\operatorname{supp}(f) \subset \mathcal{B} \subset \mathbb{R}^2$*, then for all* $k > 1$ *there is a constant* $C_k > 0$ *such that*

$$|\mathcal{SH}^h_\psi f(a,s,b)| = |\langle f, \psi_{a,s,b}\rangle| \leq C_k C(s)^2 ||f||_\infty a^{\frac{1}{4}} (1 + C(s)^{-1} a^{-1} d(b,\mathcal{B})^2)^{-k},$$

where $C(s) = \left(1 + \frac{s^2}{2} + \left(s^2 + \frac{s^2}{4}\right)^{\frac{1}{2}}\right)^{\frac{1}{2}}$ *and* $d(b,\mathcal{B})$ *is the distance from* b *to the set* $\mathcal{B}$*.*

Proposition 3.4.4 ([72]). *Let $\mathcal{F}$ and $\mathcal{D}$ be defined as in Theorem 3.4.2 and $f \in L^2(\mathbb{R}^2)$. Then:*

(i) $\operatorname{sing\,supp}(f)^c \subseteq \mathcal{F}$.

(ii) $\mathrm{WF}(f)^c \subseteq \mathcal{D}$.

Proof. (i) Let b_0 be a regular point of f. Then there exists $\phi \in C_0^\infty(\mathbb{R}^2)$ with $\phi(b_0) = 1$ on $B(b_0, \delta)$, i.e., the ball centered at b_0 with radius δ, such that $\phi f \in C^\infty(\mathbb{R}^2)$. We will show that $b_0 \in \mathcal{F}$. For this, we decompose $\mathcal{SH}_\psi f(a, s, b)$ as

$$\mathcal{SH}^h_\psi f(a, s, b) = \langle \psi_{a,s,b}, \phi f\rangle + \langle \psi_{a,s,b}, (1 - \phi)f\rangle. \tag{3.4.3}$$

Observe that, since ψ is a classical shearlet, we have

$$|\langle \psi_{a,s,b}, \phi f\rangle| \leq a^{\frac{3}{4}} \int_{\mathbb{R}^2} |\hat{\psi}_1(a\xi_1)||\hat{\psi}_2\left(\frac{1}{\sqrt{a}}\left(\frac{\xi_2}{\xi_1} - s\right)\right)||\widehat{\phi f}(\xi)|d\xi.$$

Let us now estimate this integral for $\xi_1 > 0$. The case $\xi_1 \leq 0$ is analogous. Since $\phi \in C_0^\infty(\mathbb{R}^2)$, for each $k \in \mathbb{N}$, there exists a constant C_k with $|\widehat{\phi f}(\xi)| \leq C_k|\xi|^{-2k}$. In addition, since ψ is a classical shearlet, $\operatorname{supp}(\hat{\psi}_1) \subseteqq [-\frac{1}{2}, -\frac{1}{16}] \cup [\frac{1}{16}, \frac{1}{2}]$ and $\operatorname{supp}(\hat{\psi}_2) \subseteq [-1, 1]$. Using these arguments, for $k > 2$, the first term on the right hand side of Equation (3.4.3) can be estimated as follows:

$$\begin{aligned}
&a^{\frac{3}{4}} \int_{\mathbb{R}_+ \times \mathbb{R}} |\hat{\psi}_1(a\xi_1)||\hat{\psi}_2(\frac{1}{\sqrt{a}}(\frac{\xi_2}{\xi_1} - s))||\widehat{\phi f}(\xi)|d\xi. \\
&\leq C_k||\hat{\psi}||_\infty a^{\frac{3}{4}} \int_{1/2a}^{2/a} \int_{(s-\sqrt{a})\xi_1}^{(s+\sqrt{a})\xi_1} |\xi|^{-2k} d\xi_2 d\xi_1 \\
&\leq C_k 2^{-k}||\hat{\psi}||_\infty a^{\frac{3}{4}} \int_{1/2a}^{2/a} \int_{(s-\sqrt{a})\xi_1}^{(s+\sqrt{a})\xi_1} \xi_2^{-k} d\xi_2 d\xi_1 \\
&= \frac{C_k 2^{-k}||\hat{\psi}||_\infty a^{\frac{3}{4}}}{1-k}((s + \sqrt{a})^{1-k} - (s - \sqrt{a})^{1-k}) \int_{1/2a}^{2/a} \xi_1^{1-2k} d\xi_1 \\
&\leq \frac{C_k 2^{-k}||\hat{\psi}||_\infty a^{\frac{3}{4}}}{1-k}((\sqrt{a} - s)^{1-k} - (-\sqrt{a} - s)^{1-k})\frac{1}{1-2k}((2/a)^{2-2k} - (1/a)^{2-2k}) \\
&\leq \frac{C_k 2^{-k}||\hat{\psi}||_\infty a^{\frac{3}{4}}}{k(2k-1)}(\sqrt{a} - s)^{1-k}(2/a)^{2-2k}
\end{aligned}$$

Thus, the above integral behaves as $O(a^k)$ as $a \to 0$, uniformly over $(b, s) \in B(t_0, \delta/2) \times \mathbb{R}$. Finally, using Lemma 3.4.3 of [74], we can estimate the second term of the RHS of Equation (3.4.3):

$$|\langle \psi_{a,s,b}, (1 - \phi)f\rangle| \leq C_k C(s)^2||(1 - \phi)f||_\infty a^{\frac{1}{4}}(1 + C(s)^{-1}a^{-1}d(b, B(b_0, \delta)^c)^2)^{-k},$$

where $k \in \mathbb{N}$ is arbitrary. Since $||(1-\phi)f||_\infty < \infty$, this yields

$$|\langle \psi_{a,s,b}, (1-b)f \rangle| = O(a^k), \quad \text{as} \quad a \to 0,$$

uniformly over $(b,s) \in B(b_0, \delta/2) \times [-1,1]$. In analogy, a similar estimate holds for $\mathcal{SH}^v_\psi f(a,s,t)$, proving (i).

(ii) Let $(b_0, \xi(s_0))$ be a regular directed point of f, with $s_0 \in [-1,1]$. Then there is some $\phi \in C_0^\infty(\mathbb{R}^2)$ with $\phi(b_0) = 1$ on a ball $B(b_0,\delta_1)$ such that, for each $k \in \mathbb{N}$ we have $|\widehat{\phi f}(\xi)| = O((1+|\xi|)^{-k})$ for all $\xi \in \mathbb{S}^1$ satisfying $s = \xi_2/\xi_1 \in B(s_0,\delta_2)$. We will now show that $(b_0, \xi(s_0)) \in \mathcal{D}$.

Let us first decompose $\mathcal{SH}^v_\psi f(a,s,b)$ as in (3.4.3). The second term on the RHS of (3.4.3) can be estimated as in the case (i). For the first term of (3.4.3), we only need to show that

$$\operatorname{supp}(\hat{\psi}_{a,s,b}) \subset \{\xi \in \mathbb{S}^1 : \xi_2/\xi_1 \in B(s_0,\delta_2)\} \quad \text{for all} \quad (b,s) \in B(b_0,\delta_1) \times B(s_0,\delta_2).$$

This holds since in the horizontal cone $\mathcal{C}_h$ $\widehat{\phi f}$ decays rapidly. However, so far we have only considered the case $\xi_1 > 0$; the case $\xi_1 \leq 0$ is analogous. The support of $\hat{\psi}_{a,s,b}$ in this half plane is given by

$$\{(\xi_1,\xi_2) : \xi_1 \in [1/2a, 2/a], \xi_2 \in \xi_1[s-\sqrt{a}, s+\sqrt{a}]\}.$$

Let $(b,s) \in B(b_0,\delta_1) \times B(s_0,\delta_2)$. The cone $\{\xi \in \mathbb{R}^2 : \xi_2/\xi_1 \in B(s_0,\delta_2)\}$ is bounded by the lines $\xi_2 = (s_0-\delta_2)\xi_1$ and $\xi_2 = (s_0+\delta_2)\xi$. Now let $(\xi_1,\xi_2) \in \operatorname{supp}\hat{\psi}_{a,s,b}$. Then for a sufficiently small, we have

$$|\xi_2/\xi_1 - s_0| \leq \sqrt{a} \leq \delta_2,$$

finishing the proof for $|s_0| < 1$. In the case when $|s_0| \geq 1$ (this corresponds to $|\xi_2/\xi_1| \leq 1$), the proof is analogous, using the transform $\mathcal{SH}^v_{\breve{\psi}}(a,s,b)$ instead of $\mathcal{SH}\psi^h(a,s,b)$

□

Now we just need to show the converse inclusions. For this, we need additional preliminary results. We will introduce the results, but we refer to [72] for the proofs.

Lemma 3.4.5 ([72]). *Let $S \subset \mathbb{R}$ be a compact set and $f \in L^2(\mathbb{R}^2)$ with $||f||_\infty \leq \infty$. Suppose that* $\operatorname{supp} f \subset \mathcal{B}$ *for some $\mathcal{B} \subset \mathbb{R}^2$ and define $(\mathcal{B}^\eta)^c = \{x \in \mathbb{R}^2 : d(x,\mathcal{B}) > \eta\}$. Further let $\psi_{a,s,b}$ be a classical shearlet and define $h \in L^2(\mathbb{R})$ by*

$$\hat{h}(\xi) = \int_0^\infty \int_{(\mathcal{B}^\eta)^c} \int_S \mathcal{SH}^h_\psi f(a,s,b) \hat{\psi}_{a,s,b}(\xi) ds db \frac{da}{a^3}.$$

Then $\hat{h}(\xi)$ decays rapidly as $|\xi| \to \infty$ with constants dependent only on $||f||_\infty$ and η.

Lemma 3.4.6 ([72])**.** *Let $S \subset \mathbb{R}$ and $\mathcal{B} \subset \mathbb{R}^2$ be compact sets. Suppose that $G(a, s, b)$ decays rapidly as $a \to 0$ uniformly for $(b, s) \in S \times \mathcal{B}$. Define $h \in L^2(\mathbb{R}^2)$ by*

$$\hat{h}(\xi) = \int_0^\infty \int_{\mathcal{B}} \int_S G(a, s, b)\hat{\psi}_{a,s,b}(\xi) ds dt \frac{da}{a^3}.$$

Then $\hat{h}(\xi)$ decays rapidly as $|\xi| \to \infty$.

Lemma 3.4.7 ([72])**.** *Suppose $0 \leq a_0 \leq a_1$ and $|s| \leq s_0$. Then for $K > 1$, there is a constant C_K, dependent on K only, such that:*

$$|\langle \psi_{a_0,s,b}, \psi_{a_1,s',b'} \rangle| \leq C_K \left(1 + \frac{a_1}{a_0}\right)^{-K} \left(1 + \frac{|s - s'|^2}{a_1}\right)^{-K} \left(1 + \frac{||(t - t')||^2}{a_1}\right)^{-K}.$$

Lemma 3.4.8. *Let $\phi_1 \in C^\infty(\mathbb{R}^2)$ be supported in $B(0, 1)$ (the unitary ball centered at 0), $a_\psi \in \mathbb{R}_+$ be a positive constant and define $\phi(x) := \phi_1(a_\phi^{-1}(x - b))$.*

(i) Suppose $0 \leq \sqrt{a_0} \leq \sqrt{a_1} \leq a_\phi < 1$. Then for $K > 0$,

$$|\langle \phi\psi_{a_0,s,b}, \psi_{a_1,s',b'} \rangle| \leq C_K \left(1 + \frac{a_1}{a_0}\right)^{-K} \left(1 + \frac{|s - s'|^2}{a_1}\right)^{-K} \left(1 + \frac{||(t - t')||^2}{a_1}\right)^{-K}.$$

(ii) Suppose $0 \leq \sqrt{a_0} \leq a_\phi \leq \sqrt{a_1} < 1$, $a_1 \leq a_\phi$. Then for $K > 0$,

$$|\langle \phi\psi_{a_0,s,b}, \psi_{a_1,s',b'} \rangle| \leq C_K \left(1 + \frac{a_1}{a_0}\right)^{-K} \left(1 + \frac{|s - s'|^2}{a_\phi^2}\right)^{-K} \left(1 + \frac{||(t - t')||^2}{a_1}\right)^{-K}.$$

(iii) Suppose $0 \leq \sqrt{a_0} \leq a_\phi \leq a_1 \leq \sqrt{a_1} < 1$, $a_1 \leq a_\phi$. Then for $K > 0$,

$$|\langle \phi\psi_{a_0,s,b}, \psi_{a_1,s',b'} \rangle| \leq C_K \left(1 + \frac{a_\phi}{a_0}\right)^{-K} \left(1 + \frac{||(t - t')||^2}{a_\phi^2}\right)^{-K}.$$

Using the results presented on Lemmas 3.4.5 3.4.5, we can now prove our main theorem:

Proof of Theorem 3.4.2, [72]:
We know by Proposition 3.4.4 that $\operatorname{sing\,supp}(f)^c \subseteq \mathcal{F}$ and $\operatorname{WF}(f)^c \subseteq \mathcal{D}$. Hence, we can prove the following two claims:

(i) $\mathcal{F} \subseteq \operatorname{sing\,supp}(f)^c$.

(ii) $\mathcal{D} \subseteq \operatorname{WF}(f)^c$.

Le us first prove (i). Let $b_0 \in \mathcal{F}$, then there is a δ such that for all $b \in B(b_0, \delta)$, we have that $\mathcal{SH}^h_\psi f(a, s, b) = O(a^k)$ as $a \to 0$, for all $k \in \mathbb{N}$ with $O(\cdot)-$term uniform over $(b, s) \in B(b_0, \delta) \times [-1, 1]$. An analogous estimate holds for $\mathcal{SH}^v_{\tilde{\psi}} f(a, s, b)$.

Let $\phi \in C^\infty(\mathbb{R}^2)$ which is supported in a ball $B(b_0, \nu)$ with $\nu \ll \delta$ and let $\eta = \frac{\delta}{2}$. Set $g = \phi f$ and consider the decomposition

$$\widehat{\phi f}(\xi) = \hat{g}_0(\xi) + \hat{g}_1(\xi) + \hat{g}_2(\xi) + \hat{g}_4(\xi),$$

where $\hat{g}_0(\xi) = (\widehat{\phi P(f)})(\xi)$. In addition, $P(f) = \int_{\mathbb{R}} \langle f, T_b W\rangle T_b W db$, with W being a window function such that $\hat{W} \in C^\infty(\mathbb{R}^2)$ and

$$|\hat{W}(\xi)|^2 + \chi_{C_1}(\xi) \int_0^1 |\hat{\psi}_1(a\xi_1)|^2 \frac{da}{a} + \chi_{C_2}(\xi) \int_0^1 |\hat{\psi}_1(a\xi_2)|^2 \frac{da}{a} = 1 \quad \text{for a.e. } \xi \in \mathbb{R}^2,$$

where $C_1 = \{(\xi_1, \xi_2) \in \mathbb{R}^2 : |\xi_2/\xi_1| \leq 1\}$, and $C_2 = \{(\xi_1, \xi_2) \in \mathbb{R}^2 : |\xi_2/\xi_1| > 1\}$.

Moreover, for $i = 1, 2$ we have

$$\hat{g}_i(\xi) = \chi_{C_1}(\xi) \int_{\mathcal{Q}_i} \hat{\psi}_{a,s,b}(\xi) \mathcal{SH}_h g(a, s, b) d\mu(a, s, b),$$
$$\hat{g}_{i+2}(\xi) = \chi_{C_2}(\xi) \int_{\mathcal{Q}_i} \hat{\psi}_{a,s,b}(\xi) \mathcal{SH}_v g(a, s, b) d\mu(a, s, b),$$

where $d\mu(a, s, b) = \frac{da}{a^3} dsdb$, $\mathcal{Q}_1 = [0, 1] \times [-1, 1] \times B(b_0, \nu)$ and $\mathcal{Q}_2 = [0, 1] \times [-1, 1] \times B(b_0, \nu)^c$.

The term $\hat{g}_0(\xi)$ decays rapidly as $|\xi| \to \infty$ since $\phi, P(f) \in C^\infty(\mathbb{R}^2)$. Moreover, the term $\hat{g}_1(\xi)$ decays rapidly as $|\xi| \to \infty$ by Lemma 3.4.5. In addition, Lemma 3.4.6 shows that $\hat{g}_1(\xi)$ decays rapidly as $|\xi| \to \infty$ provided that $\mathcal{SH}^h_\psi g$ decays rapidly as $a \to 0$ uniformly over $(b, s) \in B(b_0, \nu) \times [-1, 1]$. We will consider only the analysis of the terms $\hat{g}_i$, for $i = 1, 2$; the cases $i = 3, 4$ are analogous.

We will show that $\mathcal{SH}^h_\psi g$ indeed decays rapidly as $a \to 0$ uniformly over $B(b_0, \nu) \times [-1, 1]$. In order to prove this, we decompose f as $f = P(f) + P_{C_1} f + P_{C_2} f$, where $\widehat{P_{C_1} f} = \hat{f} \chi_{C_1}$ and $\widehat{P_{C_2} f} = \hat{f} \chi_{C_2}$. It is clear that $\mathcal{SH}_h \phi P(f)$ decays rapidly by the smoothness of ϕ and $P(f)$. Next, we examine the term $P_{C_1} f$. The analysis of $P_{C_2} f$ is very similar and will be omitted. We use the decomposition $P_{C_1} f = f_1 + f_2$ where

$$f_i(x) = \int_{\mathcal{Q}_i} \psi_{a,s,b}(x) \mathcal{SH}^h_\psi f_i(a, s, b) d\mu(a, s, b), \quad i = 1, 2.$$

Let us start by considering the term corresponding to f_1. We have

$$\mathcal{SH}^h_\psi(\phi f_1)(a, s, b) = \langle \phi f_1, \psi_{a,s,b}\rangle = \int_{\mathcal{Q}_1} \langle \phi \psi_{a,s,b}, \psi_{a',s',b'} \mathcal{SH}^h_\psi f_1(a', s', b')\rangle d\mu(a', s', b'). \tag{3.4.4}$$

We will decompose $\mathcal{Q}_1 = \mathcal{Q}_{10} \cup \mathcal{Q}_{11} \cup \mathcal{Q}_{12}$, corresponding to $a' > \delta', a' \leq \delta\sqrt{a'}$ and $\sqrt{a'} \leq \delta$, respectively. In case $\sqrt{a}, \sqrt{a'} \leq \delta$, by Lemma 3.4.8 we have that

$$|\langle \phi \psi_{a,s,b}, \psi_{a',s',b'}\rangle| \leq C_K \left(1 + \frac{a_1}{a_0}\right)^{-K} \left(1 + \frac{||(b - b')||^2}{a_1}\right)^{-K}. \tag{3.4.5}$$

We claim that, for $m > 4$ and $K \geq m-1$

$$\int_0^\delta \left(1+\frac{a_1}{a_0}\right)^{-K} (a')^m \frac{da'}{(a')^3} \leq C_{m,K} a^{m-2}, \quad 0 < a < \delta. \tag{3.4.6}$$

Indeed, for $a' = a_0 \leq a = a_1$,

$$\int_0^a \left(1+\frac{a}{a'}\right)^{-K} (a')^m \frac{da'}{(a')^3} = a^{m-2} \int_0^1 (1+x)^{-K} dx = C_K a^{m-2}.$$

For $a = a_0 \leq a' = a_1$,

$$\begin{aligned}\int_0^\delta \left(1+\frac{a'}{a}\right)^{-K} (a')^m \frac{da'}{(a')^3} &= a^{m-2} \int_1^{\delta/a} x^{m-3}(1+x)^{-K} dx \\ &\leq a^{m-2} \int_1^\infty x^{m-3}(1+x)^{-K} dx = C_{K,m} a^{m-2}.\end{aligned}$$

Thus (3.4.6) follows from the last two estimates. Using (3.4.6) it follows that

$$\begin{aligned}&\int_{\mathcal{Q}_{12}} \langle \phi\psi_{a,s,b}, \psi_{a',s',b'} \rangle \mathcal{SH}_\psi^h f(a',s',b') d\mu(a',s',b') \\ &\quad \leq C' \int_{-2}^2 \int_{B(b_0,\nu_0)} \int_0^\delta \left(1+\frac{a_1}{a_0}\right)^{-K} (a')^m \frac{da'}{(a')^3} db' ds'\end{aligned}$$

for all $m > 4$. Using the other cases of Lemma 3.4.8 one can show similar estimates for the integrals over the set $\mathcal{Q}_{10}$ and $\mathcal{Q}_{11}$. This shows that $\mathcal{SH}_\psi^h \phi f_1(a,s,b)$ decays rapidly for $a \to 0$ uniformly over $B(b_0,\eta) \times [-1,1]$.

Let us consider now the term corresponding to f_2:

$$\mathcal{SH}_\psi^h \phi f_2(a,s,b) = \langle \phi f_2, \psi_{a,s,b} \rangle = \int_{\mathcal{Q}_1} \langle \phi\psi_{a,s,b}, \psi_{a',s',b'} \rangle \mathcal{SH}_\psi^h f_2(a',s',b')\rangle d\mu(a',s',b').$$

We will decompose $\mathcal{Q}_2 = \mathcal{Q}_{21} \cup \mathcal{Q}_{22}$, corresponding to $||(t-t')|| > \eta$ and $||(t-t')|| \leq \eta$, respectively. Observe that, for $||(b-b')|| > \eta$ and $K > 1$,

$$\int_{B(b_0,\eta)^c} \left(1+\frac{||(b-b')||^2}{a_1}\right)^{-K} db' \leq \int_\eta^\infty \left(1+\frac{r^2}{a_1}\right)^{-K} r dr \leq C' a_1 \left(1+\frac{\eta}{a_1}\right)^{-K+2}.$$

In addition, on the region $\mathcal{Q}_{21}$, the function $\mathcal{SH}_\psi^h f_2(a',s',b')$ is bounded by $C'(a')^{3/4}$ since f is bounded. Thus

$$\begin{aligned}&\int_{\mathcal{Q}_{21}} \langle \phi\psi_{a,s,b}, \psi_{a',s',b'} \rangle \mathcal{SH}_\psi^h f_2(a',s',b') d\mu(a',s',b') \\ &\quad \leq C' \int_{-2}^2 \int_0^\delta \int_{B(b_0,\eta)^c} \left(1+\frac{||(b-b')||^2}{a_1}\right)^{-K} db' \left(1+\frac{a_1}{a_0}\right)^{-K} (a')^{3/4} \frac{da'}{(a')^3} db' \\ &\quad \leq C' \int_0^\eta a_1 \left(1+\frac{\eta}{a_1}\right) \left(1+\frac{a_1}{a_0}\right)^{-K} \frac{da'}{(a')^{9/4}}.\end{aligned}$$

The above term decays rapidly as $a \to 0$, uniformly over $\mathcal{Q}_{21}$. As for the region $\mathcal{Q}_{22}$, if $b \in B(b_0, \eta)$ and $||(b - b')|| > \eta$, then $b' \in B(b_0, \delta)$ and thus the function $\mathcal{SH}_h f$ decays rapidly, for $a \to 0$, over this region. Repeating the analysis as in the case $\mathcal{Q}_{12}$, we can prove that $\int_{\mathcal{Q}_{22}} \langle \phi\psi_{a,s,b}, \mathcal{SH}_h f(a', s', b') d\mu(a', s', b')\rangle$ is of rapid decay, as $a \to 0$ uniformly over $\mathcal{Q}_{22}$. Combining these observations, we conclude that $\mathcal{SH}_h \psi f_2(a, s, b)$ decays rapidly as $a \to 0$ uniformly over $B(b_0, \eta) \times [-1, 1]$.

It follows that $\mathcal{SH}^h_\psi g(a, s, b)$ decays rapidly as $a \to 0$ uniformly for all $(b, s) \in B(b_0, \eta) \times [-1, 1]$ and, thus, by Lemma 3.4.6 $\hat{g}_1(\xi)$ decays rapidly as $|\xi| \to \infty$. We can now conclude that $\hat{g}$ decays rapidly as $|\xi| \to \infty$, hence completing the proof of (i).

In order to show part (ii), we only sketch the idea of the proof, since it is very similar to part (i). Let $(b_0, s_0) \in \mathcal{D}$. We consider separately the case $|s_0| \leq 1$ and $|s_0| \geq 1$. In the first case, for all $b \in B(b_0, \delta)$ and $s \in B(s_0, \delta)$, we have that $|\mathcal{SH}^h_\psi f(a, s, b)| = O(a^k)$, as $a \to 0$, for all $k \in \mathbb{N}$ with $O(\cdot)-$term uniform over $(b, s) \in B(b_0, \delta) \times B(s_0, \delta)$.

Choose $\phi \in L^2(\mathbb{R}^2)$ which is supported in ball $B(b_0, \nu)$ with $\nu \ll \delta$ and let $\eta = \frac{\delta}{2}$. Then the proof proceeds as in part (i), replacing $B(b_0, \delta) \times [-1, 1]$ with $B(b_0, \delta) \times B(s_0, \delta)$. Also, for the estimates involving inner products of $\psi_{a,s,b}$ and $\psi_{a',s',b'}$ we will now use Lemma 3.4.8 including the directionally sensitive term. For example, when $\sqrt{a}, \sqrt{a'} \leq \delta$, by Lemma 3.4.8 we will use the estimate

$$|\langle \phi\psi_{a,s,b}, \psi_{a',s',b'}\rangle| \leq C_K \left(1 + \frac{a_1}{a_0}\right)^{-K} \left(1 + \frac{|s - s'|^2}{a_1}^{-K}\right) \left(1 + \frac{||(b - b')||^2}{a_1}\right)^{-K}$$

instead than (3.4.5). We can proceed similarly for the other estimates. The proof for the case $|s_0| \geq 1$ is exactly the same, with the transform $\mathcal{SH}^v_{\tilde{\psi}} f(a, s, b)$ replacing $\mathcal{SH}^h_\psi f(a, s, b)$. This finishes our proof. □

This result shows how (cone-adapted) classical shearlets are capable to resolve the wavefront set. This eases the computation of the wavefront set over its original definition (Definition 2.2.6). In particular, Theorem 3.4.2 establishes a rule for the localization and microlocalization procedure, which was originally unspecified. As we can see, the proof of this important result is strongly based on the optimal representation quality of the shearlet system (see Theorem 3.3.10). Similarly, curvelets and other optimal multiscale directional systems satisfy similar estimates for wavefront set resolution. In this case, our choice in the shearlet system is based on its faithful discretization, i.e., it has a uniform treatment of the continuous and digital realm.

We can observe that Theorem 3.4.2 requires the shearlet system to be generated by classical shearlets and the target function to be a tempered distribution. In 2011, Grohs [48] showed that the same result can be attained with weaker assumptions, namely, the generator functions need only sufficiently many anisotropic vanishing moments. In these two cases, the shearlets are band-limited, i.e., compactly supported in the Fourier domain.

In this thesis, we aim to combine these results on shearlet based wavefront set resolution with image reconstruction methods. This is motivated by the amount of information contained in the wavefront set of an image, which can be used as a prior in a reconstruction method. Due to the uncertainty principle, band-limited shearlets, lack of good resolution in the spatial domain. For this reason, we are choosing compactly supported shearlets as our main system. We can choose this setting since Grohs and Kereta [49] extended the results on wavefront set resolution to compactly supported shearlets. Later in Chapter 6 we make use a special case of band-limited shearlets to digitize Fourier Integral Operators. In the following, we will present the results for weaker assumptions which include the compactly supported case.

3.4.1 Beyond classical shearlets

Since we are working from now on in the cone-adapted setting, where we need two shearlet generators ψ and $\tilde{\psi}$, we will assume from now on that $\psi = \tilde{\psi}$. This will reduce the notation by just introducing ψ with the premise that we are still talking about the cone-adapted system. The possibility for weaker assumptions in shearlet-based wavefront set resolution is mainly based on the fact that shearlets that hold the estimate of Theorem 3.4.1 exist in abundance, but the cone-adapted construction is very specific. This motivates us to ask, what is needed for generating functions ψ so a shearlet system results in a representation like (3.3.4) and at the same time holds the properties of Theorem 3.4.2. Grohs showed in [48] that the only restriction on ψ to have these two properties is the vanishing moments in x_1-directions, as shown in the next theorem. In our work, the main advantage of this approach is that it can be also applied to tempered distributions.

Theorem 3.4.9 ([48]). *Let ψ be a Schwartz function with infinitely many vanishing moments in x_1-direction. In addition, let f be a tempered distribution and $\mathcal{D} = \mathcal{D}_1 \cup \mathcal{D}_2$, where*

$$\begin{aligned}\mathcal{D}_1 := \{(b_0, \xi(s_0)) : & \text{ there exists a neighborhood } U \text{ of } (b_0, s_0) \in \mathbb{R}^2 \times [-1,1], \text{ such that,} \\ & \text{for every } (b,s) \in U,\ |\mathcal{SH}^v_\psi f(a,s,b)| = O(a^{-k}) \text{ as } a \to 0, \text{ for all } k \in \mathbb{N}, \\ & \text{with the } O(\cdot)\text{-terms uniform over } (b,s) \in U\},\end{aligned}$$

and

$$\begin{aligned}\mathcal{D}_2 := \{(b_0, \xi(s_0)) : & \text{ there exists a neighborhood } U \text{ of } (b_0, s_0) \in \mathbb{R}^2 \times [1,\infty], \text{ such that,} \\ & \text{for every } (b,1/s) \in U,\ |\mathcal{SH}^h_\psi f(a,s,b)| = O(a^{-k}) \text{ as } a \to 0, \text{ for all } k \in \mathbb{N}, \\ & \text{with the } O(\cdot)\text{-terms uniform over } (b,1/s) \in U\}.\end{aligned}$$

Then

$$\mathrm{WF}(f)^c = \mathcal{D}.$$

This result also holds when ψ has finitely many vanishing moments. As in the case of classical shearlets, we need to first introduce some preliminary results in order to fully prove Theorem 3.4.9. We will follow the steps presented in [48], divided into two parts, corresponding to the set inclusions. In this sense, we will first show that for an $N-$regular directed point $(b_0, \xi(s_0))$ of a function $f \in L^2(\mathbb{R}^2)$ the shearlet coefficients with respect to any generating function ψ with sufficiently many vanishing moments in the x_1-direction decay quickly around $b = b_0$ and $\xi(s) = \xi(s_0)$.

Theorem 3.4.10 (Direct Theorem, [48])**.** *Assume that $f \in L^2(\mathbb{R}^2)$ and that $(b_0, \xi(s_0)) \in \mathbb{R}^2 \times \mathbb{S}^1$ is an $N-$regular directed point of f. Let $\psi \in H_{(0,L)}(\mathbb{R}^2)$ (see Theorem 3.4.1), $\hat{\psi} \in L^1(\mathbb{R}^2)$ be a shearlet with M moments which satisfies a decay rate of the form*

$$\psi(x) = O((1+|x|)^{-P}). \tag{3.4.7}$$

Then there exists a neighborhood $U(b_0)$ of b_0 and $V(s_0)$ of s such that for any $1/2 < \alpha < 1$, $t \in U(b_0)$ and $s \in V(s_0)$ we have the decay estimate

$$\mathcal{SH}_h f(a,s,b) = O(a^{-3/4+P/2} + a^{(1-\alpha)M} + a^{-3/4+\alpha N} + a^{(\alpha-1/2)L}), \quad \text{as } a \to 0. \tag{3.4.8}$$

Proof. We will first show that we can assume that f is already localized around b_0 without loss of generality, i.e. $f = \Psi f$ where Ψ is the cutoff function from Definition 2.2.6. To prove this, we will show that

$$\langle (1-\Psi)f, \psi_{a,s,b}\rangle = O(a^{-3/4+P/2}). \tag{3.4.9}$$

By definition, we have that

$$\psi_{a,s,b}(x_1,x_2) = a^{-3/4}\psi\left(\frac{(x_1-b_1)+s(s_2-b_2)}{a}, \frac{x_2-b_2}{a^{1/2}}\right). \tag{3.4.10}$$

Now we note that in computing the inner product (3.4.9) we can assume that $|x-b| > \delta$ for some $\delta > 0$ and b in a small neighborhood $U(b_0)$ of b_0 since by definition of the cut-off function $(1-\Psi)f = 0$ around b_0. Using Equation (3.4.7) we have

$$\begin{aligned} |\psi_{a,s,b}(x)| &\leq Ca^{-3/4}\left(1 + \left|\begin{pmatrix} a^{-1} & sa^{-1} \\ 0 & a^{-1/2}\end{pmatrix}(x-b)\right|\right)^{-P} \\ &\leq Ca^{-3/4}\left(1 + \left|\left|\begin{pmatrix} a^{-1} & sa^{-1} \\ 0 & a^{-1/2}\end{pmatrix}\right|\right|^{-1} |x-b|\right)^{-P} \\ &\leq Ca^{-3/4}(1+C(s)a^{-1/2}|x-b|)^{-P} \\ &= O(a^{-3/4+P/2}|x-b|^{-P}) \end{aligned}$$

for $|x-b| > \delta$ and $C(s) = (1 + \frac{s^2}{2} + (s^2 + \frac{s^2}{4})^{1/2})^{1/2}$. We can now estimate

$$\begin{aligned} \langle (1-\Psi)f, \psi_{a,s,b}\rangle &\leq Ca^{-3/4+P/2}\int_{|x-b|\geq\delta} |x-b|^{-P}|1-\Psi(x)||f(x)|dx \\ &= O(a^{-3/2+P/2}) \end{aligned} \tag{3.4.11}$$

for $b \in U(b_0)$ implying (3.4.9). Now, let us assume that $f = \Psi f$ is localized, and let us estimate the shearlet coefficients $|\langle f, \psi_{a,s,b}\rangle|$. First note that the Fourier transform of $\psi_{a,s,b}$ is given by

$$\hat{\psi}_{a,s,b}(\xi) = a^{3/4}e^{-2\pi i t\xi}\hat{\psi}(a\xi_1, a^{1/2}(\xi_2 - s\xi_1)).$$

Now pick $\frac{1}{2} < \alpha < 1$ and write

$$\begin{aligned}|\langle f, \psi_{a,s,b}\rangle| =& |\langle \hat{f}, \hat{\psi}_{a,s,b}\rangle| \leq a^{3/4} \int_{\mathbb{R}^2} |\hat{f}(\xi_1, \xi_2)||\hat{\psi}(a\xi_1, a^{1/2}(\xi_2 - s\xi_1))d\xi \\ =& \underbrace{a^{3/4} \int_{|\xi_1|<a^{-\alpha}} |\hat{f}(\xi_1, \xi_2)||\hat{\psi}(a\xi_1, a^{1/2}(\xi_2 - s\xi_1))d\xi}_{A} \\ &+ \underbrace{a^{3/4} \int_{|\xi_1|>a^{-\alpha}} |\hat{f}(\xi_1, \xi_2)||\hat{\psi}(a\xi_1, a^{1/2}(\xi_2 - s\xi_1))d\xi}_{B}\end{aligned}$$

Since ψ possesses M moments in the x_1-direction which means that $\hat{\psi}(\xi_1, \xi_2) = \xi_1^M \hat{\theta}(\xi_1, \xi_2)$ with some $\theta \in L^2(\mathbb{R}^2)$, we can estimate A as

$$\begin{aligned} A &= a^{3/4} \int_{|\xi_1|<a^{-\alpha}} |\hat{f}(\xi_1, \xi_2)||\hat{\psi}(a\xi_1, a^{1/2}(\xi_2 - s\xi_1))|d\xi \\ &= a^{3/4} \int_{|\xi_1|<a^{-\alpha}} a^M |\xi_1|^M |\hat{f}(\xi_1, \xi_2)||\hat{\theta}(a\xi_1, a^{1/2}(\xi_2 - s\xi_1))|d\xi \\ &\leq a^{M(1-\alpha)} a^{3/4} \int_{|\xi_1|<a^{-\alpha}} |\hat{f}(\xi_1, \xi_2)||\hat{\theta}(a\xi_1, a^{1/2}(\xi_2 - s\xi_1))|d\xi \\ &\leq a^{(1-\alpha)M} \langle |\hat{f}|, |\hat{\theta}_{a,s,b}|\rangle \leq a^{(1-\alpha)M} ||\hat{f}||_2 ||\hat{\theta}_{a,s,b}||_2 = a^{(1-\alpha)M} ||f||_2 ||\theta||_2. \end{aligned} \tag{3.4.12}$$

To estimate B we will make the following substitution:

$$\begin{pmatrix} a & 0 \\ -a^{1/2} & a^{1/2} \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} = \begin{pmatrix} \tilde{\xi}_1 \\ \tilde{\xi}_2 \end{pmatrix}, \quad d\xi_1 d\xi_2 = a^{-3/2} d\tilde{\xi}_1 d\tilde{\xi}_2.$$

Therefore

$$B = a^{-3/4} \int_{\frac{|\tilde{\xi}_1|}{a} > a^{-\alpha}} |\hat{f}(\frac{\tilde{\xi}_1}{a}, \frac{s}{a}\tilde{\xi}_1 + a^{-1/2}\tilde{\xi}_2)||\hat{\psi}(\tilde{\xi}_1, \tilde{\xi}_2)|d\xi. \tag{3.4.13}$$

Now we can use the fact that $(b_0, \xi(s_0))$ is a regular directed point of f. This means that there is a neighborhood $(s_0 - \epsilon, s_0 + \epsilon)$ such that

$$\hat{f}(\eta_1, \eta_2) \leq C(1 + |\eta|)^{-N} \quad \text{for all} \quad \frac{\eta_2}{\eta_1} \in (s_0 - \epsilon, s_0 + \epsilon). \tag{3.4.14}$$

By using Equation 3.4.13 and considering $\frac{\eta_2}{\eta_1}$ with $\eta_1 := \frac{\tilde{\xi}_1}{a}$, $\eta_2 := \frac{s}{a}\tilde{\xi}_1 + a^{-1/2}\tilde{\xi}_2$ and $\frac{\tilde{\xi}_1}{a} > a^{-\alpha}$ we have that

$$s - a^{a-1/2}\tilde{\xi}_2 \leq \frac{\eta_2}{\eta_1} = s + a^{-1/2}\tilde{\xi}_2 \frac{a}{\tilde{\xi}_1} \leq s + a^{\alpha-1/2}\tilde{\xi}_2. \tag{3.4.15}$$

Equation (3.4.14) implies that

$$\left|\hat{f}\left(\frac{\tilde{\xi}_1}{a}, \frac{s}{a}\tilde{\xi}_1 + a^{-1/2}\tilde{\xi}_2\right)\right| \leq C\left(1 + \frac{|\tilde{\xi}_1|}{a}\right)^{-N}. \tag{3.4.16}$$

for s in a neighborhood $V(s_0)$ of s_0, $\frac{|\tilde{\xi}_1|}{a} > a^{-\alpha}$ and $|\tilde{\xi}_2| < \epsilon' a^{1/2-\alpha}$ for some $\epsilon' < \epsilon$. Now we first split the integral B according to

$$\begin{aligned} B =& a^{-3/4}\int_{|\tilde{\xi}_1|/a \geq a^{-\alpha}} \left|\hat{f}\left(\frac{\tilde{\xi}_1}{a}, \frac{s}{a}\tilde{\xi}_1 + a^{-1/2}\tilde{\xi}_2\right)\right| |\hat{\psi}(\tilde{\xi}_1, \tilde{\xi}_2) d\tilde{\xi}_1 d\tilde{\xi}_2| \\ =& \underbrace{a^{-3/4}\int_{|\tilde{\xi}_1|/a \geq a^{-\alpha}, |\tilde{\xi}_2| < \epsilon' a^{1/2-\alpha}} \left(\frac{\tilde{\xi}_1}{a}, \frac{s}{a}\tilde{\xi}_1 + a^{-1/2}\tilde{\xi}_2\right)\right| |\hat{\psi}(\tilde{\xi}_1, \tilde{\xi}_2) d\tilde{\xi}_1 d\tilde{\xi}_2|}_{B_1} \\ &+ \underbrace{a^{-3/4}\int_{|\tilde{\xi}_1|/a \geq a^{-\alpha}, |\tilde{\xi}_2| > \epsilon' a^{1/2-\alpha}} \left(\frac{\tilde{\xi}_1}{a}, \frac{s}{a}\tilde{\xi}_1 + a^{-1/2}\tilde{\xi}_2\right)\right| |\hat{\psi}(\tilde{\xi}_1, \tilde{\xi}_2) d\tilde{\xi}_1 d\tilde{\xi}_2|}_{B_2} \end{aligned} \tag{3.4.17}$$

By (3.4.16) we can estimate B_1 according to

$$B_1 \leq C a^{aN-3/4} ||\hat{\psi}||_1. \tag{3.4.18}$$

It only remains to compute B_2. For this we will use the fact that $\frac{\partial^L}{\partial x_2^L}\psi \in L^2(\mathbb{R}^2)$. This implies finally that

$$\begin{aligned} B_2 &\leq a^{-3/4}\int_{|\tilde{\xi}_1|/a \geq a^{-\alpha}, ||\tilde{\xi}_2| > \epsilon' a^{1/2-\alpha}} |\hat{f}(\tilde{\xi}_1/a, \frac{a}{s}\tilde{\xi}_1 + a^{-1/2}\tilde{\xi}_2)\hat{\psi}(\tilde{\xi}_1, \tilde{\xi}_2)| d\tilde{\xi}_1 d\tilde{\xi}_2 \\ &= a^{-3/4}\int_{|\tilde{\xi}_1|/a \geq a^{-\alpha}, ||\tilde{\xi}_2| > \epsilon' a^{1/2-\alpha}} |\hat{f}(\tilde{\xi}_1/a, \frac{a}{s}\tilde{\xi}_1 + a^{-1/2}\tilde{\xi}_2)\tilde{\xi}_2^{-L}\widehat{(\frac{\partial^L}{\partial x_2^L}\psi)}(\tilde{\xi}_1, \tilde{\xi}_2)| d\tilde{\xi}_1 d\tilde{\xi}_2 \\ &\leq (\epsilon')^{-L} a^{-3/4+(\alpha-1/2)L}\int_{\mathbb{R}^2} |\hat{f}(\tilde{\xi}_1/a, \frac{a}{s}\tilde{\xi}_1 + a^{-1/2}\tilde{\xi}_2)||\widehat{(\frac{\partial^L}{\partial x_2^L}\psi)}(\tilde{\xi}_1, \tilde{\xi}_2)| d\tilde{\xi}_1 d\tilde{\xi}_2 \\ &= (\epsilon')^{-L} a^{(\alpha-1/2)L} |\langle|\hat{f}|, |\widehat{(\frac{\partial^L}{\partial x_2^L}\psi_{a,s,b})}|\rangle| \leq (\epsilon')^{-L} a^{(\alpha-1/2)L} ||f||_2 ||\frac{\partial^L}{\partial x_2^L}\psi||. \end{aligned} \tag{3.4.19}$$

Combining the estimates (3.4.11), (3.4.12), (3.4.18) and (3.4.19) we get the desired estimate (3.4.8). □

This shows one of the directions of the inclusions of Theorem 3.4.9, i.e. $\mathcal{D} \subseteq \mathrm{WF}(f)^c$. To prove the other direction of the inclusion, the inverse theorem, we need to show that if the shearlet coefficients of a function around $(b_0, \xi(s_0)) \in \mathbb{R}^2 \times \mathbb{S}^1$ decay sufficiently fast when $a \to 0$, then $(b_0, \xi(s_0))$ is a regular directed point.

Following the procedure presented in [48] we need to introduce some preliminary localization results. Let us first show that a frequency projection on a conical set $\mathcal{C}_{u,v}$ preserves the wavefront set, where

$$\mathcal{C}_{u,v} := \{\xi \in \mathbb{R}^2 : |\xi_1| \geq u, |\xi_2| \leq v|\xi_1|\}.$$

Theorem 3.4.11 ([48]). *Let $s_0 < v$, then the next two statements are equivalent:*

(i) The point $(b_0, \xi(s_0)) \in \mathbb{R}^2 \times \mathbb{S}^1$ is an $N-$regular directed point of $g \in L^2(\mathbb{R}^2)$.

(ii) $(b_0, \xi(s_0))$ is an $N-$regular directed point of $\mathcal{P}_{\mathcal{C}_{u,v}}$.

Proof. Let us first show the (i) ⇐ (ii) part. Write $\mathcal{P}_{\mathcal{C}} g = g - \mathcal{P}_{\mathcal{C}^c_{u,v}} g$, where $\mathcal{P}_{\mathcal{C}^c_{u,v}}$ is the orthogonal projection onto the (closure of the) complement of $\mathcal{C}_{u,v}$. By assumption there exists a cutoff function function Ψ supported around b_0 such that

$$\widehat{(\Psi g)}(\xi) = O(|\xi|^{-N}) \quad \text{for all} \quad \xi_2/\xi_1 \in (s_0 - \delta, s_0 + \delta)$$

for some $\delta > 0$. Clearly, since $s_0 \in (-v, v)$ the point $(b_0, \xi(s_0))$ is an $N-$regular point of $\mathcal{P}_{\mathcal{C}^c_{u,v}} g$. Therefore the same estimate as above also holds for $(\widehat{\mathcal{P}_{\mathcal{C}^c_{u,v}} g})$. Using Lemma 2.2 of [48] the same estimate holds also for $(\widehat{\Psi \mathcal{P}_{\mathcal{C}^c_{u,v}} g})$, and therefore an analogous estimate holds for $(\widehat{\Psi \mathcal{P}_{\mathcal{C}_{u,v}} g})$. This proves the (⇐) part.

To prove the (i)⇒(ii) part we estimate $(\widehat{\Psi \mathcal{P}_{\mathcal{C}^c_{u,v}} g})$ using the method presented in the proof of Lemma 2.2 of [48] and notice that it is negligible for the decay properties of $\widehat{(\Psi g)}$ restricted to a small cone around the line with slope s_0. □

For the next result, let us recall the definition of the $N-$th fractional derivative of a function $f \in L^4 \infty(\mathbb{R})$ given by

$$f^{(N)}(x) := \left(\frac{\partial}{\partial x}\right)^N f(x) := (\omega^N \hat{I}(\omega))^\vee(u) \quad \text{for } N \in \mathbb{R}.$$

The next lemma states some well-known results for fractional derivatives with $N \notin \mathbb{N}$.

Lemma 3.4.12 ([48]). *Let $f \in L^4(\mathbb{R})$, then*

$$(f(\cdot/a))^{(N)}(x) = a^{-N} f^{(N)}(x/a) \tag{3.4.20}$$

and

$$||(fg)^{(N)}||_2 \leq C(||f^{(N)}||_4 ||g||_4 + ||f||_4 ||g^{(N)}||_4) \quad N < 1. \tag{3.4.21}$$

In the next lemma, we show that when studying the regularity of f around b_0 only the shearlet coefficients of f around b_0 are relevant.

Lemma 3.4.13 ([48])**.** *Let $f \in L^2(\mathbb{R}^2)$, Ψ be a smooth bump function supported in a small neighborhood $V(b_0)$ of some $b_0 \in \mathbb{R}^2$ and let $U(b_0)$ be another neighborhood of b_0 with $V(b_0) + \delta B \subset U(b_0)$ for some $\delta > 0$. Here, B denotes the unit disc in $\mathbb{R}^2$ and $+$ denotes the Minkowki sum of two sets, i.e.*

$$A + B = \{a + b : \textit{ for } a \in A \textit{ and } b \in B\}.$$

Consider the function $g : \mathbb{R}^2 \to \mathbb{R}$, given by

$$g(x) = \int_{b \in U(b_0)^c, s \in [-\Xi, \Xi], a \in [0, \Gamma]} \langle f, \tilde{\psi}_{a,s,b} \rangle \Psi(x) \psi_{a,s,b}(x) a^{-3} dadsdt \tag{3.4.22}$$

Then for all u, v

$$\hat{g}(\xi) = O(|\xi|^{-N}), \quad |\xi| < \infty, \quad \textit{for } \xi \in \mathcal{C}_{u,v} \tag{3.4.23}$$

provided that

$$\theta^j(x) := \left(\frac{\partial}{\partial x_1}\right)^j \psi(x) = O(|x|^{-P_j}) \quad \textit{with } P_j/2 - 3/4 > j + 2, \quad j = 0, \ldots, N. \tag{3.4.24}$$

Proof. Consider the Radon transform in the special form:

$$I(u) := \int_{\mathbb{R}} g(u - sx_2, x_2) dx_2, \quad |s| \leq v.$$

We will show that $I^{(N)} \in L^1(\mathbb{R})$ which implies that

$$\widehat{(I^{(N)})}(\omega) = \omega^N \hat{I}(\omega), \quad \hat{I} \in L^\infty(\mathbb{R}),$$

where $\omega \in \mathbb{R}$ is in the frequency space. Using the projection slice theorem (Theorem 2.4.1) with $\xi = \omega(1, s)$ we obtain

$$|\hat{g}(\xi)| = |\hat{I}(\omega)| \leq ||I^{(N)}||_{L^1(\mathbb{R})} |\omega|^{-N} \leq ||I^{(N)}||_{L^1(\mathbb{R})} \sqrt{1 + s^2} |\xi|^{-N},$$

which proves the statement. Now, let us show that $I^{(N)} \in L^1(\mathbb{R})$. Since I is of compact support, for this, we only need to show that $I^{(N)}$ is bounded. We have

$$\begin{aligned}
I^{(N)}(u) &= \int_{\mathbb{R}^2} \int_{\mathbb{R}} \int_{\mathbb{R}_+} \langle f, \tilde{\psi}_{a,s,b} \rangle (\frac{\partial}{\partial u})^N \int_{\mathbb{R}} \Psi(u - sx, x) \psi_{a,s,b}(u - sx, x) dx d\mu(a, s, b) \\
&= \sum_{j=0}^{N} \binom{N}{j} \int_{\mathbb{R}^2} \int_{\mathbb{R}} \int_{\mathbb{R}_+} \langle f, \tilde{\psi}_{a,s,b} \rangle \int_{\mathbb{R}} (\frac{\partial}{\partial u})^{N-j} \Psi(u - sx, x) (\frac{\partial}{\partial u})^j \psi_{a,s,b}(u - sx, x) dx d\mu(a, s, b) \\
&= \sum_{j=0}^{N} \binom{N}{j} \int_{\mathbb{R}^2} \int_{\mathbb{R}} \int_{\mathbb{R}_+} \langle f, \tilde{\psi}_{a,s,b} \rangle a^{-j} \int_{\mathbb{R}} (\frac{\partial}{\partial x_1})^{N-j} \Psi(u - sx, x) \theta^j_{a,s,b}(u - sx, x) dx d\mu(a, s, b),
\end{aligned} \tag{3.4.25}$$

where $\theta^j = (\frac{\partial}{\partial x_1})^j \psi$. Using a similar argument as in the localization part at the beginning of the proof of Theorem 3.4.10, we obtain

$$|\theta^{j}_{a,s,b}(x)| = O(a^{-3/4+P_j/2}|x-b|^{-P_j}), \quad \text{as } a \to 0.$$

Since $(\frac{\partial}{\partial x_1})^{N-j}\psi\theta^j$ has small support around b_0 and the parameter b varies in a set far away from the support of $V(b_0)$ and $(\frac{\partial}{\partial x_1})^{N-j}\Psi\theta^j$, we can estimate

$$\left|\left(\frac{\partial}{\partial x_1}\right)^{N-h}\Psi(x)\theta^{j}_{a,s,b}(x)\right| = O\left(\left|\left(\frac{\partial}{\partial x_1}\right)^{N-j}\Psi(x)\right|a^{-3/4+P_j/2}|b-b_0|^{-P_j}\right), \quad \text{as } a \to 0$$

for $b \in U(b_0)^c$. By plugging the estimate above in (3.4.25) and using (3.4.23) we obtain the desired result. □

We still need to introduce two important results before introducing the inverse theorem. The first result describes how smoothness and the vanishing moments conditions have to interact with the shearlet transform to get fast decay rates. The second result describes the conditions needed for the shearlet system to form a tight frame for $L^2(\mathcal{C}_{u,v})^\vee$, where

$$L^2(\mathcal{C}_{u,v})^\vee := \{f \in L^2(\mathbb{R}^2) : \mathrm{supp}(f) \subset \mathcal{C}_{u,v}\}. \tag{3.4.26}$$

Lemma 3.4.14 ([48]). *Let $W : \mathbb{R}^2 \to R$ be given by*

$$\Delta_{u,v}(\psi)(\xi) + |\hat{W}(\xi)|^2 = C_\psi \chi_{\mathcal{C}_{u,v}}(\xi) \tag{3.4.27}$$

where $\Delta_{u,v}(\psi)(\xi) = \partial^2_{\xi_1}\psi(\xi) + \partial^2_{\xi_2}\psi(\xi)$. Assume that $\Xi > v$, $u \geq 0$ and that $\psi = \frac{\partial^M}{\partial x_1^M}\theta$ has M vanishing moments, Fourier decay of order L_1 in the first variable (the Fourier transform on the x_1-direction decays accordingly) and that θ has Fourier decay of order L_2 in the second variable such that

$$2M - 1/2 > L_2 > M > 1/2. \tag{3.4.28}$$

Then

$$|\hat{W}(\xi)|^2 = O(|\xi|^{-2\min(L_1, L_2-M)}), \quad \text{as } \xi \to \infty.$$

In particular, if ψ is sufficiently smooth and has sufficiently many vanishing, moments then W is a smooth function.

We can now prove the next theorem:

Theorem 3.4.15 ([48]). *With the assumptions of Lemma 3.4.14 and W defined as in (3.4.27), the system*

$$(P_{\mathcal{C}_{u,v}}\psi_{a,s,b})_{a\in[0,\Gamma], s\in[-\Xi,\Xi], b\in\mathbb{R}^2} \cup (T_b P_{\mathcal{C}_{u,v}} W)_{b\in\mathbb{R}^2}$$

constitutes a tight frame for $L^2(\mathcal{C}_{u,v})^\vee$ with frame constant C_ψ. We have the representation

$$\begin{aligned} f(x) &= \frac{1}{C_\psi}\int_{\mathbb{R}^2}\langle f, T_bW\rangle T_bP_{\mathcal{C}_{u,v}}Wdb \\ &+ \frac{1}{C_\psi}\int_{b\in\mathbb{R}^2}\int_{s\in[-\Xi,\Xi]}\int_{a\in[0,\Gamma]}\mathcal{SH}^h_\psi f(a,s,b)P_{\mathcal{C}_{u,v}}\psi_{a,s,b}(x)a^{-3}dadsdb \end{aligned} \tag{3.4.29}$$

for $x\in\mathbb{R}^2$. The window function W satisfies the Fourier decay estimates from Lemma 3.4.14.

Proof. The frame operator is given as the Fourier multiplier with the function $\Delta_{u,v}(\psi)(\xi)+\xi_{\mathcal{C}_{u,v}}(\xi)|\hat{W}|^2=\chi_{\mathcal{C}_{u,v}}(\xi)C_\psi$. A Fourier multiplier is an operator that alters the Fourier transform of a function by multiplying it against another function, the multiplier or symbol. Let Id be the identity mapping, then it follows that the frame operator is given by $C_\psi P_{\mathcal{C}_{u,v}}=C_\psi Id$ on $L^2(\mathcal{C}_{u,v})$ (where Id is the identity). □

We have now all the needed preliminary results to prove the inverse theorem. As in Lemma 3.4.13, we are going to prove the result for $N\in\mathbb{N}$, the generalization for $N\in\mathbb{R}$ is achieved via Lemma 3.4.12.

Theorem 3.4.16 (Inverse theorem, [48])**.** *Let $f\in L^2(\mathbb{R}^2)$ be such that $\hat{f}\in L^2(C_{u,v})$, with $0<u,v<\infty$. Assume that there exist neighborhoods $U(b_0)\subset\mathbb{R}^2$ of b_0 and $(s_0-\epsilon,s_0+\epsilon)\subset[-s_0,s_0]$ of s_0 such that*

$$\mathcal{SH}^h_\psi f(a,s,t)=O(a^K)\quad \textit{for all}\quad (b,s)\in U(b_0)\times[-s_0,s_0],\quad \textit{as } a\to 0 \tag{3.4.30}$$

with the implied constant uniform over b and s. Then $(b_0,\xi(s_0))$ is an $N-$regular directed point of f for all N with (3.4.24) such that $\psi\in H_{(N,L)}(\mathbb{R}^2)$, and $\hat{\theta}^j,\omega_1^{-M}\hat{\psi}(\omega),(\widehat{\frac{\partial^L}{\partial x_2^L}\theta^j})\in L^1(\mathbb{R}^2)$, $j=0,\dots,N$, and for some $1/2<\alpha<1$,

$$N+2<\min(K-3/4,(1-\alpha)(M+N)-\frac{3}{4},(\alpha-1/2)L-\frac{3}{4},2(L_2-M+1),2(L_1+1)), \tag{3.4.31}$$

where $M>1$ is the number of vanishing moments of ψ, L is the Fourier decay of ψ in the second coordinate and L_1,L_2 are defined as in Lemma 3.4.14 such that (3.4.28) holds.

Proof. Let us choose Γ,Ξ such that the system

$$(P_{\mathcal{C}_{u,v+\kappa}}\psi_{a,s,b})_{a\in[0,\Gamma],s\in[-\Xi,\Xi],b\in\mathbb{R}^2}\cup(T_bP_{\mathcal{C}_{u,v+\kappa}}W)_{b\in\mathbb{R}^2}$$

is a tight frame (see Definition 3.2.3) for $L^2(\mathcal{C}_{u,v+\kappa})$ and $v+\kappa>s_0$, with W chosen according to Lemma 3.4.14. We are going to prove that for a localized version of $\tilde{f}$ of

$$g=\int_{b\in\mathbb{R}^2,s\in(-\Xi,\Xi),a\in(0,\Gamma)}\langle f,\psi_{a,s,b}\rangle\psi_{a,s,b}a^{-3}dadsdb \tag{3.4.32}$$

around b_0 the Fourier transform of the $I(u) := \mathcal{R}\tilde{f}(u, b_0)$ decays with order $|\omega|^{-N}$ for $|\omega| \to \infty$. By the projection slice theorem this would prove that $(b_0, \xi(s_0))$ is a regular directed point of g. To show that this already implies that $(b_0, \xi(s_0))$ is a regular directed point of f, we argue as follows: By Theorem 3.4.15 we have the representation

$$f = \frac{1}{C_\psi} P_{\mathcal{C}_{u,v+\kappa}}(g + \int_{b\in\mathbb{R}^2} \langle f, T_b W\rangle T_b W db).$$

It follows from Theorem 3.4.11 that $(b_0, \xi(s_0))$ is a $N-$regular directed point of f if it is an $N-$regular directed point of $g + \int_{b\in\mathbb{R}^2} \langle f, T_b W\rangle T_b W f b$. By Lemma 3.4.14 and (3.4.31), $(b_0, \xi(s_0))$ is a $N-$regular point of $\int_{b\in\mathbb{R}^2} \langle f, T_b W\rangle T_b W db$, and therefore we only need to verify regularity for g, which we will now do.

First note that by Lemma 3.4.13 we can without loss of generality restrict the parameter b in the integral (3.4.32) to $U(b_0)$ if we multiply by a suitable cutoff function Ψ. Therefore we need to study the regularity properties of

$$\tilde{f} = \int_{b\in U(b_0), s\in(-\Xi,\Xi), a\in(0,\Gamma)} \langle f, \psi_{a,s,b}\rangle \Psi \psi_{a,s,b} a^{-3} da ds db,$$

where Ψ is supported in a small neighborhood $V_0(s_0) \subset U(b_0)$ around b_0. Let us denote by $I(u)$ the function

$$I(u) := \tilde{\mathcal{R}}\tilde{f}(u, s_0)$$

with

$$\tilde{\mathcal{R}}\tilde{f}(u, s_0) = \int_{b\in U(b_0), s\in(-\Xi,\Xi), a\in(0,\Gamma)} \langle f, \psi_{a,s,b}\rangle \tilde{\mathcal{R}}\Psi\psi_{a,s,b}(u, s_0) a^{-3} da ds db,$$

and

$$\tilde{\mathcal{R}}\Psi\psi_{a,s,b}(u, s_0) = a^{-3/4} \int_{\mathbb{R}} \Psi(u - s_0 x_2, x_2)\psi(\frac{u - s_0 x_2 - b_1 + s_0(x_2 - t_2)}{a}, \frac{x_2}{a^{1/2}}) dx_2.$$

To prove our goal that $\hat{I}(\omega) = O(|\omega|^{-N})$ we need to show that $\omega^N \hat{I}(\omega) \in L^\infty(\mathbb{R})$ or the stronger statement, that the fractional derivative $I^{(N)}$ of I defined by

$$I^{(N)}(u) := (\frac{\partial}{\partial u})^N I(u) := (\omega^N \hat{I}(\omega))^\vee(u)$$

is in $L^1(\mathbb{R})$.

Unless stated otherwise in what follows the variables a, s, b are allowed to vary over the sets $[0, \Gamma], [-\Xi, \Xi]$ and $U(b_0)$, respectively. Using the product rule and the definition of $\tilde{\mathcal{R}}$ the quantity $||I^{(N)}||_1$ can be estimated by

$$C \max_{j=0,\ldots,N} \int_{\mathbb{R}} \int_{a,s,b} |\langle f, \psi_{a,s,b}\rangle| a^{-j} \tilde{\mathcal{R}}(\frac{\partial^{N-j}}{\partial x_1^{N-j}} \Psi \frac{\partial^j}{\partial x_1^j} \psi_{a,s,b})(u, s_0)| d\mu(a, s, b) du.$$

We only treat the case $j = N$, the other cases can be done analogously.

Assume first that the function Ψ is zero outside a cube of sidelength η around b_0. Then the support of I^N is contained in the interval $I_U = [(b_0)_1 - s_0(b_0)_2 - 2\eta, (b_0)_1 - s_0(b_0)_2 - 2\eta]$ and the integration variable x_2 from the definition of $\tilde{\mathcal{R}}$ can be restricted to the interval $I_X = [(b_0)_2 - \eta, (b_0)_2 + \eta]$, where $b_0 = ((b_0)_1, (b_0)_2)$. We now separate the integral as follows

$$A + B := \int_{I_U} \int_{\mathbb{R}^2} \int_{\mathbb{R}} \int_{\mathbb{R}_+} |\langle f, \psi_{a,s,b} \rangle| a^{-N} |\tilde{\mathcal{R}}(\Psi \frac{\partial^N}{\partial x_1^N} \psi_{a,s,b})(u, s_0)| d\mu(a, s, b) du$$

where

$$A = \int_{I_U} \int_{\mathbb{R}^2} \int_{s_0-\epsilon}^{s_0+\epsilon} \int_{\mathbb{R}_+} |\langle f, \psi_{a,s,b} \rangle| a^{-N} |\tilde{\mathcal{R}}(\psi \frac{\partial^N}{\partial x_1^N} \psi_{a,s,b})(u, s_0)| d\mu(a, s, b) du$$

and

$$B = \int_{I_U} \int_{\mathbb{R}^2} \int_{\setminus(s_0-\epsilon, s_0+\epsilon), b} \int_{\mathbb{R}_+} |\langle f, \psi_{a,s,b} \rangle| a^{-N} |\tilde{\mathcal{R}}(\Psi \frac{\partial^N}{\partial x_1^n} \psi_{a,s,b})(u, s_0)| d\mu(a, s, b) du$$

In order to estimate B we note that

$$\tilde{\mathcal{R}} \left(\Psi \frac{\partial^N}{\partial x_1^N} \psi_{a,s,b} \right) (u, s_0) = \int_{I_X} \Psi \frac{\partial^N}{\partial x_1^N} \psi_a, s, b(u - s_0 x_2, x_2) dx_2 = \mathcal{SH}_\theta^h \Psi \delta_{x_1 + s_0 x_2 - u}(a, s, b).$$

It is well known and easy to show that for $s \in (-\Xi, \Xi) \setminus [s_0 - \epsilon, s_0 + \epsilon]$, the point $(b, \xi(s))$ is and $R-$ regular directed point of $\delta_{x_1 + s_0 x_2 - u}$, therefore it is also an $R-$regular directed point of the localized version $\Psi \delta_{x_1 + s_0 x_2 - u}$. By using the same arguments as in the proof of the Theorem 3.4.10 we see that for any $1/2 < \alpha < 1$ the estimate

$$\mathcal{SH}_\theta^h \Psi \delta_{x_1 + s_0 x_2 - u}(a, s, b) = O(a^{(1-\alpha)(M+N) - 3/4} + a^{-3/4 + (\alpha - 1/2)L}) \tag{3.4.33}$$

holds with the implied constant uniform over $b \in U(b_0), s \in (-\Xi, \Xi) \setminus [s_0 - \epsilon, s_0 + \epsilon]$. The details of these can be found in the appendix of [48]. Since by assumption there exists $1/2 < \alpha < 1$ such that

$$N + 2 < \min((1 - \alpha)(M + N) - 3/4, (\alpha - 1/2)L - 3/4),$$

the expression B is bounded. In order to estimate A we use the fast decay of the shearlet coefficients of f around (b_0, s_0). By our assumptions on N, K, the coefficients $\langle f, \psi_{a,s,b} \rangle$ decay of order greater than $a^{N+2+3/4}$, and therefore A is bounded. This finishes with the proof of the inverse theorem. □

Due to the technicality of Theorem 3.4.16, in order to get some intuition we are motivated to present a more informal version in the following:

Corollary 3.4.17 (Inverse theorem, informal version)**.** *Assume that ψ has sufficiently many vanishing moments in the x_1-direction, is sufficiently smooth and sufficiently well-localized in space. Assume further that* (3.4.30) *holds for some f. Then (b_0, s_0) is an $N-$regular directed point of f for all $N < K - 11/4$.*

Having the direct and inverse theorem, we are ready to combine them and introduce our extended version of the shearlet-based resolution of the wavefront set.

Theorem 3.4.18 (Resolution of the Wavefront set II, [48])**.** *Let $f \in L^2(\mathbb{R}^2)$, $N \in \mathbb{R}$ and $\epsilon > 0$. Then there exist P, M, L, L_1, L_2 such that for all functions $\psi \in H_{(N,0)}(\mathbb{R}^2)$ with M vanishing moments in x_1-direction, decay of order P towards infinity, C^L regularity in the second coordinate and L_1, L_2 as in Lemma 3.4.14 the following holds: Set $\mathcal{D} = \mathcal{D}_1 \cup \mathcal{D}_2$, where*

$$\begin{aligned}\mathcal{D}_1 :=&\{(b_0, \xi(s_0)) : (b_0, s_0) \in \mathbb{R}^2 \times [-1, 1] \text{ and } (b, s) \text{ in a neighborhood } U \text{ of } (b_0, s_0),\\ &|\mathcal{SH}^v_\psi f(a, s, b)| = O(a^k) \text{ as } a \to 0, \text{ for all } k \in \mathbb{N}, \text{ with the } O(\cdot)\text{-terms uniform}\\ &\text{over } (b, s) \in U\}.\end{aligned}$$

and

$$\begin{aligned}\mathcal{D}_2 :=&\{(b_0, \xi(s_0)) : (b_0, s_0) \in \mathbb{R}^2 \times [1, \infty] \text{ and } (b, 1/s) \text{ in a neighborhood } U \text{ of } (b_0, s_0),\\ &|\mathcal{SH}^h_\psi f(a, s, b)| = O(a^k) \text{ as } a \to 0, \text{ for all } k \in \mathbb{N}, \text{ with the } O(\cdot)\text{-terms uniform}\\ &\text{over } (b, 1/s) \in U\}.\end{aligned}$$

Then

$$\mathrm{WF}^{N+3/4+\epsilon}(f)^c \subseteq \mathcal{D} \subseteq \mathrm{WF}^{N-11/4-\epsilon}(f)^c. \tag{3.4.34}$$

The precise values of P, M, L, L_1, L_2 are given in Theorems 3.4.10 and 3.4.16.

Proof. First, we need to show that if

$$|\mathcal{SH}^h_\psi f(a, s, b)| = O(a^N), \quad \text{as } a \to 0$$

for $|s| \leq 1$, then

$$|\mathcal{SH}^h_\psi f(a, s, b)| = O(a^N), \quad \text{as } a \to 0$$

with some suitable cone with $v > 1$ and P, L, M large enough. For this, we can estimate the integral

$$\int_{\mathcal{C}^c_{u,v}} \hat{f}(\xi)\overline{\hat{\psi}_{a,s,b}}(\xi)d\xi = O(a^N)$$

using the same estimates (3.4.14) to (3.4.16) in the proof of Theorem 3.4.10. This shows that $|\mathcal{SH}^h_\psi P_{\mathcal{C}_{u,v}} f(a, s, b)| = O(a^N)$. The reverse implication can also be shown using the same argument. We then obtain:

$$|\mathcal{SH}^h_\psi f(a,s,b)| = O(a^N) \quad \text{for} \quad |s| \leq 1 \Longleftrightarrow |\mathcal{SH}^h_\psi P_{\mathcal{C}_{u,v}} f(a,s,b)| = O(a^N)$$

as $a \to 0$, for a cone $C_{u,v}$ with $v > 1$ and P, L, M large enough. The case $s > 1$ is similar. We also need the fact that for $s \leq 1$ the point $(b, \xi(s))$ is an $N-$regular directed point of $P_{\mathcal{C}_{u,v}} f$ if and only if $(b, \xi(s))$ is an $N-$regular directed point of f (see Theorem 3.4.11). Finally, the statement follows directly from Theorems 3.4.10 and 3.4.16. □

As one can observe, the results presented in Theorem 3.4.18 are applied to the wavefront set of finite degree N, $\mathrm{WF}^N(f)$. The results also hold for the full wavefront set, given by

$$\mathrm{WF}(f) = \bigcup_{N=1}^{\infty} WF^N(f).$$

In order to finish this chapter with the full wavefront set resolution in the desired form we will present two last theorems. For the next results we will extend the notion of the shearlet base wavefront set resolution from $L^2(\mathbb{R}^2)$ to the space of tempered distributions $\mathcal{S}'(\mathbb{R}^2)$.

Remark 3.4.19. *We can extend the results from $L^2(\mathbb{R}^2)$ to the space of tempered distributions $\mathcal{S}'(\mathbb{R}^2)$ by simply taking shearlet generating functions in $\mathcal{S}(\mathbb{R}^2)$ and using duality as we did with the Radon transform (see Remark 2.3.4).*

Theorem 3.4.20 ([48])**.** *Assume that $\psi \in \mathcal{S}(\mathbb{R}^2)$ is a Schwartz test function (see Section 2.1) with infinitely many vanishing moments in x_1-direction. Then*

$$\mathrm{WF}(f) = \{(b, \xi(s)) \subset \mathbb{R}^2 \times \mathbb{S}^1 : \mathcal{SH}^h_\psi f(a,s,b) \textit{ does not decay rapidly around } (b,s)\}$$

for any tempered distribution $f \in \mathcal{S}'(\mathbb{R}^2)$ (see Section 2.1) with frequency support in $\mathcal{C}_{u,v}$ for $u, v > 0$.

Proof. We have already proven this result for $f \in L^2(\mathcal{C}_{u,v})^\vee$. Since ψ is a test function, the generalization to tempered distributions follows easily by just using the same arguments. □

Finally, the following theorem is an extension of Theorem 3.4.2 for weaker assumptions over ψ.

Theorem 3.4.21 (Resolution of the Wavefront set III, [48])**.** *Let $\psi \in \mathcal{S}(\mathbb{R}^2)$ be a Schwartz function with infinitely many vanishing moments in x_1-direction. Let $f \in \mathcal{S}'(\mathbb{R}^2)$ be a tempered distribution and $\mathcal{D} = \mathcal{D}_1 \cup \mathcal{D}_2$, where*

$$\begin{aligned}\mathcal{D}_1 := &\{(b_0, \xi(s_0)) : (b_0, s_0) \in \mathbb{R}^2 \times [-1,1] \textit{ and } (b,s) \textit{ in a neighborhood } U \textit{ of } (b_0,s_0),\\ &|\mathcal{SH}^v_\psi f(a,s,b)| = O(a^k) \textit{ as } a \to 0, \textit{ for all } k \in \mathbb{N}, \textit{ with the } O(\cdot)\textit{-terms uniform}\\ &\textit{over } (b,s) \in U\}\end{aligned}$$

and

$$\begin{aligned}\mathcal{D}_2 := \{&(b_0,\xi(s_0)) : (b_0,s_0)\in\mathbb{R}^2\times[1,\infty] \textit{ and } (b,1/s) \textit{ in a neighborhood } U \textit{ of } (b_0,s_0),\\ &|\mathcal{SH}^h_\psi f(a,s,b)| = O(a^k) \textit{ as } a\to 0, \textit{ for all } k\in\mathbb{N}, \textit{ with the } O(\cdot)\textit{-terms uniform}\\ &\textit{over } (b,1/s)\in U\}.\end{aligned}$$

Then

$$\mathrm{WF}(f)^c = \mathcal{D}. \tag{3.4.35}$$

Proof. This is an immediate consequence of Theorems 3.4.10 and 3.4.16 making the usual adaptations to handle general tempered distributions. □

We have now introduced the basic assumptions for the shearlet generators, in order to be able to perform the wavefront set resolution of tempered distributions. This allows us to have a specified procedure to localize and microlocalize functions. In addition, the weaker assumptions presented in Theorem 3.4.21 make it possible to use compactly supported shearlets as our generators.

As mentioned before, compactly supported shearlets have high-resolution properties in the spatial domain, making them suitable for tomographic reconstruction problems as well as other inverse problems. If we want to use the shearlet-based wavefront set resolution in real-world problems, we still need to translate the estimate 3.4.35 to the digital realm. This is not trivial, mainly due to the lack of notion of oriented singularities in the digital domain.

Remark 3.4.22. *Since we are working with compactly supported shearlets, we can restrict our attention to tempered distributions over a open domain $\Omega\subset\mathbb{R}^2$. This can be done using the same principle as in Definition 2.1.12, where Schwartz functions over Ω can be extended by 0 to all $\mathbb{R}^2$ resulting on a function in $\mathcal{S}(\mathbb{R}^2)$.*

Remark 3.4.22 we can modify Theorem 3.4.21 to detect wavefront sets in $\mathcal{S}'(\Omega)$.

Theorem 3.4.23 (Resolution of the Wavefront set IV). *Let $\Omega\subset\mathbb{R}^2$ be an open domain. In addition, let $\psi\in\mathcal{S}(\Omega)$ (see Definition 2.1.12) be a Schwartz function with infinitely many vanishing moments in x_1-direction. Let $f\in\mathcal{S}'(\Omega)$ be a tempered distribution and $\mathcal{D}=\mathcal{D}_1\cup\mathcal{D}_2$, where*

$$\begin{aligned}\mathcal{D}_1 := \{&(b_0,\xi(s_0)) : (b_0,s_0)\in\mathbb{R}^2\times[-1,1] \textit{ and } (b,s) \textit{ in a neighborhood } U \textit{ of } (b_0,s_0),\\ &|\mathcal{SH}^v_\psi f(a,s,b)| = O(a^k) \textit{ as } a\to 0, \textit{ for all } k\in\mathbb{N}, \textit{ with the } O(\cdot)\textit{-terms uniform}\\ &\textit{over } (b,s)\in U\}\end{aligned}$$

and

$$\begin{aligned}\mathcal{D}_2 := \{&(b_0,\xi(s_0)) : (b_0,s_0)\in\mathbb{R}^2\times[1,\infty] \textit{ and } (b,1/s) \textit{ in a neighborhood } U \textit{ of } (b_0,s_0),\\ &|\mathcal{SH}^h_\psi f(a,s,b)| = O(a^k) \textit{ as } a\to 0, \textit{ for all } k\in\mathbb{N}, \textit{ with the } O(\cdot)\textit{-terms uniform}\\ &\textit{over } (b,1/s)\in U\}.\end{aligned}$$

Then

$$\mathrm{WF}(f)^c = \mathcal{D}. \tag{3.4.36}$$

Proof. This is an immediate consequence of Theorems 3.4.10 and 3.4.16 making the usual adaptations to handle general tempered distributions. □

In Chapter 5 we will propose a novel method to use the wavefront set resolution capabilities of the compactly supported shearlet systems, combined with convolutional neural networks in order to perform digital wavefront set extraction. But before, we need to work in the continuous setting to study the microlocal behavior of continuous convolutional neural networks, this will allow us to use microlocal analysis to perform inverse problem regularization in tomographic reconstruction (Chapter 8) using the rich information provided by the wavefront set. For this, in the next chapter, we will introduce a novel approach to studying deep convolutional neural networks in the context of microlocal analysis by introducing the notion of continuous convolutional neural networks as non-linear operators on functional spaces.

4 Microlocal analysis of continuum convolutional residual neural networks

In this chapter we present the theory of microlocal analysis for deep convolutional neural networks in the continuum setting. In this setting, the neural network architecture is seen as an operator between functional (or distributional) spaces. Although this theory can be extended from Fourier integral operators to a variety of convolutional neural networks, we will focus on convolutional residual neural networks (conv-ResNets), due to their extensive applications to real-world problems and their simple and elegant mathematical structure.

In the context of the continuum setting, a convolutional neural network is interpreted as an operator between Hilbert spaces. In the following we will analyze the different components of residual convolutional neural networks separately, in particular, we will study the microlocal behavior of the convolutional operator, the residual layer and the pointwise ReLU operator. We refer as microlocal behavior of an operator to its action upon the singularities of functions (distributions). In the case of Fourier integral operators, such action is described by the microcanonical relation [70]. As we will show in the following sections, the continuum convolutional layers are pseudodifferential operators (see Definition 2.3.9), where the amplitude, also known as the symbol, can be explicitly computed, therefore the standard microlocal analysis presented in Chapter 2 can be applied. In the case of the ReLU activation function, which is a nonlinear operator, a nonstandard approach needs to be taken.

My own contribution: This chapter results from numerous discussions with my supervisor, Gitta Kutyniok, and collaborators Ozan Öktem and Philipp Petersen which were later published as [9]. The main ideas were developed by me but formalized and corrected with the help of my co-authors. In particular, Philipp Petersen was extremely helpful on the formalization of the notion of ReLU and the Heaviside function on tempered distributions. In addition, Ozan Öktem helped with the extension of the Radon transform and back-projection operators to tempered distributions. This work was the final result of my research throughout my PhD studies. The actual writing was moslty done by myself with the help of Philipp Petersen for the main results in form of theorems and lemmas.

4.1 Continuum convolutional residual neural networks

Before we discuss residual neural networks, we would like to refer the reader to Section 1.2 for the basic concepts of deep learning and deep convolutional neural networks. Residual neural networks have shown state-of-the-art performance in image classification, achieving first place on the ImageNet challenge first in [71], and more recently in [110]. The residual neural network architecture was first introduced in 2015 by He et al. [56] with the goal to ease the training of general convolutional neural networks. This architecture allows one to increase the depth of the networks without a significant computational cost in the training. He et al. explicitly reformulated the layers as learning residual functions with reference to the layer inputs, instead of learning unreferenced functions. This action is reflected in a skip connection of a residual block which combines the input and the output. On the inner weight layers, the Residual Linear Unit (ReLU) activation function is applied. This function is given by the element-wise application of the $\mathrm{ReLU}(x) = \max\{0, x\}$ function. Figure 4.1 depicts the basic block of a ResNet.

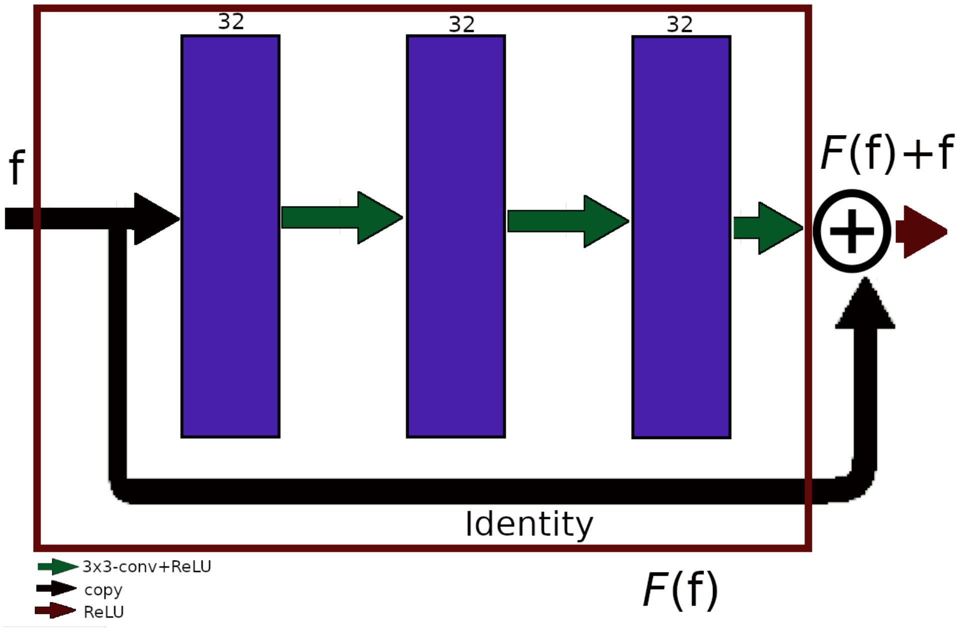

Figure 4.1: Illustration of the principal block in ResNet, namely the skip connection from the input to the output is the main characteristic of this architectures.

There are different reasons why residual representations are relevant in image regression and classification. Let us consider $\mathcal{H} : \mathcal{S}(\mathbb{R}^2) \to \mathcal{S}(\mathbb{R}^2)$ where $\mathcal{H}(f)$ is regarded as an underlying mapping to be learned by a few stacked layers, with $f \in \mathcal{S}(\mathbb{R}^2)$ denoting the input of the first of these layers. Later, in Equation (4.1.1) we will show an explicit example of such mapping. Let us hypothesize that multiple nonlinear layers can asymptotically approximate the complicated function $f \mapsto \mathcal{H}(f)$. Thus, the layers can asymptotically approximate the residual functions, $\mathcal{H}(f) - f$, if the input f and output $\mathcal{H}(f)$ are of the same dimension (see [56]). This means that rather than expecting stacked layers to approximate $\mathcal{H}(f)$, we explicitly let these layers approximate a residual function

$$\mathcal{F}(f) := \mathcal{H}(f) - f.$$

The original input function f thus becomes $\mathcal{F}(f) + f$. In their original work, He et al. [56] showed that if we have a shallow network and add layers, which can be constructed as identity mappings, a deeper model requires to have a training error not greater than its shallower version. The so-called degradation problem [56] suggests that the solvers might have difficulties in approximating identity mappings by multiple nonlinear layers. The ResNet formulation is also known as *residual learning*, if identity mappings are optimal, i.e., the solution is close to the identity, the residual is trained to be zero.

Although, it is unlikely that identity mappings are optimal, the residual reformulation might help to precondition the problem. If the optimal function is closer to an identity mapping than to a zero mapping, it should be easier for the optimizer to find the perturbations with reference to an identity mapping, than to learn the function as a new one. In other words, in each training step one does not need to learn the entire update, but only a small offset from the identity.

The aforementioned argument plays an important role in the design of architectures intended to solve inverse problems. A relevant example for our own purposes is the case of the learned primal-dual architecture introduced by Öktem and Adler in 2017 [3]. This architecture is based on a primal-dual iterative method [25] where the proximal operators (see Section 4.3) are learned. In this case, the proximal operators are typically close to the identity, therefore residual blocks are well suited to approximate them.

When working with images, as in our case, a common choice for the inner layers of the residual block (see Figure 4.1) are the convolutional layers. This architecture is traditionally expressed in the discrete setting, defined as follows.

Definition 4.1.1 (Discrete two-dimensional convolutional ResNet, [9, Definition 2.1])**.** *Consider matrices in* $\mathbb{R}^{n_1 \times n_2}$ *representing functions on* $\mathbb{R}^2$ *that are discretized at* $n_1 \times n_2$ *sample points. Next, let* $k_0, k_1, k_2, k_3, k_4 \in \mathbb{N}$ *denote the* numbers of channels per layer *with where* $k_4 = 1$*. Furthermore, let*

$$\boldsymbol{\theta}_j := (\boldsymbol{\theta}_j^{l,k})_{l=1,k=1}^{n_{j-1},n_j} \in (\mathbb{R}^{3\times 3})^{n_{j-1}\times n_{j-1}} \quad \textit{for } j = 1, \dots, 4$$

denote a set of set of filters and $b_j \in (\mathbb{R}^{n_1 \times n_2})^{k_j}$ *the channel-wise bias. We define the convolutional affine operator* $\mathcal{W}_{\boldsymbol{\theta}_j} : \left(\mathbb{R}^{N_1\times N_2}\right)^{n_{j-1}} \to \left(\mathbb{R}^{N_1 \times N_2}\right)^{n_j}$ *as*

$$\mathcal{W}_{\boldsymbol{\theta}_j, b_j}(\boldsymbol{f})(i_1, i_2, k) = b_j^k(i_1, i_2) + \sum_{l=1}^{n_{j-1}} (\boldsymbol{\theta}_j^{l,k} * \boldsymbol{f})(i_1, i_2) \quad \textit{for } k \in \{1, \dots, n_j\} \textit{ and } \quad \boldsymbol{f} \in \left(\mathbb{R}^{N_1 \times N_2}\right)^{n_{j-1}}. \tag{4.1.1}$$

The ResNet operator $\mathrm{ResNet} : (\mathbb{R}^{N_1 \times N_2})^{n_0} \to \mathbb{R}^{N_1 \times N_2}$ *is given by*

$$\mathrm{ResNet}(\boldsymbol{f}_1, \dots, \boldsymbol{f}_{n_0}) = \boldsymbol{f}_1 + \mathcal{F}(\boldsymbol{f}_1, \dots, \boldsymbol{f}_{n_0}) \quad \textit{for } \boldsymbol{f}_1, \dots, \boldsymbol{f}_{n_0} \in \mathbb{R}^{N_1 \times N_2},$$

where $\mathcal{F} \colon (\mathbb{R}^{N_1 \times N_2})^{n_0} \to \mathbb{R}^{N_1 \times N_2}$ *is the operator*

$$\mathcal{F}(\boldsymbol{f}_1, \dots, \boldsymbol{f}_{n_0}) = \left(\mathcal{W}_{\boldsymbol{\theta}_4, b_4} \circ \mathrm{ReLU} \circ \mathcal{W}_{\boldsymbol{\theta}_3, b_3} \circ \mathrm{ReLU} \circ \mathcal{W}_{\boldsymbol{\theta}_2, b_2} \circ \mathrm{ReLU} \circ \mathcal{W}_{\boldsymbol{\theta}_1, b_1}\right)(\boldsymbol{f}_1, \dots, \boldsymbol{f}_{n_0})$$

for $\boldsymbol{f}_1, \dots, \boldsymbol{f}_{n_0} \in \mathbb{R}^{N_1 \times N_2}$*. Here* $\mathrm{ReLU}(x) := \max\{x, 0\}$ *is applied coordinate-wise.*

Notice that in Definition 4.1.1 we have three inner weight layers instead of two as in the original architecture (see Figure 4.1). We depict this scenario in Figure 4.2.

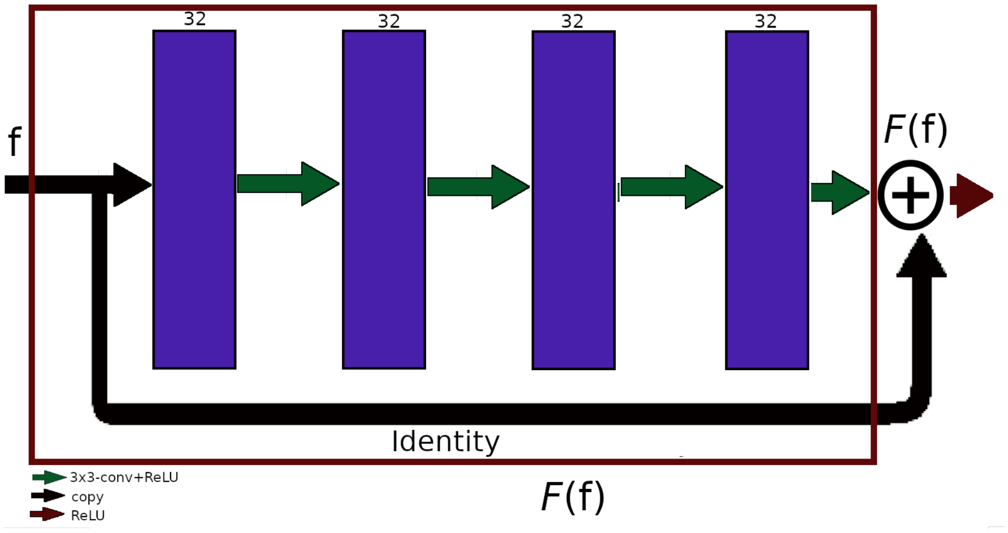

Figure 4.2: Principal ResNet block used in this thesis, with three convolutional inner layers.

The ResNet architecture has been used in many image classification applications (see [71]), but it has also be used in image reconstruction tasks. In particular, convolutional ResNets are the main ingredient of the learned primal-dual architecture, used for image reconstruction in inverse problems [3]. In Section 4.3 we explore the learned primal-dual architecture in more detail.

Since we would like to analyze the wavefront set propagation performed by the ResNet architecture, we need to introduce the notion of the ResNet architecture in the continuum setting. For that purpose, we will rewrite each layer in ResNet as an operator between functional (distributional) space. In Sections 4.1.1, 4.1.2 and 4.2.5 we introduce the continuum setting of the three main components of convolutional residual neural networks, that is, the convolutional, residual and ReLU layers. This allows us to define the operator given by the ResNet architecture. In addition, we also analyze the microlocal behavior of each component individually. In other words, we compute the microcanonical relation of such operators. In this case, we are extending the notion of microcanonical relation as a mapping describing the wavefront set propagation when an operator is applied. We will later use this analysis to characterize the singularity propagation within the learned primal-dual architecture.

4.1.1 The continuum convolutional operator

In general, in the continuum setting, the operator defined by the convolution of a function with a smooth filter will be pseudodifferential. The main problem with this approach is that the convolution with a smooth kernel will vanish all the singularities of the functions it acts upon. Since this does not happen in practice when working with discrete convolutions, this suggests that we need to find another interpretation for the continuum counterpart of the convolution in (4.1.1).

To find an appropriate representation in the continuum setting to the discrete convolution step in (4.1.1), and inspired by [101], we can interpret the discrete convolution as a discretization of a differential operator. Consider a continuum image $f \in L^2(\mathbb{R}^2)$ with discretization $\boldsymbol{f} \in \mathbb{R}^{N\times N}$, namely

$$\boldsymbol{f} = \begin{pmatrix} f_{11} & \cdots & f_{1N} \\ \vdots & \ddots & \vdots \\ f_{N1} & \cdots & f_{NN} \end{pmatrix}.$$

Moreover, let $\mathcal{K}_{\boldsymbol{\theta}}$ be a 3×3 convolutional operator parametrized by the filter

$$\boldsymbol{\theta} = \begin{pmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \\ \theta_{31} & \theta_{32} & \theta_{33} \end{pmatrix}, \tag{4.1.2}$$

where $\theta_{ij} \in \mathbb{R}$. Thus, applying such an operator is given by the discrete convolution

$$\mathcal{K}^d_{\boldsymbol{\theta}} \boldsymbol{f} := \boldsymbol{\theta} * \boldsymbol{f}, \tag{4.1.3}$$

where

$$\boldsymbol{\theta} * \boldsymbol{f}[i,j] = \sum_{l=1}^{3}\sum_{k=1}^{3} \boldsymbol{\theta}[l,k]\boldsymbol{f}[i-l,j-k], \quad \text{for } i,j \in \{1,\dots,N\}.$$

Next, note that the filter $\boldsymbol{\theta}$ can be expressed in terms of a basis of $\mathbb{R}^{3\times 3}$ as follows

$$\begin{aligned} \boldsymbol{\theta} = \boldsymbol{\theta}(\beta) = \beta_{11}\boldsymbol{\Delta}_{11} + \frac{\beta_{12}}{2h}\boldsymbol{\Delta}_{12} + \frac{\beta_{21}}{2h}\boldsymbol{\Delta}_{21} + \frac{\beta_{22}}{4h^2}\boldsymbol{\Delta}_{22} + \frac{\beta_{13}}{h^2}\boldsymbol{\Delta}_{13} \\ + \frac{\beta_{31}}{h^2}\boldsymbol{\Delta}_{31} + \frac{\beta_{32}}{2h^3}\boldsymbol{\Delta}_{32} + \frac{\beta_{23}}{2h^3}\boldsymbol{\Delta}_{23} + \frac{\beta_{33}}{h^4}\boldsymbol{\Delta}_{33}, \end{aligned} \tag{4.1.4}$$

where $\Delta_{ij} \in \mathbb{R}^{3\times 3}$ are the basis elements of $\mathbb{R}^{3\times 3}$, given by

$$\begin{aligned} &\boldsymbol{\Delta}_{11} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad \boldsymbol{\Delta}_{12} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & -1 & 0 \end{pmatrix}, \quad \boldsymbol{\Delta}_{13} = \begin{pmatrix} 0 & -1 & 0 \\ 0 & 2 & 0 \\ 0 & -1 & 0 \end{pmatrix}, \\ &\boldsymbol{\Delta}_{21} = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & -1 \\ 0 & 0 & 0 \end{pmatrix}, \quad \boldsymbol{\Delta}_{22} = \begin{pmatrix} 1 & 0 & -1 \\ 0 & 0 & 0 \\ -1 & 0 & 1 \end{pmatrix}, \quad \boldsymbol{\Delta}_{23} = \begin{pmatrix} 1 & -2 & 1 \\ 0 & 0 & 0 \\ -1 & 2 & -1 \end{pmatrix}, \\ &\boldsymbol{\Delta}_{31} = \begin{pmatrix} 0 & 0 & 0 \\ 1 & -2 & 1 \\ 0 & 0 & 0 \end{pmatrix}, \quad \boldsymbol{\Delta}_{32} = \begin{pmatrix} 1 & 0 & -1 \\ -2 & 0 & 2 \\ 1 & 0 & -1 \end{pmatrix}, \quad \boldsymbol{\Delta}_{33} = \begin{pmatrix} -1 & 2 & -1 \\ 2 & -4 & 2 \\ -1 & 2 & -1 \end{pmatrix}. \end{aligned} \tag{4.1.5}$$

Note that the 3×3 matrices Δ_{ij} represent the matrix form of finite difference discretization of the partial derivatives, when the grid spacing is $h = 1$ in (4.1.4) around a central point in $\boldsymbol{f}$. For a smooth function f, we derive that if the discretization h goes to zero, then

$$\begin{aligned}\boldsymbol{\theta} * \boldsymbol{f}(i,j) \to \big(&\beta_{11}f + \beta_{12}\partial_2 f + \beta_{21}\partial_1 f + \beta_{22}\partial_1\partial_2 f + \beta_{13}\partial_2^2 f \\ &+ \beta_{31}\partial_1^2 f + \beta_{23}\partial_2^2\partial_1 f + \beta_{32}\partial_1^2\partial_2 f + \beta_{33}\partial_1^2\partial_2^2 f\big)\,((x_i, y_j)),\end{aligned}$$

where $(x_i, y_j)_{i,j=1}^N$ are the points on the corresponding discretization grid. For an open set $\Omega \subset \mathbb{R}^2$, we obtain the operator $\mathcal{K}_{\boldsymbol{\theta}}$ in (4.1.3) defined on $\mathcal{S}(\Omega)$.

Definition 4.1.2. *Let $\boldsymbol{\theta} = \{\theta_{i,j}\}_{i,j=1}^3 \in \mathbb{R}^{3\times3}$ be a discrete convolutional kernel. In addition, let $\Omega \subset \mathbb{R}^2$ be an open domain. The* continuum convolutional operator*, $\mathcal{K}_{\boldsymbol{\theta}} : \mathcal{S}(\Omega) \to \mathcal{S}(\Omega)$, is then given by*

$$\begin{aligned}\mathcal{K}_{\boldsymbol{\theta}}(f) =&\beta_{11}f + \beta_{12}\partial_2 f + \beta_{21}\partial_1 f + \beta_{22}\partial_1\partial_2 f + \beta_{13}\partial_2^2 f \\ &+ \beta_{31}\partial_1^2 f + \beta_{23}\partial_2^2\partial_1 f + \beta_{32}\partial_1^2\partial_2 f + \beta_{33}\partial_1^2\partial_2^2 f.\end{aligned} \tag{4.1.6}$$

and $\{\beta_{ij}\}_{i,j=1}^3$ are the corresponding coefficients of the base expansion (4.1.4)*. By duality, we can extend $\mathcal{K}_{\boldsymbol{\theta}}$ to tempered distributions $\mathcal{S}'(\Omega)$.*

In addition, when applying convolutions in neural networks, one typically works with multiple channels. In our analysis above, we are simply using a filter with one channel. This analysis can be easily extended to multiple channels by simply applying the operator $\mathcal{K}_{\boldsymbol{\theta}}$ in (4.1.6) to each channel, individually.

Finally, the convolution operator $\mathcal{K}_{\boldsymbol{\theta}}$ in (4.1.3) with a 3×3 kernel $\boldsymbol{\theta}$ can be seen as a discretization of a 2nd order linear differential operator, defined in (4.1.6). The coefficients $\{\beta_{i,j}\}_{1\leq i,j\leq 3}$ of this linear differential operator are defined by the values of the filter $\boldsymbol{\theta}$ using the change of basis in (4.1.4).

Remark 4.1.3. *Since every linear differential operator is a pseudodifferential operator. From Equation* (4.1.6)*, we write the operator $\mathcal{K}_{\boldsymbol{\theta}} : \mathcal{S}(\mathbb{R}^2) \to \mathcal{S}(\mathbb{R}^2)$ in its pseudodifferential form as*

$$\mathcal{K}_{\boldsymbol{\theta}} f(x) = \frac{1}{4\pi^2}\int_{\mathbb{R}^2}\int_{\mathbb{R}^2} e^{i(x-y)\cdot\xi} p_{\boldsymbol{\theta}}(\xi) f(y)\,dy\,d\xi,$$

where the amplitude function or symbol p is given by the polynomial:

$$\begin{aligned}p_{\boldsymbol{\theta}}(\xi) =&\beta_{11} + \beta_{12}\xi_2 + \beta_{21}\xi_1 + \beta_{22}\xi_1\xi_2 \\ &+ \beta_{13}\xi_2^2 + \beta_{31}\xi_1^2 + \beta_{23}\xi_2^2\xi_1 + \beta_{32}\xi_1^2\xi_2 + \beta_{33}\xi_1^2\xi_2^2.\end{aligned} \tag{4.1.7}$$

This also holds since the symbol is of order m.

Notice that, the interpretation of a discrete convolutional operator, which takes non-smooth images as inputs, needs to be define in the continuum in a distributional sense.

This means that also all the other operations will need to be applicable to tempered distributions. In Section 4.2.1 we study the wavefront set propagation done by $\mathcal{K}_{\boldsymbol{\theta}}$. In addition, Section 4.2.1 discusses the microlocal analysis of the operator $\mathcal{K}_{\boldsymbol{\theta}}$.

As stated in (4.1.1), after applying the convolutional layer on the input of the conv-ResNet block we have to pass the output to the ReLU non-linearity. In the following section we will introduce the continuum version of the pointwise ReLU application, this is defined in form of an operator acting on the space $\mathcal{S}'(\Omega)$.

4.1.2 The ReLU operator on tempered distributions

As discussed in Section 1.2 non-linear activation functions play an important role on the design of deep neural networks. In this thesis we will focus in the rectified linear unit (ReLU). This non-linearity is used in many neural networks, in particular residual neural networks. ReLU is defined as the pointwise application of the following real-valued function.

Definition 4.1.4. *The* ReLU *activation function is the function* $\mathrm{ReLU}:\mathbb{R}\to\mathbb{R}$ *given by*

$$\mathrm{ReLU}(x) := \max\{0,x\} \quad \textit{for all } x\in\mathbb{R}^2.$$

In addition if $y\in\mathbb{R}^{N\times M}$ *is an array, where* $N,M\in\mathbb{N}$ *the entry-wise application of the* ReLU *activation function is the array* $\mathrm{ReLU}(y)\in\mathbb{R}^{N\times M}$ *given by*

$$\mathrm{ReLU}(y)[i,j] := \max\{0,y[i,j]\} \quad \textit{for all } 1\le i\le N, \textit{ and } 1\le j\le M$$

where $y[i,j]$ *is the entry* i,j *of* y.

Our aim is to extend the ReLU to an operator that acts on tempered distributions in order to study its microlocal behavior. For that, we will use the next remark on the classical ReLU.

Remark 4.1.5. *Let* $\mathrm{ReLU}:\mathbb{R}\to\mathbb{R}$ *be the classical* ReLU *function (Definition 4.1.4) and* $H:\mathbb{R}\to\mathbb{R}$ *be the Heaviside step function given by*

$$H(x) := \begin{cases} 1, & \textit{if } x>0, \\ 0, & \textit{if } x\le 0. \end{cases}$$

Hence we can write the ReLU *function as*

$$\mathrm{ReLU}(x) = H(x)x,$$

The above remark can be used to extend ReLU to $f\in C^\infty(\Omega)$, by simply defining

$$\mathsf{ReLU}(f)(x) := \mathrm{ReLU}\big(f(x)\big) = \mathrm{H}\big(f(x)\big)f(x) = \begin{cases} f(x), & \text{if } f(x)>0, \\ 0, & \text{if } f(x)\le 0. \end{cases} \tag{4.1.8}$$

We only know that $\mathsf{ReLU}\colon \mathcal{S}(\Omega)\to L^\infty(\Omega)$ and in fact $\mathsf{ReLU}(f)$ may not be smooth for $f\in\mathcal{S}(\Omega)$. Thus, ReLU does not necessarily map $\mathcal{S}(\Omega)$ to $\mathcal{S}(\Omega)$, i.e., we cannot use duality to define ReLU on distributions, as we did previously with the convolutional operator. Using the characterization in (4.1.8) to extend ReLU to distributions involves extending the Heaviside function to tempered distributions and to make sure the consequent multiplication is well-defined. For that purpose, we first need to introduce the notion of *essential*, *positive*, and *negative support* of a tempered distribution.

Definition 4.1.6 (Essential support [99])**.** *Let $\Omega \subset \mathbb{R}^n$ be a domain. The* essential support *of $f \in \mathcal{S}'(\Omega)$ is defined as the set*

$$\operatorname{ess\,supp}(f) := \mathbb{R}^n \setminus \bigcup_{U \in \mathfrak{U}} U,$$

where $\mathfrak{U} := \big\{U \subset \Omega : U \text{ is open and } f\big|_U = 0\big\}$ with $f\big|_U$ denoting the restriction of f to $U \subset \Omega$.

Note that if $f \in \mathcal{C}(\Omega)$, then $\operatorname{ess\,supp}(f) = \operatorname{supp}(f)$, i.e., the essential support coincides with the classical notion of support. We also introduce the notion of *positive* and *negative supports* of a distribution.

Definition 4.1.7 (Positive support, [99])**.** *Let $\Omega \subset \mathbb{R}^n$ be a domain. The* positive support *of $f \in \mathcal{S}'(\Omega)$ is defined as the set*

$$\operatorname{supp}_+(f) := \overline{\bigcup_{U \in \mathfrak{U}} U},$$

where $\mathfrak{U} := \big\{U \subset \Omega : f(\phi) > 0 \text{ for all } \phi \in \mathcal{S}(\Omega) \setminus \{0\}, \operatorname{supp}\phi \subset U, \phi \geq 0\big\}$. Finally, the negative support *of f is defined as $\operatorname{supp}_-(f) := \overline{(\operatorname{supp}_+(f))^c}$.*

We can now define the *Heaviside operator* acting on tempered distributions.

Definition 4.1.8. *Let $\Omega \subset$ be open. The* Heaviside operator $\mathsf{H} \colon \mathcal{S}'(\Omega) \to L^\infty(\Omega)$ *is defined as*

$$\mathsf{H}(f) := \mathbb{1}_{\operatorname{supp}_+(f)}, \quad \text{for } f \in \mathcal{S}'(\Omega), \tag{4.1.9}$$

where $\operatorname{supp}_+(f) \subset \Omega$ is the positive support of f (Definition 4.1.7) and $\mathbb{1}_{\operatorname{supp}_+(f)}$ denotes the characteristic function of $\operatorname{supp}_+(f)$.

Before proceeding with the extension of the ReLU operator to tempered distributions, we would like to list a set of desirable properties. In particular, $\mathsf{ReLU} \colon \mathcal{S}'(\Omega) \to \mathcal{S}'(\Omega)$ should preferably have the following properties.

Remark 4.1.9.

1. ***Negative support is mapped to zero:*** $\operatorname{ess\,supp} \mathsf{ReLU}(f) \subset \operatorname{supp}_+(f)$ *(see Definition 4.1.7 for a definition of the essential support* $\operatorname{ess\,supp}$*),*
2. ***Positive support does not change:*** $\mathsf{ReLU}(f)(\phi) = f(\phi)$ *for all test functions $\phi \in \mathcal{S}(\Omega)$ supported in* $\operatorname{supp}_+(h)$*,*
3. ***Relation with the Heaviside operator:*** $\mathsf{ReLU}(f) = \mathsf{H}(f)\, f$ *whenever $f \in \mathcal{S}'(\Omega)$.*

Having extended the Heaviside function to distributions as in (4.1.9), we need to make sure that the multiplication between the distribution $f \in \mathcal{S}'(\Omega)$ and $\mathsf{H}(f) \in L^\infty(\Omega)$ is well-defined. By Definition 2.2.3 and the Hörmander condition (Theorem 2.2.11), this is indeed the case if $(x, -\lambda) \notin \mathrm{WF}(f)$ for all $(x, \lambda) \in \mathrm{WF}\big(\mathsf{H}(f)\big)$, i.e., we can define $\mathsf{ReLU}(f)$ by (4.1.8) for any $f \in \mathcal{S}'(\Omega)$ that satisfies this criteria. However, the multiplication is not necessarily well-defined, if there exists $(x, \lambda) \in \mathrm{WF}\big(\mathsf{H}(f)\big)$, where $(x, -\lambda) \in \mathrm{WF}(f)$. The next remark shows a particular case when it is well defined.

Remark 4.1.10. *If $u, v \in L^2_{\mathrm{loc}}(\mathbb{R})$ and we define $uv(x) = u(x)v(x)$ almost everywhere, then the multiplication of u and v defined in* (2.2.2) *coincides with uv almost everywhere. This holds even if there exist $(x, \lambda) \in \mathrm{WF}(u)$ such that $(x, -\lambda) \in \mathrm{WF}(v)$.*

To see this, let $x \in \mathbb{R}^2$ and ψ be as in Definition 2.2.3. Then

$$\widehat{\psi^2(uv)}(\xi) = \int_{\mathbb{R}^2} \widehat{\psi u}(\xi)\widehat{\psi v}(\nu - \xi)d\xi$$

holds in an L^2 sense. Moreover, by Plancherel's identity, we have that $\widehat{\psi u}, \widehat{\psi v} \in L^2(\mathbb{R}^2)$, which yields with the Cauchy-Schwarz identity, that

$$\int_{\mathbb{R}^2} \big|\widehat{\psi u}(\xi)\widehat{\psi v}(\nu - \xi)\big|d\xi \leq \|\widehat{\psi u}\|_2 \, \big\|\widehat{\psi v}(\nu - \cdot)\big\|_2 = \|\widehat{\psi u}\|_2 \, \|\widehat{\psi v}\|_2 < \infty.$$

This yields absolute convergence in (2.2.2)*.*

Thus, all we know is that the multiplication of $\mathsf{H}(f)$ and f is well-defined whenever $f \in \mathcal{S}'_{\mathrm{loc}}(\Omega)$ (Remark 4.1.10). Since we aim to compute WF(ReLU(f)), one idea is therefore to locally smooth-out f close to points, where we cannot define the multiplication of f with $\mathsf{H}(f)$. The next notion will play an important role in the definition of ReLU on distributions.

Definition 4.1.11 ([9, Definition A.6])**.** *Let $f \in \mathcal{S}'(\Omega)$ be a tempered distribution. Then, the L^2-support of f is defined as the largest open set on Ω where f is given by an L^2-function:*

$$\mathrm{supp}_{L^2}(h) := \bigcup \Big\{ U \subset \Omega \ \textit{open} \ : f\big|_U \in \mathcal{S}'(U) \Big\}.$$

This means that if $x \in \mathrm{supp}_{L^2}(h)$ then there is an open set $x \in U \subset \Omega$ and $f_U \in L^2(U)$ such that $f(\phi) = \int_U f_U(x)\phi(x)dx$ for all $\phi \in \mathcal{S}(U)$.

This leads to the following definition of ReLU on distributions.

Definition 4.1.12 ([9, Definition 4.1])**.** *Let $\Omega \subset \mathbb{R}^2$ be open, $\kappa > 0$, and $\phi_\kappa \in \mathcal{S}(\mathbb{R}^2)$ be a function that integrates to 1, is positive and is supported on a compact subset of $B_\kappa(0)$. Then define*

$$\mathsf{ReLU}_{\kappa,\phi_\kappa}(f) := \mathsf{H}(f)f^s, \quad \textit{for } f \in \mathcal{S}'(\Omega), \tag{4.1.10}$$

where $f^s := (1 - \theta_\kappa)f$. Here $\theta_\kappa := \mathbb{1}_X * \phi_\kappa$ with

$$X := \Big\{ x \in \mathbb{R}^2 \setminus \mathrm{supp}_{L^2}(f) \colon (x, \lambda) \in \mathrm{WF}(H(h)), (x, -\lambda) \in \mathrm{WF}(h) \ \textit{for a} \ \lambda \in \mathbb{S}^1 \Big\} + B_\kappa(0).$$

In the above, $\mathrm{supp}_{L^2}(f) \subset \Omega$ denotes the L^2-support of f (Definition 4.1.11).

We next show that Definition 4.1.12 could be used to extend the ReLU function to distributions.

Proposition 4.1.13 ([9, Proposition 4.2])**.** *Let $\Omega \subset \mathbb{R}^2$ be open. Consider $\mathsf{ReLU}_{\kappa,\phi_\kappa}$ defined in* (4.1.10) *for some $\kappa > 0$ and $\phi_\kappa \in \mathcal{S}(\mathbb{R}^2)$ that integrates to* 1*, is positive, and is supported on a compact subset of $B_\kappa(0)$. Then $\mathsf{ReLU}_{\kappa,\phi_\kappa} \colon \mathcal{S}'(\Omega) \to \mathcal{S}'(\Omega)$.*

Proof. We need to show that $\mathsf{ReLU}_{\kappa,\phi_\kappa}(f) \in \mathcal{S}'(\Omega)$, whenever $f \in \mathcal{S}'(\Omega)$. To see this, note first that $1 - \theta_\kappa$ is smooth and vanishes on a neighborhood of every $x \in \Omega \setminus \operatorname{supp}_{L^2}(f)$, where

$$(x, \lambda) \in \mathrm{WF}\big(\mathsf{H}(f)\big) \quad \text{and} \quad (x, -\lambda) \in \mathrm{WF}(f) \quad \text{for some } \lambda \in \mathbb{S}^1.$$

Hence, the product $(1 - \theta_\kappa)f$ is well-defined and by Theorem 2.2.11, there does not exist an $x \in \Omega \setminus \operatorname{supp}_{L^2}(f)$ such that

$$(x, \lambda) \in \mathrm{WF}\big(\mathsf{H}(f)\big) \quad \text{and} \quad (x, -\lambda) \in \mathrm{WF}((1 - \theta_\kappa)f).$$

Theorem 2.2.11 and Remark 4.1.10 now imply that $\mathsf{ReLU}_{\kappa,\phi_\kappa}(f) \in \mathcal{S}'(\Omega)$ whenever $f \in \mathcal{S}'(\Omega)$, which concludes the proof. □

Remark 4.1.14. *The set X in Definition 4.1.12 is a neighborhood of the set on which the definition of* $\mathrm{ReLU}(h)$ *via the multiplication $H(h)h$ is not well defined. To understand the nature of this set, we consider three examples:*

1. *$f \in \mathcal{S}'(\Omega)$. Then $X = \varnothing$, and hence $\mathsf{ReLU}_{\kappa,\phi_\kappa}(f) = \mathrm{ReLU}(f)$. In particular, if $f = \mathrm{H}(h)$ for some $h \in \mathcal{S}'(\Omega)$, then $\mathsf{ReLU}_{\kappa,\phi_\kappa}(f) = \mathrm{ReLU}(f) = f$.*
2. *$f = P(\mathbb{1}_B)$ for some domain $B \subset \Omega$ and P is an elliptic linear pseudodifferential operator of order at least one. Then $\operatorname{ess\,supp}(f) \subset \partial B$, so $\mathsf{H}(f) = 0$ which in turn implies that $X = \varnothing$ and $\mathrm{ReLU}_{\kappa,\phi_\kappa}(f) = 0$.*
3. *$f = P(\mathbb{1}_B + h)$ for some domain $B \subset \Omega$ and P is an elliptic linear differential operator. Assume furthermore that $h \in \mathcal{C}^\infty(\Omega)$ is such that $P(h)$ is positive on B. Then $X = \partial B + B_\kappa(0)$, since f is not a function at ∂B and $\mathsf{H}(f) = \mathbb{1}_B$. Thus $\mathrm{WF}\big(\mathrm{H}(f)\big) = \mathrm{WF}(f)$.*

Finally, notice that $\mathsf{ReLU}_{\kappa,\phi_\kappa} \colon \mathcal{S}'(\Omega) \to \mathcal{S}'(\Omega)$ in Definition 4.1.12 fulfills the first and third properties from Remark 4.1.9. Furthermore, in relation with the second property, $\mathsf{ReLU}_{\kappa,\phi_\kappa}(f)(\phi) = f(\phi)$ holds for all $\phi \in \mathcal{S}(\Omega)$ with a support that has a distance of more than 2κ from $\mathrm{WF}\big(\mathsf{H}(f)\big) \subset \partial \operatorname{supp}_+(f)$.

In Section 4.2.2 we will explore the computation of $\mathrm{WF}(H(f))$ and $\mathrm{WF}(\mathsf{ReLU}_{\kappa,\phi_\kappa}(f))$. In contrast to the convolutional operator, we are note able to precisely compute the wavefront set of $\mathrm{WF}(\mathsf{ReLU}_{\kappa,\phi_\kappa}(f))$. This means that, for the rest of points where such wavefront set cannot be computed we will need to make use of the DeNSE wavefront set extraction [8].

4.1.3 The continuum convolutional residual operator

Finally, following Definitions 4.1.2 and 4.1.12 we can define the *continuum convolutional ResNet operator* as follows.

Definition 4.1.15 (Continuum two-dimensional convolutional ResNet, [9, Definition 4.4])**.** *Let $\Omega \subset \mathbb{R}^2$ be open and let $N \in \mathbb{N}$, $j \in \{1,2,3,4\}$, where $n_4 = 1$ be the* numbers of channels per layer*. Further, for $j = 1,\dots,4$, let $\boldsymbol{\theta}_j := (\boldsymbol{\theta}_j^{l,k})_{l=1,k=1}^{n_{j-1},n_j} \subset (\mathbb{R}^{3\times 3})^{n_{j-1}\times n_{j-1}}$ be a* set of coefficients*. Let $\kappa > 0$ and let $\phi_\kappa \in \mathcal{S}(\Omega)$ be a function that integrates to 1, is positive and is supported on a compact subset of $B_\kappa(0)$.*

We define the continuum convolutional affine operator $W^c_{\boldsymbol{\theta}_j} : (\mathcal{S}'(\Omega))^{n_{j-1}} \to (\mathcal{S}'(\Omega))^{n_j}$ *as*

$$W_{\boldsymbol{\theta}_j}(f)_k = \sum_{l=1}^{n_{j-1}} \mathcal{K}_{\boldsymbol{\theta}_j^{l,k}}(f) \quad \text{for } k \in \{1,\dots,n_j\} \text{ and } f \in (\mathcal{S}'(\Omega))^{n_{j-1}}. \tag{4.1.11}$$

The continuum ResNet operator $\mathrm{ResNet}_{\kappa,\phi_\kappa} \colon (\mathcal{S}'(\Omega))^{n_0} \to \mathcal{S}'(\Omega)$ *is then given by*

$$\mathrm{ResNet}_{\kappa,\phi_\kappa}(f_1,\dots,f_{n_0}) := f_1 + \mathcal{G}(f_1,\dots,f_{n_0}) \quad \text{for } f_1,\dots,f_{n_0} \in \mathcal{S}'(\Omega),$$

where $\mathcal{G} : (\mathcal{S}'(\Omega))^{n_0} \to \mathcal{S}'(\Omega)$ is the operator

$$\mathcal{G}(f_1,\dots,f_{n_0}) = \big(W_{\boldsymbol{\theta}_4} \circ \mathrm{ReLU}_{\kappa,\phi_\kappa} \circ \mathcal{W}_{\boldsymbol{\theta}_3} \circ \mathrm{ReLU}_{\kappa,\phi_\kappa} \circ \mathcal{W}_{\boldsymbol{\theta}_2} \circ \mathrm{ReLU}_{\kappa,\phi_\kappa} \circ \mathcal{W}_{\boldsymbol{\theta}_1}\big)(f_1,\dots,f_{n_0})$$

for $f_1,\dots,f_{n_0} \in \mathcal{S}'(\Omega)$.

Remark 4.1.16.

1. *In contrast with Definition 4.1.1, we do not include a bias term in the definition of the continuum ResNet above, since no such term will appear in our implementation. We also show that the absence of the bias term does not have negative impact in the performance of the algorithm.*

2. *Note that besides the previously defined operators $\mathcal{K}_{\boldsymbol{\theta}_j^{l,k}}$ and $\mathrm{ReLU}_{\kappa,\phi_\kappa}$, only addition is applied in the continuum ResNet. Since the set of distributions is a linear space, we conclude that $\mathrm{ResNet}_{\kappa,\phi_\kappa}$ is a well-defined operator from $\mathcal{S}'(\Omega)$ to $\mathcal{S}'(\Omega)$.*

In the next section we explore the microlocal analysis of the $\mathrm{ResNet}_{\kappa,\phi_\kappa}$ operator, for that, we first analyze the microlocal behavior of each of its components, i.e., $\mathcal{K}_{\boldsymbol{\theta}}$, $\mathsf{ReLU}_{\kappa,\phi_\kappa}$ and the residual layer.

4.2 Microlocal analysis of continuum convolutional residual neural networks

Now that we have defined the residual network architecture in the continuum setting, in form of the operator $\mathrm{ResNet}_{\kappa,\phi_\kappa}$, we aim to study the propagation of singularities that such operator performs on its input. For that purpose, in this section we study the propagation of singularities performed by each of its components, separately.

4.2.1 Differential operator

The pseudodifferential nature of the convolutional operator, $\mathcal{K}_{\boldsymbol{\theta}}$, acting on $\mathcal{S}'(\Omega)$ (Definition 4.1.2) allows us study its microlocal behavior using standard microlocal analysis techniques [70]. In particular, the pseudo-local property (Theorem 2.3.10) can be used to prove the next proposition.

Proposition 4.2.1. *Let $\boldsymbol{\theta} \in \mathbb{R}^{3\times 3}$ be a discrete convolutional kernel and $\Omega \subset \mathbb{R}^2$ and $\Xi \subset \mathbb{R} \times (0,\pi)$ be open sets. In addition, let $\mathcal{K}_{\boldsymbol{\theta}} : \mathcal{S}'(\Omega) \to \mathcal{S}'(\Omega)$ be the convolutional kernel from Definition 4.1.2. Then the microcanonical relation of $\mathcal{K}_{\boldsymbol{\theta}}$ is given by*

$$\mathrm{WF}(\mathcal{K}_{\boldsymbol{\theta}} f) \subset \mathrm{WF}(f). \tag{4.2.1}$$

This means that $\mathcal{K}_{\boldsymbol{\theta}}$ will not introduce new singularities. In addition, let $p_{\boldsymbol{\theta}}$ be the amplitude of $\mathcal{K}_{\boldsymbol{\theta}}$ defined in (4.1.7). *If $p_{\boldsymbol{\theta}}$ satisfies*

$$0 < |p_{\boldsymbol{\theta}}(\xi)| \textit{ for all } ||\xi|| \neq 0,$$

then $\mathcal{K}_{\boldsymbol{\theta}}$ preserves the wavefront set, i.e., we have

$$\mathrm{WF}(\mathcal{K}_{\boldsymbol{\theta}} f) = \mathrm{WF}(f).$$

Proof. Following Remark 4.1.3 we know that $\mathcal{K}_{\boldsymbol{\theta}}$ is a pseudodifferential operator with amplitude $p_{\boldsymbol{\theta}}$ given by (4.1.7). By Theorem 2.3.10, we have that

$$\mathrm{WF}(\mathcal{K}_{\boldsymbol{\theta}} f) \subset \mathrm{WF}(f).$$

In addition, if

$$0 < |p_{\boldsymbol{\theta}}(\xi)| \text{ for all } ||\xi|| \neq 0$$

then $\mathcal{K}_{\boldsymbol{\theta}}$ is an elliptic pseudodifferential operator, which means it preserves the wavefront set. □

Proposition 4.2.1 allows us to propagate the wavefront set of a tempered distribution $f \in \mathcal{S}'(\Omega)$ through the convolutional layer of the continuum ResNet architecture. In the following section we will explore the other related operators.

4.2.2 ReLU application

In this section we will explore the microlocal behavior of $\mathsf{ReLU}_{\kappa,\phi_\kappa} : \mathcal{S}'(\Omega) \to \mathcal{S}'(\Omega)$ defined in (4.1.10). The strategy for this analysis is based on the Hörmander condition, also known as the product theorem (Theorem 2.2.11).

4.2.3 The wavefront set of $\mathsf{H}(f)$

For a function $g \in \mathcal{S}(\Omega)$, the wavefront set of $\mathsf{H}(g)$ is determined through the following factors: A point x, with a neighborhood where g is almost always positive will be mapped to a constant by the Heaviside function. Since constant functions are smooth, this operation deletes the wavefront set associated to a neighborhood of x. The same argument applies on neighborhoods where g is negative almost everywhere. Points x' in which g vanishes have the potential to create new singularities since the Heaviside function has a jump in 0. If g is smooth in x' and also has non-vanishing gradient, then the implicit function theorem tells us the form of the discontinuity of $\mathsf{H}(g)$. Following these statements, we have the next proposition.

Proposition 4.2.2 ([9, Proposition 4.8])**.** *Let $g \in L^2(\mathbb{R}^2)$. Let further*

$$\begin{aligned} R_g :=&\{x \in \mathbb{R}^2 \colon x \in \partial(\mathrm{supp}_+(g)), x \notin \mathrm{sing\,supp}(g), \nabla g(x) \neq 0\},\\ C_g :=&\{x \in \mathbb{R}^2 \colon x \in \partial(\mathrm{supp}_+(g)), x \notin \mathrm{sing\,supp}(g), \nabla g(x) = 0\},\\ S_g :=&\{x \in \mathbb{R}^2 \colon x \in \partial(\mathrm{supp}_+(g)), x \in \mathrm{sing\,supp}(g)\}. \end{aligned}$$

If for an $\alpha \neq 0$

$$x \in R_g \text{ and } \lambda = \alpha \nabla_x(g) \tag{4.2.2}$$

then, $(x, \lambda) \in \mathrm{WF}(\mathsf{H}(g))$. Moreover, $(x, \lambda) \in \mathrm{WF}(\mathsf{H}(g))$ only if (4.2.2) holds or $x \in C_g \cup S_g$.

Proof. We start with the "only if" part. The statement is clear if $\mathrm{WF}(\mathsf{H}(g)) = \varnothing$. Otherwise, we choose $(x, \lambda) \in \mathrm{WF}(\mathsf{H}(g))$. Assume first that $x \in \partial(\mathrm{supp}_+(g))^c$. Then either $x \in \mathrm{supp}_{-,0}(g)$ or $x \in \mathrm{supp}_+(f)^\circ$. Since both $\mathrm{supp}_{-,0}(g)$ and $\mathrm{supp}_+(g)^\circ$ are open sets, we have that there exists an open neighborhood U of x such that $U \subset \mathrm{supp}_{-,0}(g)$ or $U \subset \mathrm{supp}_+(g)^\circ$. As a result, $\mathsf{H}(g)$ is constant on U. Therefore, (x, λ) cannot be in $\mathrm{WF}(\mathsf{H}(g))$, which yields a contradiction.

Hence, we can assume that $(x, \lambda) \in \mathrm{WF}(\mathsf{H}(g))$ and $x \in \partial(\mathrm{supp}_+(g))$. In addition, we assume that $x \notin C_g \cup S_g$. Then, $x \notin \mathrm{sing\,supp}(g)$. Therefore, there exists a neighborhood U' of x, where g is smooth and ∇g does not vanish.

We wish to show now that on U' the set $\{g = 0\}$ is a smooth curve with normal $\nabla_x g$ at x. For this, we invoke a *smooth version of the implicit function theorem* [77, Theorem 2.1]. In this form, the theorem considers a smooth function $\underline{g} \colon \Omega \to \mathbb{R}$ such that

$$0 = \underline{g}(x_1^*, x_2^*),$$

for $(x_1^*, x_2^*) \in \Omega$. If $\frac{\partial g}{\partial x_2} \neq 0$ then there exists a smooth κ defined on a neighborhood of x_1^* such that locally, i.e., for x_1 in an open neighborhood of x_1^*,

$$\underline{g}(x_1, \kappa(x_1)) = 0$$

also $\kappa'(x_1) = \frac{\partial g}{\partial x_1}(x_1) / \frac{\partial g}{\partial x_2}(x_1)$. Moreover, in an open neighborhood of x_1^*, x_2^* every (x_1, x_2) such that $\underline{g}(x_1, x_2) = 0$ is of the form $(x_1, \kappa(x_1))$. Applying the implicit function theorem to g if $\frac{\partial g}{\partial x_2} \neq 0$ yields that $\eta_x = \nabla g(x)/\|\nabla g\|$ is a normal at the zero-level set of g at x. By swapping variables, the same argument can be made if $\frac{\partial g}{\partial x_1} \neq 0$. We obtain that locally $\mathsf{H}(g) = \mathbb{1}_\Omega$ with $\partial\Omega$ being a smooth curve that has normal η_x at x. By arguments in [72], this implies that $(x, \lambda) \in \mathrm{WF}(\mathbb{1}_\Omega)$ if $\lambda = \alpha\eta_x$ for an $\alpha \neq 0$. This concludes the proof of the "only if" part.

For the "if" part, we notice again that if $x \in R_g$, then $x \in \partial(\mathrm{supp}_+(g))$. Thus, $x \notin S_g$ which implies that $x \notin \mathrm{sing\,supp}(g)$. Therefore, and since $x \notin C_g$, the implicit function theorem is applicable. The same argument as before yields that (4.2.2) holds. □

Remark 4.2.3. *It is important to ask whether Proposition 4.2.2 is tight. To improve our intuition, we show an example for each of the cases of Proposition 4.2.2 that may lead to the* creation of wavefront set. *Here we mean by* creation of wavefront set *the introduction of new oriented singularities to the function where the operator acts upon.*

1. Creation of wavefront set according to (4.2.2).

   Let $g(x) = 1 - \|x\|^2$. The squared Euclidean norm is a smooth function. It holds that
   $$\{g = 0\} = \{x \colon \|x\| = 1\}$$
   is the unit circle. Moreover, $\mathsf{H}(g) = \mathbb{1}_{B_1}$ is the indicator of the unit ball. It is not hard to see that the wavefront set of this function is $\{(x, x) \colon x \in \mathbb{S}^1\}$. Also
   $$\nabla_x(1 - \|x\|^2) = \begin{pmatrix} 2x_1 \\ 2x_2 \end{pmatrix} = 2x.$$

2. $x \in C_g$ and $x \in \mathrm{sing\,supp}(\mathsf{H}(g))$.

   Let g_1 be a positive C^∞ function, supported on a set Ω_1 that contains $(0,0)$. Let g_2 be another such function, however, with $\Omega_1 \cap \Omega_2 = \{(0,0)\}$. If $\Omega_1 \cup \Omega_2$ is not an open neighborhood of $(0,0)$, which is possible, then $\mathsf{H}(g_1 + g_2)$ is discontinuum at $(0,0)$ implying that $(0,0)$ is a singular point of $\mathsf{H}(g_1 + g_2)$. One concrete example would be given by $\Omega_1 = [-1,0]^2$, $\Omega_2 = [0,1]^2$. In this case, $\partial(\mathrm{supp}_+(g_1 + g_2))$ is not given by a single curve in the neighborhood of $(0,0)$. Note that, necessarily by the smoothness of g_1, g_2 it holds that $\nabla_x(g_1 + g_2) = 0$ for $x = (0,0)$.

3. $x \in C_g$ and $x \notin \mathrm{sing\,supp}(\mathsf{H}(g))$. *Let g be a smooth compactly supported positive function. Then every $x \in \partial\,\mathrm{supp}(g)$ satisfies that $x \in C_g$ and $x \notin \mathrm{sing\,supp}(\mathsf{H}(g)) = \mathrm{sing\,supp}(g)$.*

4. $x \in S_g$ and $x \in \mathrm{sing\,supp}(\mathsf{H}(g))$.

   Let $g(x) = 1/|x|^\alpha$, with α such that $g \in L^2$. Then we have that $h = g\mathbb{1}_{\mathbb{R}^+ \times \mathbb{R}}$ is square-integrable and also $(0,0) \times (\mathbb{S}^1 \cap \mathbb{R}^+ \times \mathbb{R}) \subset \mathrm{WF}(h) = \mathrm{WF}(\mathsf{H}(h))$, where the last equality holds since g and hence h are non-negative.

5. $x \in S_g$ and $x \notin \mathrm{sing\,supp}(\mathsf{H}(g))$.

Let $\phi\colon \mathbb{R} \to \mathbb{R}$ be a C^∞ function with compact support on $\mathbb{R}^+$. The function $g(x) = \phi(x_1) + \mathbb{1}_{\mathbb{R}^-}(x_1)x_1^3$ is not smooth since it has a jump in its third derivative at the $x_1 = 0$ axis. At the same time $\{(x_1, x_2) \in \mathbb{R}^2 : x_1 = 0\} = \partial(\mathrm{supp}_+(g))$. Finally, we observe that $\mathsf{H}(g)(x) = \phi(x)$ and hence the wavefront set of $\mathsf{H}(g)$ is empty.

Notice that Proposition 4.2.2 is not yet a precise characterization of the wavefront set of $\mathsf{H}(f)$. It implies that all singularities must be in one of the sets R_g, C_g or S_g but there is a closed-form of the orientations of the singularities only if $x \in R_g$. In the next section we explore the microlocal behavior of the ReLU operator expressed as the product $H(\cdot)\cdot$.

4.2.4 Wavefront set of $\mathsf{ReLU}(f) = \mathsf{H}(f)f$

In this subsection, we choose a fixed $\kappa > 0$ and $\phi_\kappa \in \mathcal{S}(\mathbb{R}^2)$ that integrates to 1, is positive, and is supported on a compact subset of $B_\kappa(0)$. To reduce the computation of $\mathrm{WF}(\mathsf{ReLU}_{\kappa,\phi_\kappa}(f))$ to that of $\mathrm{WF}(\mathsf{H}(f))$, we will make use of the following version of the product theorem.

Theorem 4.2.4 ([16, Theorem 13]). *Let u and v be distributions in $\mathcal{S}'(U)$ for an open domain U. Assume that for no point (x, λ) in $\mathrm{WF}(u)$ we have $(x, -\lambda) \in \mathrm{WF}(v)$. Then, $uv \in \mathcal{S}'(U)$ and*

$$\mathrm{WF}(uv) = S_+ \cup S_u \cup S_v,$$

where

$$\begin{aligned} S_+ &:= \{(x, \lambda + \mu)\colon (x, \lambda) \in \mathrm{WF}(u), (x, \mu) \in \mathrm{WF}(v)\}, \\ S_u &:= \{(x, \lambda)\colon (x, \lambda) \in \mathrm{WF}(u), x \in \mathrm{ess\,supp}(v)\}, \\ S_v &:= \{(x, \lambda)\colon (x, \lambda) \in \mathrm{WF}(v), x \in \mathrm{ess\,supp}(u)\}. \end{aligned}$$

In particular, for $g \in \mathcal{S}'(\mathbb{R}^2)$ and $f \in \mathcal{C}^\infty(\mathbb{R}^2)$ where $\mathrm{supp}(f)$ is compact, we have that $\mathrm{WF}(fg) \subset \mathrm{WF}(g) \cap (\mathrm{supp}(f) \times \mathbb{R}^2)$.

Let $\Omega \subset \mathbb{R}^2$, $f \in \mathcal{S}'(\Omega)$, and let us assume that $\mathrm{WF}(f)$ is known. In addition, using the results of Subsection 4.2.3 we have also access to $\mathrm{WF}(\mathsf{H}(f))$. We denote as in Definition 4.1.12

$$X_h := \{x \in \mathbb{R}^2 \setminus \mathrm{supp}_{L^2}(h)\colon (x, \lambda) \in \mathrm{WF}(H(h)), (x, -\lambda) \in \mathrm{WF}(h) \text{ for } \lambda \in \mathbb{S}^1\} + B_\kappa(0) \tag{4.2.3}$$

and

$$X_h^{3\kappa} := \{x \in \mathbb{R}^2 \setminus \mathrm{supp}_{L^2}(h)\colon (x, \lambda) \in \mathrm{WF}(H(h)), (x, -\lambda) \in \mathrm{WF}(h) \text{ for } \lambda \in \mathbb{S}^1\} + B_{3\kappa}(0). \tag{4.2.4}$$

Note that by Definition 4.1.12 we have that $\theta_\kappa = 0$ on $(X_h^{3\kappa})^c$ and hence $h^s = h$ on $(X_h^{3\kappa})^c$.

Now we have collected all necessary ingredients to be able to compute the wavefront set of $\mathsf{ReLU}_{\kappa,\phi_\kappa}(f) = \mathsf{H}(f)f$, which is expressed in the form of the following theorem.

Theorem 4.2.5 ([9, Theorem 4.11])**.** *Let $\Omega \subset \mathbb{R}^2$ be open and let $f \in \mathcal{S}'(\Omega)$. In addition, let*

$$\begin{aligned} \mathcal{A}_f :=& \, \mathrm{WF}(f) \cap (\mathrm{supp}_+(f)^{\mathrm{o}} \times \mathbb{S}^1), & (4.2.5)\\ \mathcal{R}_f :=& \{(x,\lambda) \in R_f \times \mathbb{S}^1 : (x,\lambda) \text{ follows } (4.2.2)\}, & (4.2.6) \end{aligned}$$

where R_f is defined as in Proposition 4.2.2. Moreover, $\mathcal{CS}_f$ is given by

$$\mathcal{CS}_f := \{(x,\xi) \in (S_f \cup C_f) \times \mathbb{S}^1 \colon (x,\xi) \in \mathrm{WF}(\mathsf{ReLU}_{\kappa,\phi_\kappa}(f))\}, \tag{4.2.7}$$

where C_f and S_f are defined as in Proposition 4.2.2. Then $\mathrm{WF}(\mathsf{ReLU}_{\kappa,\phi_\kappa}(f))$ *is given by*

$$\begin{aligned} \mathrm{WF}(\mathsf{ReLU}_{\kappa,\phi_\kappa}(f)) \cap (X_h^{3\kappa} \times \mathbb{S}^1)^c &= (\mathcal{A}_f \cup \mathcal{R}_f \cup \mathcal{CS}_f) \cap (X_h^{3\kappa} \times \mathbb{S}^1)^c, & (4.2.8)\\ \mathrm{WF}(\mathsf{ReLU}_{\kappa,\phi_\kappa}(f)) \cap (X_h^{3\kappa} \times \mathbb{S}^1) &\subset (\mathcal{A}_f \cup \mathcal{R}_f \cup \mathcal{CS}_f) \cap (X_h^{3\kappa} \times \mathbb{S}^1), & (4.2.9) \end{aligned}$$

In particular, we have

$$\mathrm{WF}(\mathsf{ReLU}_{\kappa,\phi_\kappa}(f)) \subset \mathcal{A}_f \cup \mathcal{R}_f \cup \left((C_f \cup S_f) \times \mathbb{S}^1\right). \tag{4.2.10}$$

Proof. Since $\mathbb{R}^2$ can be decomposed as

$$\mathbb{R}^2 = \mathrm{supp}_+(f)^{\mathrm{o}} \cup \partial(\mathrm{supp}_+(f)) \cup \mathrm{supp}_{-,0}(f),$$

we have that $\mathrm{WF}(\mathsf{ReLU}_{\kappa,\phi_\kappa}(f))$ can be decomposed as

$$\mathrm{WF}(\mathsf{ReLU}_{\kappa,\phi_\kappa}(f)) = \mathrm{A}_{f,\kappa} \cup \mathrm{B}_{f,\kappa} \cup \mathrm{D}_{f,\kappa}, \tag{4.2.11}$$

where

$$\begin{aligned} \mathrm{A}_{f,\kappa} :=& \, \mathrm{WF}(\mathsf{ReLU}_{\kappa,\phi_\kappa}(f)) \cap (\mathrm{supp}_+(f)^{\mathrm{o}} \times \mathbb{S}^1),\\ \mathrm{B}_{f,\kappa} :=& \, \mathrm{WF}(\mathsf{ReLU}_{\kappa,\phi_\kappa}(f)) \cap (\mathrm{supp}_{-,0}(f) \times \mathbb{S}^1),\\ \mathrm{D}_{f,\kappa} :=& \, \mathrm{WF}(\mathsf{ReLU}_{\kappa,\phi_\kappa}(f)) \cap (\partial(\mathrm{supp}_+(f)) \times \mathbb{S}^1). \end{aligned}$$

Notice in addition that $\mathrm{A}_{f,\kappa}$, $\mathrm{B}_{f,\kappa}$ and $\mathrm{D}_{f,\kappa}$ are disjoint. Now, since $\mathrm{supp}_+(f)^{\mathrm{o}}$ is open, we find for every $x \in \mathrm{supp}_+(f)^{\mathrm{o}}$ with $x \notin X_h^{3\kappa}$ an open neighborhood U of x such that $\mathsf{ReLU}_{\kappa,\phi_\kappa}(f)|_U = \mathsf{H}(f)|_U f^s|_U = f|_U$, since $\mathsf{H}(f)(x) = 1$ for every $x \in U$ and $h^s = h$ on $(X_h^{3\kappa})^c$. Thus

$$\mathrm{A}_{f,\kappa} \cap (X_h^{3\kappa} \times \mathbb{S}^1)^c = \mathrm{WF}(f) \cap (\mathrm{supp}_+(f)^{\mathrm{o}} \times \mathbb{S}^1) \cap (X_h^{3\kappa} \times \mathbb{S}^1)^c = \mathcal{A}_f \cap (X_h^{3\kappa} \times \mathbb{S}^1)^c, \tag{4.2.12}$$

where $\mathcal{A}_f$ is as in the statement of the proposition. Moreover, for every $x \in \mathrm{supp}_+(f)^{\mathrm{o}}$ with $x \in X_h^{3\kappa}$ there is an open neighborhood U' of x such that $\mathsf{ReLU}_{\kappa,\phi_\kappa}(f)|_{U'} = \mathsf{H}(f)|_{U'} f^s|_{U'} = f^s|_{U'}$. Therefore,

$$\begin{aligned} \mathrm{A}_{f,\kappa} \cap (X_h^{3\kappa} \times \mathbb{S}^1) &= \mathrm{WF}(f^s) \cap (\mathrm{supp}_+(f)^{\mathrm{o}} \times \mathbb{S}^1) \cap (X_h^{3\kappa} \times \mathbb{S}^1)\\ &\subset \mathrm{WF}(f) \cap (X_h^{3\kappa} \times \mathbb{S}^1) = \mathcal{A}_f \cap (X_h^{3\kappa} \times \mathbb{S}^1). \end{aligned} \tag{4.2.13}$$

Next, since $\operatorname{ess\,supp}(\mathsf{H}(f)) = \operatorname{supp}_+(f)$, by Theorem 4.2.4, we can conclude that

$$\operatorname{supp}_{-,0}(f) \subset (\operatorname{sing\,supp}(\mathsf{H}(f)f^s))^c = (\operatorname{sing\,supp}(\mathsf{ReLU}_{\kappa,\phi_\kappa}(f)))^c.$$

Then, we have

$$\mathrm{B}_{f,\kappa} = \mathrm{WF}(\mathsf{ReLU}_{\kappa,\phi_\kappa}(f)) \cap (\operatorname{supp}_{-,0}(f) \times \mathbb{S}^1) = \varnothing. \tag{4.2.14}$$

Let us now study the set $\mathrm{D}_{f,\kappa} := \mathrm{WF}(\mathsf{ReLU}_{\kappa,\phi_\kappa}(f)) \cap (\partial(\operatorname{supp}_+(f)) \times \mathbb{S}^1)$. Following the notation of Proposition 4.2.2, we can decompose the set $\partial(\operatorname{supp}_+(f))$ as

$$\partial(\operatorname{supp}_+(f)) = R_f \cup C_f \cup S_f. \tag{4.2.15}$$

Using this decomposition, we can write $\mathrm{D}_{f,\kappa}$ as

$$\mathrm{D}_{f,\kappa} = \mathrm{R}_{f,\kappa} \cup \mathrm{CS}_{f,\kappa},$$

where

$$\begin{aligned}\mathrm{R}_{f,\kappa} &:= \mathrm{WF}(\mathsf{ReLU}_{\kappa,\phi_\kappa}(f)) \cap (R_f \times \mathbb{S}^1),\\ \mathrm{CS}_{f,\kappa} &:= \mathrm{WF}(\mathsf{ReLU}_{\kappa,\phi_\kappa}(f)) \cap ((C_f \cup S_f) \times \mathbb{S}^1).\end{aligned}$$

Next, we would like to show that

$$\begin{aligned}\mathrm{R}_{f,\kappa} \cap (X_h^{3\kappa} \times \mathbb{S}^1)^c &= \{(x,\lambda) \in R_f \times \mathbb{S}^1 \colon (x,\lambda) \text{ follows (4.2.2)}\} \cap (X_h^{3\kappa} \times \mathbb{S}^1)^c\\ &= \mathcal{R}_f \cap (X_h^{3\kappa} \times \mathbb{S}^1)^c,\end{aligned} \tag{4.2.16}$$

$$\begin{aligned}\mathrm{R}_{f,\kappa} \cap (X_h^{3\kappa} \times \mathbb{S}^1) &\subset \{(x,\lambda) \in R_f \times \mathbb{S}^1 \colon (x,\lambda) \text{ follows (4.2.2)}\} \cap (X_h^{3\kappa} \times \mathbb{S}^1)\\ &= \mathcal{R}_f \cap (X_h^{3\kappa} \times \mathbb{S}^1).\end{aligned} \tag{4.2.17}$$

Let us start with (4.2.16). Consider first $(x,\lambda) \in \mathcal{R}_f$, $x \notin X_h^{3\kappa}$. Then, $x \in R_f$ and thus $x \notin \operatorname{sing\,supp} f$. In particular, $x \notin \operatorname{sing\,supp} f^s$. Moreover, since $\nabla_x f \neq 0$, we conclude that $x \in \operatorname{ess\,supp}(f)$ and therefore $x \in \operatorname{ess\,supp}(f^s)$. Using Theorem 4.2.4, we conclude that $(x,\lambda) \in \mathrm{WF}(\mathsf{ReLU}_{\kappa,\phi_\kappa}(f)) = \mathrm{WF}(f^s\,\mathsf{H}(f))$ if and only if $(x,\lambda) \in \mathrm{WF}(\mathsf{H}(f))$. Since $x \notin C_f \cup S_f$, we conclude from Proposition 4.2.2 that (x,λ) satisfies (4.2.2). To show the converse embedding, assume that (x,λ) is such that $x \in R_f$ and (x,λ) satisfies (4.2.2). By Proposition 4.2.2, we have that $(x,\lambda) \in \mathrm{WF}(\mathsf{H}(f))$. Furthermore, $x \notin \operatorname{sing\,supp} f$ and $\nabla_x f \neq 0$, which implies that $x \in \operatorname{ess\,supp} f$. We conclude by Theorem 4.2.4 that $(x,\lambda) \in \mathrm{WF}(f^s\,\mathsf{H}(f)) = \mathrm{WF}(\mathsf{ReLU}_{\kappa,\phi_\kappa}(f))$. This shows (4.2.16).

To show (4.2.17) it suffices to observe that $x \in R_f$ implies that $x \notin \operatorname{sing\,supp} f$ and hence $x \notin \operatorname{sing\,supp}(1-\theta_\kappa)f$. Therefore, we conclude that

$$\mathrm{R}_{f,\kappa} \cap (X_h^{3\kappa} \times \mathbb{S}^1) \subset \mathrm{WF}(\mathsf{H}(f)) \cap (R_f \times \mathbb{S}^1) \cap (X_h^{3\kappa} \times \mathbb{S}^1) = \mathcal{R}_f \cap (X_h^{3\kappa} \times \mathbb{S}^1),$$

where the last equality follows from Proposition 4.2.2. These yields (4.2.17).

The full result now follows by considering the decomposition (4.2.11). The part associated with $\mathrm{A}_{f,\kappa}$ is estimated via (4.2.12) and (4.2.13). The part associated with $\mathrm{B}_{f,\kappa}$ vanishes due to (4.2.14). Finally, the part associated with $\mathrm{D}_{\kappa,\phi_\kappa}$ is estimated via the decomposition (4.2.15), where the $\mathrm{R}_{f,\kappa}$ part is estimated via (4.2.16) and (4.2.17) and $\mathcal{CS}_f = \mathrm{CS}_{f,\kappa}$ holds per definition. □

Remark 4.2.6. *Theorem 4.2.5 only presents an estimate for the wavefront set associated with the set* $(X_h^\kappa)^c$*. In the sequel, since we are using the continuum relations to obtain certain properties of digital relations, we will assume that* κ *is chosen very small and* X_h^κ *is not seen by the discretization. In other words, in practice, we compute the wavefront set only via* (4.2.8)*.*

Although on $(X_h^\kappa)^c$, Theorem 4.2.5 does not entirely free us from computing $\mathrm{WF}(\mathsf{ReLU}_{\kappa,\phi_\kappa}(f))$, it restricts the necessity for such a computation to the cases where $x \in S_f \cup C_f$. The set $\mathcal{CS}_f$ can also be further split up: For $x \in \mathbb{R}^2 \setminus X_h^\kappa$, we denote by $\mathrm{WF}(f)_x$ the x-slice of the wavefront set of f defined as $\Lambda \subset \mathbb{R}^2$ such that $(x,\lambda) \in \{x\} \times \Lambda$ for all $(x,\lambda) \in \mathrm{WF}(f)$.

Proposition 4.2.7 ([9, Proposition 4.13])**.** *Let* $\Omega \in \mathbb{R}^2$ *be open,* $f \in \mathcal{S}'(\Omega)$ *be a distribution and* $\mathcal{CS}_f$ *be as in Theorem 4.2.5. Then*

$$\begin{aligned}
&\mathcal{CS}_f \bigcap (X_h^\kappa \times \mathbb{S}^1)^c = \Big\{(x,\lambda) : x \in C_g \setminus X_h^\kappa \textit{ and } (x,\lambda) \in \mathrm{WF}\big(\mathsf{H}(f)\big)\Big\} \\
&\bigcup \Big\{(x,\lambda) : x \in S_g \setminus X_h^\kappa, \mathrm{WF}(f)_x \cap -\mathrm{WF}\big(\mathsf{H}(f)\big)_x = \varnothing, \lambda \in (\mathrm{WF}(f)_x + \mathrm{WF}\big(\mathsf{H}(f)\big)_x) \setminus \{0\}\Big\} \\
&\bigcup \Big\{(x,\lambda) : x \in S_g \setminus X_h^\kappa, \mathrm{WF}(f)_x \cap -\mathrm{WF}\big(\mathsf{H}(f)\big)_x \neq \varnothing, (x,\lambda) \in \mathrm{WF}\big(\mathsf{ReLU}_{\kappa,\phi_\kappa}(f)\big)\Big\}.
\end{aligned} \tag{4.2.18}$$

Proof. The result follows immediately from Theorem 4.2.4. □

Theorem 4.2.5 presents two ways to estimate the wavefront set of $\mathsf{ReLU}_{\kappa,\phi_\kappa}(f)$ on $(X_h^\kappa)^c$. On the one hand, it can be precisely computed by (4.2.8). This, however, may require us to compute $\mathcal{CS}_f$ via Proposition 4.2.7. This computation could be performed according to (4.2.18), by using a method such as DeNSE to find $\mathrm{WF}(\mathsf{ReLU}_{\kappa,\phi_\kappa}(f))$ if required.

Next, the wavefront set on $(X_h^\kappa)^c$ can be estimated using (4.2.10). Since we expect that it is not problematic to overestimate the wavefront set slightly, we decided to make use of the second option and cast this wavefront set extraction algorithm as Algorithm 1.

4.2.5 Microlocal analysis of the residual layer and sum-taking

In the continuum setting, residual neural networks are operators $\mathcal{H} \colon \mathcal{S}'(\Omega) \to \mathcal{S}'(\Omega)$ of the form

$$\mathcal{H}(f) = f + \mathcal{F}(f), \tag{4.2.19}$$

where $\mathcal{F} \colon \mathcal{S}'(\Omega) \to \mathcal{S}'(\Omega)$ is continuum two-dimensional convolutional ResNet according to Definition 4.1.15. In addition, in Subsection 4.1.3, we allow summing over the channels of a convolutional block.

Algorithm 1: Wavefront set classifier of $\mathsf{ReLU}_{\kappa,\phi_\kappa}(f)$.

```
Input: Distribution $f \in \mathcal{S}'(\Omega)$, $\mathrm{WF}(f)$, $x \in \Omega$.
Output: Estimate $\mathrm{WF}(\mathsf{ReLU}_{\kappa,\phi_\kappa}(f))_x \subset \Omega$.
initialisation;
if $x \in \mathrm{supp}_+(f)^{\mathrm{o}}$ then
    return $\Lambda_x = \mathrm{WF}(f)_x$;
end
if $x \in R_f$ then
    return $\Lambda_x = \{\pm\nabla_x(f)/\|\nabla_x(f)\|\}$;
end
if $x \in C_f \cup S_f$ then
    return $\Lambda_x = \Omega$;
end
```

Because of this, we would like to identify $\mathrm{WF}(f+g)$ for two distributions with known wavefront set. The following two results yield a complete description thereof.

Theorem 4.2.8 ([92, Page 93])**.** *Let $\Omega \subset \mathbb{R}^2$ be open, let $f, g \in \mathcal{S}'(\Omega)$ and $(x;\lambda)$ be a regular directed point of f and g, then $(x;\lambda)$ is a regular directed point of $f+g$. In particular, if $(x,\lambda) \in \mathrm{WF}(f)$ and $(x;\lambda)$ is a regular directed point of g, then $(x,\lambda) \in \mathrm{WF}(f+g)$.*

Corollary 4.2.9 ([9, Corollary 4.15])**.** *Let $\Omega \subset \mathbb{R}^2$ be open and let $f, g \in \mathcal{S}'(\Omega)$, then $\mathrm{WF}(f+g)$ is given by*

$$\mathrm{WF}(f+g) = \mathcal{A}_f \cup \mathcal{A}_g \cup \mathcal{A}_{f+g}, \tag{4.2.20}$$

where

$$\mathcal{A}_f := \{(x;\lambda) \in \mathrm{WF}(f) : x \notin \mathrm{WF}(g)\}, \qquad \mathcal{A}_{f+g} := ((\mathrm{WF}(f) \cap \mathrm{WF}(g))) \cap \mathrm{WF}(f+g),$$
$$\mathcal{A}_g := \{(x;\lambda) \in \mathrm{WF}(g) : x \notin \mathrm{WF}(f)\}$$

In particular, we have that

$$\mathcal{A}_{f+g} \subset \mathrm{WF}(f) \cap \mathrm{WF}(g). \tag{4.2.21}$$

Proof. The result follows immediately from Theorem 4.2.8. □

In a similar fashion as we did for the wavefront set of $\mathsf{ReLU}(f)$, using Corollary 4.2.9 we can find two ways to extract the wavefront set of $f+g$ via Corollary 4.2.9. We express the one that yields a superset of the wavefront set of $f+g$ via (4.2.21) in the form of Algorithm 2.

Algorithm 2: Wavefront set classifier of $f + g$.

```
Input: Distribution $f, g \in \mathcal{S}'(\Omega)$, $\mathrm{WF}(f), \mathrm{WF}(g)$, $x \in \Omega$.
Output: Estimate $\mathrm{WF}(f+g)_x \subset \Omega$.
initialisation;
if $x \in \mathrm{WF}(f) \cap \mathrm{WF}(g)^c$ then
    return $\Lambda_x = \mathrm{WF}(f)_x$;
end
if $x \in \mathrm{WF}(f)^c \cap \mathrm{WF}(g)$ then
    return $\Lambda_x = \mathrm{WF}(g)_x$;
end
if $x \in \mathrm{WF}(f) \cap \mathrm{WF}(g)$ then
    return $\Lambda_x = \mathrm{WF}(g)_x \cup \mathrm{WF}(f)_x$;
end
```

4.2.6 Microlocal analysis of continuum residual neural network

Let us first notice that the continuum residual neural network operator as presented in Definition 4.1.15 has four basic components, (1) the differential (convolutional) layers, (2) summation over channels, (3) application of the ReLU, and (4) the residual connection. We have seen in Section 4.2.5 that the effect on the wavefront set through summation over channels and the residual connection can be described as the output of Algorithm 2. In addition, the effect of the differential layers is described by (4.2.1) and the wavefront set after an application of the ReLU can be found through an application of Algorithm 1. Overall, there is an algorithm that produces for every continuum convolutional ResNet and every input distribution of which the wavefront set is known, an estimate of the wavefront set of the output.

In the next section we will study the microlocal behavior of a neural network architecture which uses convolutional ResNet blocks as its backbone. The *learned primal-dual architecture* was introduced by Adler and Öktem in [3] and is mainly used for image reconstruction tasks associated to inverse problems in particular tomographic reconstruction. Later in Chapter 7 we present a method that makes use of the learned primal-dual architecture to performs tomographic reconstruction, and in addition uses the wavefront set of its input is used as a strong-prior.

4.3 Continuum learned primal-dual architecture

In Section 1.3 we introduced the notion of hybrid methods for inverse problems, which are designed to solve a problem of the form

$$g = \mathcal{A}(f_{\mathrm{true}}) + \delta g, \tag{4.3.1}$$

aims to recover the ground truth $f_{\mathrm{true}} \in X$ from data measurements $g \in Y$. Both X (reconstruction space) and Y (data space) are Hilbert spaces. Moreover, $\mathcal{A} : X \to Y$

is the forward operator and $\delta g \in Y$ is the noise. As we know these methods combine model-based and data-driven approaches, where the data-part can be regarded as a deep neural network.

In this thesis we will focus on a specific type of hybrid method, namely *learned iterative schemes*, and in particular, the *learned primal-dual algorithm* based on the learned primal-dual architecture [3]. The learned primal-dual architecture is based on primal-dual optimization schemes coming from non-linear programming [25]. In this section we introduce the learned primal-dual architecture for inverse problems, in the continuum setting. In addition, we also discuss how the analysis from Section 4.1 can be used to study the microlocal behavior of this architecture. In our context, the use of this architecture is motivated by two main reasons, the state-of-the-art performance in tomographic reconstruction and its simple construction based on ResNet blocks. The fact that the main backbone of this architecture is based on the ResNet architecture allows us to directly extract its microlocal behavior using the techniques presented in Section 4.1.

The learned primal-dual algorithm is a hybrid method, since it combines ideas from model-based and data-driven regularization. Let us start with the model-based part of the algorithm. A classical model-based technique for solving (4.3.1) is the variational regularization approach. In this approach one seeks to minimize a regularized objective functional by solving

$$\min_{f\in X} \left(\mathcal{L}(\mathcal{A}(f), g) + \lambda\mathcal{S}(f)\right) \quad \text{for a fixed} \quad \lambda \geq 0, \tag{4.3.2}$$

where $S : X \to \mathbb{R}$ is the regularization functional, $\mathcal{L} : Y \times Y \to \mathbb{R}$ is the loss function, and λ is the regularization parameter. On the one hand, the regularization functional encodes a priori information about the ground truth f_{true} (e.g. sparsity or regularity). On the other hand, λ controls the influence of the a-priori knowledge. In this context, $\mathcal{L} : Y \times Y \to \mathbb{R}$ is a suitable affine transformation of the data log-likelihood [13].

In imaging, the minimization in (4.3.2) is a large-scale optimization problem, traditionally solved using gradient-based methods such as gradient descent. In gradient descent one uses the fact that the gradient of a function defines the direction of maximum growth to minimize 4.3.2 by updating the function in the direction of the negative gradient, i.e.,

```
Algorithm 3: Gradient descent algorithm
  Input: Function $f_0 \in X$, WF($f$) and learning rate $h > 0$.
  Output: Minimizing function $f_I$.
  for $i = 1, \dots, I-1$ do
    $f_{i+1} \longleftarrow f_i - h\nabla\left(\mathcal{L}(\mathcal{A}(f_i), g) + \lambda\mathcal{S}(f_i)\right)$
  end
```

The main issue with Algorithm 3 is not just that sometimes a-priori information is hard to encode in terms of a functional, but that many regularizers of interest result in a non-differentiable objective functional. This issue does not allow us to compute $\nabla\left(\mathcal{L}(\mathcal{A}(f_i), g) + \lambda\mathcal{S}(f_i)\right)$. As an alternative approach to this problem, there exist different methods coming from non-smooth convex optimization, of particular importance are *proximal* methods, which have been introduced in order to work with non-smooth objective functionals.

In *proximal* methods, a proximal step replaces the gradient step. Let us first define a proximal operator.

Definition 4.3.1. *Let X be a Hilbert space and $\mathcal{G} : X \to \mathbb{R}$ a potentially non-smooth functional, and $\tau \in \mathbb{R}^+$ a step size. Then, the proximal operator of $\mathcal{G}$ given the step size τ is defined as*

$$prox_{\tau\mathcal{G}}(f) = \arg\min_{f' \in X} \left(\mathcal{G}(f') + \frac{1}{2\tau} ||f' - f||_X^2 \right), \quad \textit{for every } f \in X. \tag{4.3.3}$$

Notice that Equation (4.3.3) is well-defined even for non-smooth functionals $\mathcal{G}$. Therefore, one can think of the proximal operator as an approximation to the gradient update step for the non-smooth case. A classical application of Definition 4.3.1 is the proximal point algorithm [25] for minimizing an objective functional $\mathcal{G} : X \to \mathbb{R}$, where the iterations are given by

$$f_{i+1} = \text{prox}_{\tau\mathcal{G}}(f_i). \tag{4.3.4}$$

Although one could use this algorithm directly to solve (4.3.2), this is generally impractical, since (4.3.3) does not have a closed-form solution [25]. Proximal primal-dual schemes were introduced as a solution for this issue. To discuss this notion, let us first introduce the concept of *dual space* and operator adjoint in the context of Hilbert spaces.

Definition 4.3.2. *Let X be a Hilbert space with norm $|| \cdot ||_X$. Then, the* dual space *of X, namely, X^* is given by*

$$X^* := \{F : X \to \mathbb{R} : F \textit{ is bounded and linear}\}.$$

In addition, the dual space *is equipped with the norm $|| \cdot ||_{X'} : X' \times X' \to \mathbb{R}$ defined as*

$$||F||_{X'} := \sup_{f \in X} \{|F(f) : ||f||_X = 1\}.$$

Definition 4.3.3. *Let X and Y be two Hilbert spaces with inner products $\langle \cdot, \cdot \rangle_X : X \times X \to \mathbb{R}$ and $\langle \cdot, \cdot \rangle_Y : Y \times Y \to \mathbb{R}$ respectively. In addition, let $\mathcal{A} : X \to Y$ be a linear operator. Therefore, the* adjoint *of A, namely, $A^* : Y \to X$ is the linear operator fulfilling*

$$\langle A(f), g \rangle_X = \langle f, A^*(g) \rangle_Y, \quad \textit{for every } f \in X \textit{ and } g \in Y.$$

In primal-dual schemes, an auxiliary *dual* variable in the range of the forward operator is introduced and the *primal* ($f \in X$) and dual variables are updated in an alternating manner. This is mainly motivated by the equivalence between the minimization of a functional (the primal problem) and the maximization of its adjoint (the dual problem). By moving towards the negative gradient direction of the primal and the positive gradient of the dual, one can converge to the desired minimum faster [25].

One of the most used primal-dual schemes is the primal-dual hybrid gradient (PDHG) algorithm [25], also known as the Chambolle-Pock algorithm, named after the authors. The scheme is adapted for minimization problems with the following structure:

$$\min_{f\in X}\left(\mathcal{F}(\mathcal{T}(f))+\mathcal{G}(f)\right), \tag{4.3.5}$$

where $\mathcal{T}: X\to Y$ is an operator, which can be non-linear. Here, the *primal space* X and the *dual space* Y are Hilbert spaces. In addition, $\mathcal{F}: Y\to\mathbb{R}$ and $\mathcal{G}: X\to\mathbb{R}$ are functionals on the dual/primal spaces. Notice that (4.3.2) is an special case of (4.3.5) if we set $\mathcal{F} := \mathcal{L}(\cdot, g)$, $\mathcal{T}=\mathcal{A}$ and $\mathcal{G} := \mathcal{S}$. This scheme is given by Algorithm 4.

Algorithm 4: Non-linear primal-dual algorithm [3]

```
Input: Constants $\sigma, \tau > 0$ s.t. $\sigma\tau||\mathcal{A}|| < 1$, $\gamma \in [0,1]$ and functions $f_0 \in X$,
       $h_0 \in Y$.
Output: Primal solution $f_I \in X$ and dual solution $h_I \in Y$.
for $i = 1, \dots, I-1$ do
    $h_{i+1} \longleftarrow \mathrm{prox}_{\sigma\mathcal{L}^*}(h_i + \sigma\mathcal{A}(\overline{f}_i))$;
    $f_{i+1} \longleftarrow \mathrm{prox}_{\tau\mathcal{S}}(f_i - \tau[\partial\mathcal{A}(f_i)]^*(h_{i+1}))$;
    $\overline{f}_{i+1} \longleftarrow f_{i+1} + \gamma(f_{i+1} - f_i)$;
end
```

In Algorithm 4, $\mathcal{L}^*: X' \to \mathbb{R}$ is the *convex conjugate* of $\mathcal{L}(\cdot, g): X\to\mathbb{R}$ defined as

$$\mathcal{L}^*(F) = \sup\{F(f) - \mathcal{L}(f,g) : f\in X\}, \quad \text{for every } F\in X'.$$

In addition, $f_i \in X$ and $h_i \in Y$ are the dual and primal variable at step i, respectively. Finally, $[\partial\mathcal{A}(f_i)]^* : Y\to X$ is the adjoint of the derivative of $\mathcal{A}$ at the point f_i (see Definition 4.3.3). Knowing the advantages of the primal-dual algorithm over other standard methods, one would like to make use of its structure but at the same time boost its performance by a data-driven approach. In addition, one would like to introduce a-priori information based on the physics of the problem. Also, the update steps present in Algorithm 4 are based on predefined steps σ and τ. If such steps are too small, the algorithm will take a lot of iterations to converge, but if they are too large, the algorithm will oscillate around the optimal points.

In [3], Öktem and Adler derived a learned reconstruction scheme inspired by PDHG depicted in Algorithm 4, obtaining the learned primal-dual algorithm (see Algorithm 5). For that purpose, they followed the observation in [57, 96], that proximal operators can be replaced by other operators that are not necessarily proximal operators. In the learned primal-dual algorithm, the proximal operators of the PDHG are replaced by operators parametrized by a neural network so the parameters are learned from training data, resulting in a learned reconstruction operator. In this approach the authors use convolutional residual neural networks to learn the primal and dual updates, avoiding the complication of choosing right steps σ and τ. They also make use of the forward operator $\mathcal{A}$ which is the most powerful prior we have. In addition, this algorithm separates the problem in data update (dual) and image update (primal), so the neural networks that define each step just need to learn local information, while the forward operator and its adjoint cover the global aspect of the problem

The learned primal-dual algorithm depicted in Algorithm 5 can be seen as a neural network realization of the classical PDHG algorithm with a few modifications, guided by recent advances in machine learning. More precisely, we observe the following

- Instead of working on the primal space X, extend the primal space to allow the algorithm some *memory* between the iterations
$$f = [f^{(1)}, f^{(2)}, \dots, f^{(N_{\text{primal}})}] \in X^{N_{\text{primal}}}, \quad \text{where } N_{\text{primal}} \in \mathbb{N}.$$
Similarly, one extends the dual space U to $U^{N_{\text{primal}}}$.
- Instead of explicitly enforcing updates of the form $h_i + \sigma\mathcal{K}(\overline{f}_i)$, allow the network to learn how to combine the previous update with the result of the operator evaluation.
- Instead of hard-coding the over-relaxation $\overline{f}_{i+1} \leftarrow f_{i+1} + \theta(f_{i+1} - f_i)$, let the network freely learn in what point the forward operator should be evaluated.
- Instead of using the same learned proximal operators in each iteration allow them to differ.

Algorithm 5: Learned primal-dual algorithm [3]

```
Input: Functions $f_0 \in X$, $g, h_0 \in Y$.
Output: Primal solution $f_I \in X$ and dual solution $h_I \in Y$.
for $i = 1, \dots, I-1$ do
    $h_{i+1} \longleftarrow \Gamma_{\theta^d_{i+1}}(h_i, \mathcal{A}(f_i), g)$;
    $f_{i+1} \longleftarrow \Lambda_{\theta^p_{i+1}}(f_i, [\partial\mathcal{A}(f_i)]^*(h_{i+1}))$;
end
```

In Algorithm 5 the primal ($\Lambda_{\theta^p_i}$) and dual ($\Gamma_{\theta^d_i}$) proximal operators are parametrized by convolutional residual operators of the form (4.1.11). Notice that the associated proximal operators $\Lambda_{\theta^p_i}$ and dual $\Gamma_{\theta^d_i}$ as well as the forward operator $\mathcal{A}$ are defined in X. Moreover, in the original work by Adler et al. they used biases for the operators $\mathcal{W}_{\boldsymbol{\theta}_j}$ and parametric ReLU (PReLU) as non-linearities in (4.1.11). For our purposes, we make use of unbiased operators and ReLU since we have found that they will perform like the original architecture, this is shown in Chapter 8.

Based on Definition 4.1.15 for the continuum ResNet, we can now define the continuum Learned Primal-Dual network as in Algorithm 5 with operators $\Lambda_{\theta^p_i}$ and $\Gamma_{\theta^d_i}$ given by continuum ResNets. Hence, continuum Learned Primal-Dual network is a mapping

$$\mathcal{A}^\dagger{}_\theta \colon \mathcal{S}'(\Xi) \to \mathcal{S}'(\Omega) \quad \text{where} \quad \mathcal{A}^\dagger{}_\theta(g) := f_N \quad \text{and}$$

with $f_N \in \mathcal{S}'(\Omega)$ given by the N-step iterative scheme in Algorithm 6 in which Λ_i and Γ_i are continuum two-dimensional convolutional ResNets as in Definition 4.1.15 and θ represents the weighst of the neural network.

Algorithm 6: Continuum Learned Primal-Dual network

```
**Input:** $\Omega \subset \mathbb{R}^2$, $\Xi \subset \mathbb{R} \times (0,\pi)$, $f_0 \in \mathcal{S}'(\Omega)$, $h_0 \in \mathcal{S}'(\Xi)$ and $g \in \mathcal{S}'(\Xi)$.
**Output:** Primal solution $f_N \in \mathcal{S}'(\mathbb{R}^2)$ and dual solution $h_N \in \mathcal{S}'(\Xi)$.
**for** $i = 1, \dots, N-1$ **do**
    $h_{i+1} \longleftarrow \Gamma_i(h_i, \mathcal{A}(f_i), g)$;
    $f_{i+1} \longleftarrow \Lambda_i(f_i, [\partial\mathcal{A}(f_i)]^*(h_{i+1}))$;
**end**
```

Figure 4.3 depicts the learn primal-dual with iterations $I = 10$.

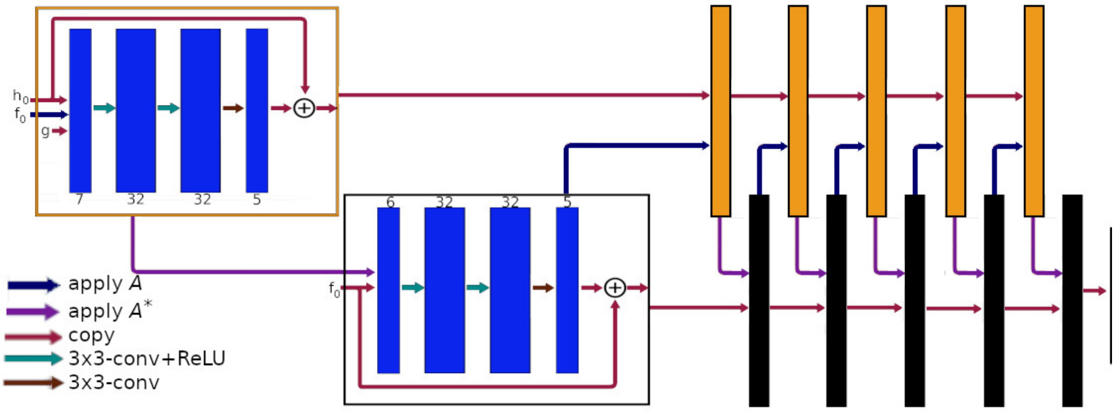

Figure 4.3: Learned primal-dual architecture from Algorithm 5.

Both the primal residual blocks (Figure 4.4) and the dual residual blocks (Figure 4.5) are of the form (4.1.11), involving solely convolutional operators $\mathcal{K}_{\boldsymbol{\theta}_j^{l,k}}(f^{(k)})$, the residual layer and the ReLU operator. In addition the dual and primal blocks are connected by the Radon transform $\mathcal{A}$ and its adjoint $\mathcal{A}^*$, in this thesis the forward operator is given by the Radon transform $\mathcal{R} : \mathcal{S}'(\Omega) \to \mathcal{S}'(\Xi)$ (Definition 2.3.1) and its adjoint, the *back-projection* $\mathcal{R}^* : \mathcal{S}'(\Xi) \to \mathcal{S}'(\Omega)$ (Definition 2.3.2). Therefore, the microlocal behavior of each individual component (convolutional layers, residual layers and non-linearities) can be analyzed by the results presented in Sections 4.1.1, 4.2.5 and 4.1.2. In addition, the microlocal behavior of the Radon transform and its adjoint is discussed in Section 2.5

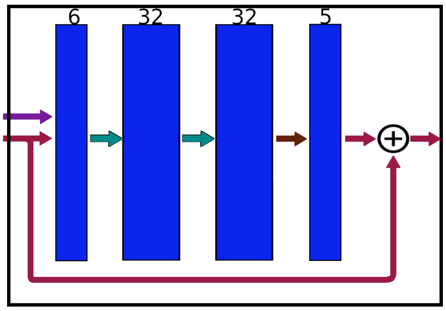

Figure 4.4: Primal convResNet block of learned primal dual architecture.

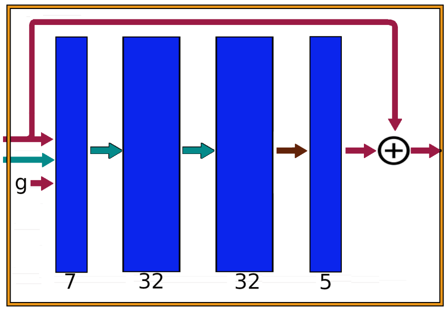

Figure 4.5: Dual convResNet block of learned primal dual architecture.

The microcanonical relation of the convolutional operators, ReLU, the residual layer, the Radon transform and its adjoint are given by (4.2.1), (4.2.8), (4.2.20), (2.5.7), and (2.5.11), respectively. This allows us to have a microcanonical relation for the entire learned primal-dual architecture in the continuum case, given by the composition of the microcanonical relations across the dual and primal blocks. We will present the explicit form for the digital case in Chapter 6.

Naturally, we would like to apply this theory to real-world problems. For this, in Chapter 6 we introduce a novel technique to digitize the analysis presented in this chapter. This is then applied in the context of task-adapted reconstruction for computed tomography in Chapter 7. Finally, the numerical experiments of the task-adapted reconstruction method is presented in Chapter 8. In Chapter 8 we will also present the numerical experiments regarding the digital wavefront set extraction (Chapter 5) and the digital wavefront set propagation in convolutional ResNets (Chapter 6). Before can to digitize the microlocal analysis of convolutional residual neural networks, we need to introduce the concept of a digital wavefront set and how to compute it. This is the goal of the next chapter.

5 Digital wavefront set extraction with shearlets and convolutional neural networks

In Chapter 3 we introduced the mathematical machinery of microlocal analysis and its relation with multiscale directional systems coming from harmonic analysis. In addition, we also discussed the main role of microlocal analysis in the understanding of the propagation of singularities under the action of Fourier integral operators, such as the Radon transform or convolutional operators. This machinery will allow us later in Chapter 6 to understand how wavefront sets are propagated in convolutional residual neural networks used for real-world applications, such as tomographic reconstruction. Before we can do that, in this chapter we will introduce the notion of digital wavefront set extraction. In addition we will also present an algorithm that combines the digital shearlet transform and a convolutional neural network architecture to extract wavefront set of a digital image. Moreover, we will introduce a general setting of shearlet-based semantic edge detection algorithms on which wavefront set extraction is a particular example. We will show that the introduction of shearlets as a feature extraction performs heavy lifting on general edge detection and classification.

As we know, large amounts of information are necessary to describe a signal is contained in the wavefront set. This makes the wavefront set suitable as a-priori information for any inverse problem reconstruction, as long as such inverse problems have a Fourier integral operator as forward model. In particular, in modern methods for regularization of inverse problems, one makes use of neural networks to define a subset of the model parameters. In Chapter 8 we will present a set of methods that use the wavefront set to regularize an inverse problem method for tomographic reconstruction. This is done by propagating the wavefront set of the data, through a neural network, in order to perform a task on the output, improving the reconstruction.

Before we proceed, we should notice that all the theory presented so far corresponds to the continuous setting, where the signals live in an infinite-dimensional space. Moreover, the definition of wavefront set, as well as the shearlet-based wavefront set resolution, involves the computation of an infinite number of Fourier and shearlet coefficients, respectively, whereas, in reality, we have access just to a limited (mostly small and finite) number of coefficients. Although the wavefront set cannot be naturally defined for digital signals, this does not imply that one is not able to describe the projection of the singularities to a discrete grid. For example, edge detection plays an important role in computer vision, where digital images, and where edges (discontinuity between objects) can be regarded as singularities. In the case of wavefront sets, we need to describe not just the location of such edges, but also their orientations.

In this chapter, we introduce the concept of a digital wavefront set for digital images. Using the ideas presented in Chapter 3, we also introduce an algorithm that makes use of the digital shearlet coefficients of an image to extract its digital wavefront set using a deep convolutional neural network. In addition, Section 5.3 presents the concept of the distracted supervision paradox, which can be applied to digital wavefront sets in the context of semantic edge detection.

My own contributions: This chapter is the product of lengthy discussions with my supervisor Gitta Kutyniok, and my collaborators Ozan Öktem and Phlipp Petersen. This ideas were mainly developed at the beginning on my PhD when we asked ourselves about the possibility of approximating the wavefront set of a distribution when digitized. The idea of using deep neural network classifier where the inputs are given by the digital shearlet coefficients was developed by me. These ideas were publish in our joint papers [8, 6]. The actual writing was mostly done by myself..

5.1 Wavefront set extraction

A necessary first step before applying techniques from microlocal analysis in real-world applications is to extract the wavefront set of functions. The wavefront set is the set of singularities and their orientations. Microlocal analysis studies how wavefront sets are transformed under the action of operators. In particular, if the operator is Fourier integral (Definition 2.4.4), it will not create new singularities, but just transformed them with a prescribed mapping (Definition 2.4.5). This gives you access to the singularities of $\mathcal{P}f$ without computing it, providing an useful tool in inverse problems involving FIOs. By the definition of the wavefront set, this involves the asymptotic analysis of the Fourier transform of the localized function in question (Definition 2.2.6).

5.1.1 Image space: distributions and functions

In Chapter 4 we have introduced a set of tools that allows us to propagate the wavefront set through the continuum convolutional residual neural network and the continuum learned primal-dual architecture. In our theory, we have assumed that the input of such architectures in the continuum setting are either Schwartz functions or tempered distributions. Since $\mathcal{S}(\mathbb{R}^2)$ is the dualspace of $\mathcal{S}'(\mathbb{R}^2)$, the wavefront set propagation results in both spaces are connected via duality. The main reason that we choose these spaces instead of the classical choice $L^2(\mathbb{R}^2)$ is the use of the differential operator $\mathcal{K}_{\boldsymbol{\theta}}$ (Definition 4.1.2) associated to the convolutional layers. Since such operator is interpreted as a differential operator, there is a need to compute derivatives of its input.

In this chapter we aim to digitize the notion of wavefront set extraction presented in Chapter 3, where we presented the shearlet based extraction for $L^2(\mathbb{R}^2)$ and $\mathcal{S}'(\mathbb{R}^2)$. For that purpose we need to define the image space, where the digitization acts upon. Definition 4.1.15 suggests that the image space in this case should be $\mathcal{S}'(\mathbb{R}^2)$. Unfortunately, the digitization of tempered distributions represents a key challenge that will significantly add computational complexity to our problem. For this reason, we have chosen $L^2(\mathbb{R}^2)$. The next remark shows that our results from Chapter 4 are still compatible with this choice.

Remark 5.1.1. *By Definiton 4.1.15, the continuum convolutional ResNet acts on the space of tempered distributions, $\mathcal{S}'(\Omega)$, where the differential operator $\mathcal{K}_{\boldsymbol{\theta}}$ is well-defined. Furthermore, notice that the rest of the related operators, namely,* ReLU *and the residual layers are well-defined in $L^2(\mathbb{R}^2)$. In the case of the continuum learned primal-dual architecture (Algorithm 6), the Radon transform and the back-projection are also well-defined in $L^2(\mathbb{R}^2)$ (Definitions 2.3.1 and 2.3.2).*

In the discrete case, the discrete convolutional operator $\mathcal{K}^d_{\boldsymbol{\theta}}$ (see (4.1.3)*). This operator is also well-defined in some discretization of functions in $L^2(\Omega)$, in particular $\ell_2(\Omega^d)$, where Ω^d is the discrete grid associated to the domain Ω. This allows us to choose the space $L^2(\Omega)$ as the image space where our discretization acts upon.*

5.1.2 Extraction from finitely many samples.

In applications, where only finitely many point samples of the underlying function are known, estimating the asymptotic behavior of the Fourier transform, required for the computation of the wavefront set, is usually not possible. Indeed, we show in Section 5.2 that *every method* that seeks to approximately and explicitly extract the wavefront from discrete samples of functions from a function class $\mathcal{C} \subset L^2(\mathbb{R}^2)$ *fails* on a dense subset of $\mathcal{C}$. This happens whenever $\mathcal{C}$ contains at least all $k-$times differentiable functions for arbitrary $k \in \mathbb{N}$, see Theorem 5.2.4.

The problem of extracting the wavefront set of functions from a function class $\mathcal{C}$ certainly becomes easier the smaller $\mathcal{C}$ gets. Indeed, in applications where, for example, $\mathcal{C}$ models a class of images, it is reasonable to assume that $\mathcal{C}$ is a very small subset of $L^2(\mathbb{R}^2)$ and does not contain all $k-$times differentiable functions for any $k \in \mathbb{N}$. Hence, in this situation, wavefront set extraction from point samples explicitly designed for the class of images could be feasible. In fact, if the set $\mathcal{C}$ is so small that every function in $\mathcal{C}$ is uniquely determined through its samples on the grid, then wavefront set detection could, in principle, be performed through a large database, which could be learned from $\mathcal{C}$. From the above, it is clear that a useful wavefront set extractor needs to be closely adapted to the underlying function class $\mathcal{C}$. This strong adaptation to $\mathcal{C}$ is referred to as our *guiding principle*.

5.1.3 Classification with applied harmonic analysis.

As presented in Chapter 3, certain transforms from applied harmonic analysis, like the curvelet and shearlet transform, offer an alternative possibility to identify the wavefront set. In particular, the connection between the behavior of these transforms and the wavefront set has been analyzed in [23, 72], with edge detection as a particular application in [120]. These approaches characterize the wavefront set through the rate of decay of the respective transforms (Theorems 3.4.2 and 3.4.9). In this way, one can transform the problem of extracting the wavefront set of a function to a classification problem on the decay of another function. While this point of view certainly makes the wavefront set more accessible, especially since it does not depend on an unspecified localization procedure, it still presents some limitations presented at the beginning of this section. In particular, it cannot produce a successful explicit wavefront set extractor acting on sampled functions, since evaluating the decay rate requires an infinite sequence of coefficients.

The task of detecting edges and classifying them is known as *semantic edge detection* [79] and has a considerable impact on computer vision. We will present this approach in detail in Section 5.5. Furthermore, we will also formalize semantic edge detection in the context of statistical decision theory in Section 5.5.

5.1.4 Data driven wavefront set extraction.

Having in mind the aforementioned guiding principle, a successful wavefront set extractor needs to be adapted to the function class of interest. The relevant function classes in applications are, however, difficult to characterize analytically. An alternative is therefore to adopt a *data-driven model* where the function class of interest is given empirically through examples.

Based on the above, we propose the following algorithm named *Deep Network Shearlet Edge Extractor* (DeNSE) [8]. First, we assemble a supervised training set consisting of images with their associated ground truth wavefront set, or a suitable surrogate. Second, we train a classifier—in our case a *deep neural network*—to predict the wavefront set from the *shearlet coefficients* of the training data. Finally, we apply the resulting classifier to unseen data. In this way, the algorithm combines two crucial elements. On the one hand, as mentioned previously, the interaction of the shearlet transform with the wavefront set is theoretically well understood and presents the microlocal information of a function in a more accessible way. On the other hand, the trained classifier allows a strong adaptation to the underlying function class, thereby complying with our guiding principle.

In Section 5.4, we present the construction of the algorithm as well as the training data that is used. In this context, we analyze the detection of edges and orientations in images as well as higher-order wavefront set detection in sinogram data. We will later show in Section 7.3, that our proposed method outperforms all conventional edge-orientation estimators as well as alternative data-driven methods including the current state-of-the-art. Moreover, we are unaware of any wavefront set extractor in the literature that goes beyond edge or ramp detection, so that the following analysis can be seen as the first advance in this direction.

5.2 Wavefront set of sampled functions

In this section we analyze whether it is possible to explicitly construct an operator that maps a finitely sampled function f to an estimate of its wavefront set. This typically arises in practical applications, e.g., images are only given as pixels representing point samples of a real-valued function. It is natural to use Shannon's sampling theorem to make the connection between a sampled function and its wavefront set more precise. The theorem is stated in Section 5.2.1 and –based on it– Section 5.2.2 introduces the notion of an approximate wavefront set extractor. Finally, in Section 5.2.3 we show that any approximate wavefront set extractor on a function class $\mathcal{C} \subset L^2(\mathbb{R}^2)$ that predicts the wavefront set of functions $f \in \mathcal{C}$ from a finite number of sample values fails on a dense subset of $\mathcal{C}$ if $\mathcal{C}$ contains at least all $C^k(\mathbb{R}^2) \cap L^2(\mathbb{R}^2)$ functions for any arbitrary $k \in \mathbb{N}$. This result holds even if the sampling density is allowed to depend on the function f.

5.2.1 Sampling theorem and Paley-Wiener spaces

The sampling theorem states that every band-limited function f can be written as a sum of shifted cardinal sine functions weighted by point samples of f. In other words, a band-limited function is fully determined by its values on a discrete grid. To give a precise statement, we introduce Paley-Wiener spaces.

Definition 5.2.1. *Given $\Lambda > 0$, the* Paley-Wiener space $\mathcal{PW}_\Lambda \subset L^2(\mathbb{R}^d)$ *is defined as*

$$\mathcal{PW}_\Lambda := \left\{ f \in L^2(\mathbb{R}^d) : \operatorname{supp}(\hat{f}) \subset [-\Lambda, \Lambda]^d \right\}.$$

We also define the $d-$dimensional sinc-function as

$$\operatorname{sinc}_d(x) := \prod_{i=1}^{d} \frac{\sin(\pi x_i)}{\pi x_i}, \quad \text{where} \quad x = (x_1, \dots, x_d) \in \mathbb{R}^d.$$

Having in mind the above notation, we now state the sampling theorem, see, e.g. [83].

Theorem 5.2.2 (Sampling theorem, [8])**.** *Let $d \in \mathbb{N}$, $f \in L^2(\mathbb{R}^d) \cap C(\mathbb{R}^d)$ and $\Lambda > 0$. Then*

$$f \in \mathcal{PW}_\Lambda \iff f(x) = \sum_{n \in \mathbb{Z}^d} f\left(\frac{n}{\Lambda}\right) \operatorname{sinc}_d(\Lambda \cdot x - k), \quad \text{for} \quad x \in \mathbb{R}^d.$$

Notice that by Theorem 5.2.2, if $f \in \mathcal{PW}_\Lambda$ then it is band-limited. Furthermore since $\operatorname{sinc}_2(m-n)$ vanishes for every $m, n \in \mathbb{Z}^2$ such that $m \neq n$, we observe that $f(m/\Lambda) = y_m$ for all $m \in \mathbb{Z}^2$, where y_m is a sequence on the grid. In other words, every sequence on a grid defines an associated interpolating band-limited function and conversely, every band-limited function is uniquely determined by its values on a discrete grid.

As a consequence, the problem of extracting the wavefront set of a function $f \in L^2(\mathbb{R}^2) \cap C(\mathbb{R}^2)$ from its discrete sampled values $\{f(m/\Lambda)\}_{m \in \mathbb{Z}^2}$ can be re-stated as extracting the wavefront set of f from its projection onto a Paley-Wiener space, i.e. $P_\Lambda(f) := P_{\mathcal{PW}_\Lambda}(f)$. Now, for functions $f \in L^2(\mathbb{R}^2)$, which are only defined almost everywhere, we can even use the sampling theorem as a definition of a point evaluation. For $f \in L^2(\mathbb{R}^2)$, we have that $P_\Lambda(f) \in C(\mathbb{R}^2)$. Hence, we set $f(m/\Lambda) := P_\Lambda(f)(m/\Lambda)$.

5.2.2 Wavefront set extractors

As already stated, the problem of extracting the wavefront set from samples on a grid is equivalent to extracting the wavefront set given the projection onto a Paley-Wiener space. There are multiple conceivable notions of a wavefront set extractor. First, for $\Lambda > 0$, we could ask for a map

$$\mathrm{DWF}_\Lambda : \mathcal{PW}_\Lambda \to 2^{\mathbb{R}^2 \times \mathbb{S}^1} \quad \text{such that } \mathrm{DWF}(P_\Lambda f) = \mathrm{WF}(f) \text{ for all } f \in L^2(\mathbb{R}^2), \quad (5.2.1)$$

where $2^{\mathbb{R}^2 \times \mathbb{S}^1}$ denotes the power set of $\mathbb{R}^2 \times \mathbb{S}^1$. Essentially, this map requests to extract the wavefront set of a function f from its samples on a fixed grid. It is clear that such a map, DWF_Λ, cannot exist, since it fails for functions f that have fine structures which cannot be detected by coarse sampling. For example, a function that vanishes on every grid point of $\mathbb{Z}/\Lambda$ while having a non-trivial wavefront set would be classified the same as the zero function.

A more reasonable model for a wavefront set extractor should give an approximate prediction of the wavefront set that eventually improves as the sampling density increases. To weaken this statement even further, we might only ask for approximate extraction of the wavefront set at one point and only for functions from a fixed-function class $\mathcal{C} \subset L^2(\mathbb{R}^2)$. For a fixed set $W \subset \mathbb{R}^2 \times \mathbb{S}^1$ and a point $x \in \mathbb{R}^2$, we define

$$W_x := \big\{\xi \in \mathbb{S}^1 : (x; \xi) \in W\big\}.$$

We can now model the approximation described above by considering a sequence of wavefront set extractors given by

$$\mathrm{DWF}_j : \mathcal{PW}_j \to 2^{(\mathbb{R}^2 \times \mathbb{S}^1)} \quad \text{for } j \in \mathbb{N} \qquad (5.2.2)$$

such that, for fixed $x \in \mathbb{R}^2$, and all $f \in \mathcal{C}$,

$$d_H(\overline{\mathrm{DWF}_j(P_j(f))_x}, \overline{\mathrm{WF}(f)_x}) \longrightarrow 0, \qquad (5.2.3)$$

where $P_j(f)$ is the projection of f into the Paley-Wiener space $\mathcal{PW}_j$. Here d_H denotes the Hausdorff distance with the convention $d_H(X, \varnothing) = d_H(\varnothing, X) := 1$ for any non-empty compact subset $X \subset \mathbb{S}^1$ and $d_H(\varnothing, \varnothing) := 0$. Recall that with this definition d_H is a metric on compact subsets of $\mathbb{S}^1$ (including the empty set). A sequence as in (5.2.2) satisfying (5.2.3) results in an approximate extraction of the wavefront set of $f \in L^2(\mathbb{R}^2)$ at $x \in \mathbb{R}^2$ from point samples of f where the sampling density may depend on f. This observation motivates the following definition.

Definition 5.2.3. *Let $\mathcal{C} \subset L^2(\mathbb{R}^2)$. A sequence $\{\mathrm{DWF}_j\}_{j\in\mathbb{N}}$ of mappings as in* (5.2.2) *is called an* approximate wavefront set extractor. *We say that an approximate wavefront set extractor is*

- clairvoyant *at $x \in \mathbb{R}^2$ if the sequence satisfies* (5.2.2) *at x for all $f \in \mathcal{C}$, and*
- ignorant *to $f \in \mathcal{C}$ at $x \in \mathbb{R}^2$ if $d_H(\overline{\mathrm{DWF}_j(P_j(f))_x}, \overline{\mathrm{WF}(f)_x}) \not\to 0$ as $j \to \infty$.*

5.2.3 Non-existence of clairvoyant approximate wavefront set extractors

In the following we observe that, for every $x \in \mathbb{R}^2$, there is no clairvoyant approximate wavefront set extractor if $\mathcal{C}$ contains at least all k-times differentiable functions, for some $k \in \mathbb{N}$. Furthermore, in this situation, every approximate wavefront set extractor is ignorant to a dense subset of $\mathcal{C}$ at x.

Theorem 5.2.4. *Let $k \in \mathbb{N}$ and $\mathcal{C} \subset L^2(\mathbb{R}^2)$ be such that $\mathcal{C} \supset C^k(\mathbb{R}^2) \cap L^2(\mathbb{R}^2)$. For every $x \in \mathbb{R}^2$ we have that, for every approximate wavefront extractor $\{\mathrm{DWF}_j\}_{j\in\mathbb{N}}$, there exists a dense subset $\mathcal{M} \subset \mathcal{C}$ such that $\{\mathrm{DWF}_j\}_{j\in\mathbb{N}}$ is ignorant to all $f \in \mathcal{M}$ at x. In particular, no approximate wavefront set extractor is clairvoyant at x.*

Proof. The proof proceeds in two steps. First, for a given approximate wavefront set extractor, $(\mathrm{DWF}_j)_{j\in\mathbb{N}}$, and a point $x \in \mathbb{R}^2$, we construct a function $q \in \mathcal{C}$ such that $(\mathrm{DWF}_j)_{j\in\mathbb{N}}$ is ignorant to q at x. Second, we show that the set of such functions is dense in $\mathcal{C}$. **Step 1:** Notice that Definition 2.2.6 implies

$$\mathrm{WF}(f_1 + f_2) = \mathrm{WF}(f_1) \quad \text{for every } f_1 \in L^2(\mathbb{R}^2) \text{ and } f_2 \in C^\infty(\mathbb{R}^2) \cap L^2(\mathbb{R}^2). \tag{5.2.4}$$

For $x \in \mathbb{R}^2$, we now choose a function $g \in C^k \cap L^2(\mathbb{R}^2)$ such that for a non-empty set V we have that $\mathrm{WF}(g) \supset \{x\} \times V$. Since $k < \infty$, such a function always exists. We can also assume hat $\|(1+|\cdot|^2)^{k/2}\widehat{g}\|_{L^1(\mathbb{R}^2)} < \infty$. Then, by (5.2.4), we can conclude that $\mathrm{WF}(g - P_j g) \supset \{x\} \times V$ holds for every $j \in \mathbb{N}$. Moreover, by construction, we have $d_H(\overline{\mathrm{WF}(g)_x}, \varnothing) = 1$. To define the desired function q, we first set

$$q_0 := P_1 g,$$

$$q_n := \begin{cases} q_{n-1} + (P_n g - P_{n-1} g) & \text{if } \mathrm{DWF}_{j-1}(P_{n-1} q_{n-1})_x = \varnothing, \\ q_{n-1} & \text{otherwise,} \end{cases} \tag{5.2.5}$$

for all $n \geq 1$. By the Riemann-Lebesgue Lemma [99], we conclude that for every $N \in \mathbb{N}$ we conclude that

$$\sum_{n \leq N} \|q_n - q_{n-1}\|_{C^k} \lesssim \sum_{n \leq N} \left\| (1+|\cdot|^2)^{\frac{k}{2}} \left(\widehat{q}_n - \widehat{q}_{n-1}\right) \right\|_{L^1} \leq \left\| (1+|\cdot|^2)^{\frac{k}{2}} \widehat{g} \right\|_{L^1} < \infty. \tag{5.2.6}$$

Moreover, by definition,

$$(P_n g - P_{n-1} g) \perp (P_m g - P_{m-1} g) \quad \text{for all } n \neq m.$$

Hence, by the Pythagorean theorem, for every $N \in \mathbb{N}$,

$$\sum_{n \leq N} \|P_n g - P_{n-1} g\|_2^2 \leq \|g\|_2^2 < \infty. \tag{5.2.7}$$

It now follows from (5.2.7), (5.2.6) and (5.2.5) that q_n is a Cauchy sequence in $C^k(\mathbb{R}^2)$ and $L^2(\mathbb{R}^2)$. Therefore q_n converges to a limit $q \in L^2(\mathbb{R}^2) \cap C^k(\mathbb{R}^2) \subset \mathcal{C}$. Furthermore, one of the following statements holds:

(1) $\overline{\mathrm{DWF}_j(P_jq_j)_x}$ does not converge for $j \to \infty$;

(2) $\overline{\mathrm{DWF}_j(P_jq_j)_x}$ converges to a limit W such that $d_H(W, \overline{\mathrm{WF}(q)_x}) \geq 1/2$;

(3) $\overline{\mathrm{DWF}_j(P_jq_j)_x}$ converges to a limit W such that $d_H(W, \overline{\mathrm{WF}(q)_x}) < 1/2$.

In cases (1) and (2), we directly obtain that DWF_j is ignorant to q at x. In case (3), we obtain that there exists some j_0 such that

$$d_H\big(\overline{\mathrm{DWF}_j(P_jq_j)_x}, \overline{\mathrm{WF}(q)_x}\big) < 1 \quad \text{for all } j \geq j_0. \tag{5.2.8}$$

We now consider the cases $\mathrm{WF}(q)_x = \varnothing$ and $\mathrm{WF}(q)_x \neq \varnothing$ separately. If $\mathrm{WF}(q)_x = \varnothing$, then (5.2.8) implies that $\mathrm{DWF}_j(P_jq_j)_x = \varnothing$ for all $j \geq j_0$, since no subset of $P(\mathbb{S}^1) \setminus \{\varnothing\}$ has a distance less than 1 to the empty set. Therefore,

$$q - P_{j_0}q = \sum_{j>j_0} (P_jg - P_{j-1}g) = g - P_{j_0}g. \tag{5.2.9}$$

We obtain from (5.2.9) that $\varnothing = \mathrm{WF}(q)_x = \mathrm{WF}(g)_x \neq \varnothing$ which is a contradiction. If $\mathrm{WF}(q)_x \neq \varnothing$, then $d_H(\overline{\mathrm{WF}(q)_x}, \varnothing) = 1$. By the triangle inequality, this yields that there exists some j_0 such that $\mathrm{DWF}_j(P_jq_j)_x \neq \varnothing$ for all $j \geq j_0$. Therefore, $q = q_{j_0} \in \mathcal{PW}_{j_0}$ by definition, which implies that $W(q)_x = \varnothing$. Hence, Case (3) does not occur, i.e., $(\mathrm{DWF}_j)_{j\in\mathbb{N}}$ is ignorant to q at x.

Step 2: For an arbitrary $f \in \mathcal{C}$, there exists some $j_1 \in \mathbb{N}$ such that

$$\|f - P_{j_1}f\|_2 \leq \frac{\epsilon}{2} \quad \text{and} \quad \|g - P_{j_1}g\|_2 \leq \frac{\epsilon}{2}.$$

Define $q_{j_1} = P_{j_1}f$ and, for every $n \geq j_1$, define q_n as in (5.2.5). It is clear that q_n converges to a limit $q_f \in C^k(\mathbb{R}^2) \cap L^2(\mathbb{R}^2)$. Also, it is straightforward to show that $\|q_f - f\|_2 \leq \epsilon$. Now using the same arguments as in Step 1, it follows that $(\mathrm{DWF}_j)_{j\in\mathbb{N}}$ is ignorant to q_f. □

Remark 5.2.5. *(1) Theorem 5.2.4 and its proof also hold when "wavefront set" is replaced by "$(k+1)$-wavefront set" or "singular support".*

(2) The arguments in the proof of Theorem 5.2.4 are independent from the domain $\mathbb{R}^2$. Indeed, the same result holds for functions defined on an open domain $\Omega \subset \mathbb{R}^2$ and $x \in \Omega$. Here we define the wavefront set of $f \in L^2(\Omega)$ as

$$\Big\{(x,\xi) \in \Omega \times \mathbb{S}^1 : (x,\xi) \in \mathrm{WF}\left(\tilde{f}\right) \text{ where } \tilde{f} = f \text{ on } \Omega \text{ and } \tilde{f} = 0 \text{ elsewhere}\Big\}.$$

(3) Theorem 5.2.4 demonstrates that there does not exist any clairvoyant approximate wavefront set extractor for sufficiently large function classes. Even more severely, for every $k \in \mathbb{N}$, every approximate wavefront set extractor fails on a dense subset of $L^2(\mathbb{R}^2) \cap C^k(\mathbb{R}^2)$. Also, if the function class is so small that every function is uniquely determined by its values on the grid, then one can construct a wavefront

> *set extractor by storing the wavefront set for each function in a database. This shows that for most classical function classes it is impossible to analytically derive a wavefront set extractor. For function classes in applications, such as sets of images, wavefront set extractors could exist but are potentially highly sensitive to the choice of $\mathcal{C}$.*

In order to formalize the cases where one is able to perform digital wavefront set extraction, we are going to present an alternative form of digital wavefront set extraction as the task of semantic edge detection (Section 5.3). This allows us formalize the task as a non-randomized decision rule in the context of statistical decision theorem. The idea of introducing the task in the picture is driven by our goal to use the wavefront set as some type of regularizer in tomographic reconstruction, this is going to be explored later in Chapter 7.

5.3 Semantic edge detection

The characterization of the wavefront set of a characteristic function $f = \chi_\Omega$ presented in Proposition 2.2.14 implies that detecting the wavefront set of a piece-wise smooth function with singularities along a smooth curve is equivalent to detecting edges and their normal directions. Edge detection, which is one of the most well-studied problems in image processing, is therefore a sub-problem of wavefront set extraction, which one could call *singular support extraction*. In the case where we want to detect orientations in a discrete set, for example, a set of different disjoint intervals of angles, the task of wavefront set extraction is equivalent to detect edges in images and classify them. In this setting, the classes corresponds to the orientations of the particular edge points. This is known in computer vision as *semantic edge detection*.

In short, semantic edge detection is the task of detecting edges and object boundaries in natural images and classifying the points in those edges from a finite set of classes. These classes could represent, for instance, the objects the edges belong to [121] or the orientation of the edge at that particular point [5]. The recent interest from the research community in semantic edge detection is mainly driven by its far-reaching applications in imaging-related tasks such as object recognition, semantic segmentation, and image reconstruction. Semantic edge detection combines two different classification tasks. The first is classical category-agnostic edge detection, which can be viewed as a pixel-wise binary classification for determining whether a pixel belong to an edge or not. The second is the recognition of the classes of pixels in an image that belongs to edges. One can perform semantic edge detection using a model-based or a data-driven approach; each comes with its strengths and shortcomings. Due to our guiding principle, we are going to introduce a method that combines both approaches.

In this subsection, we review some of the existing semantic edge detection methods, both model-based and data-driven, as well as a fundamental limitation of this task, namely, the distracted supervision paradox.

5.3.1 Model-based semantic edge detection

Many existing approaches to identify singularities in images are model-based. These methods usually involve two steps: a filtering step to enhance edge-like features and a classification step to identify pixels belonging to edges. The aforementioned features are extracted using simple rules and heuristics, e.g., convolution with local difference filters correspond to operating on the image with Roberts [95], Sobel [107], and Prewitt [88] operators. In a similar manner, the well-known Canny edge detector [24] corresponds to convolving the image with a Gaussian kernel to further identify those pixels where the gradient is high.

There have been attempts in the past to also model the semantic information of detected edges as in [54]. This work was in fact among the first publications to propose a principled way to combine generic object detectors with bottom-up contours for semantic edge detection. Determining the orientation of an edge is particularly important in inverse problems, since this information is essential in relating edges in data to those in the signal [70]. Wavefront set extraction refers to semantic edge detection, where the classification of the edges is based on their orientation. The continuous theory of wavefront set resolution via multiscale directional systems (e.g., shearlets [72]) allows one to design model-based approaches for wavefront set extraction. These are essentially a digital implementation of the continuous theory, which filters the images before performing the corresponding classification.

An example of such an approach is the shearlet-based algorithm in [120], which uses digital shearlets to filter an image in order to highlight the features corresponding to different orientations and scales. One then performs a simple clustering classification algorithm to classify the corresponding directions. A more recent approach is [93, 94], where a general directional system, known as symmetric molecules, is used to filter the directional features of images to then classify them to be edge, ridge, or blob.

On the one hand, these model-based approaches for semantic edge detection rely on 'first principles' from approximation theory and are easy to interpret, hence it can also be easier to improve upon. On the other hand, the use of rigid heuristics regarding the characterization of singularities makes it difficult to utilize these methods in real-world applications, where the data represents empirically defined function classes.

5.3.2 Data-driven semantic edge detection

More recently, as part of the success stories of machine learning and its successes in addressing various tasks in modern computer vision, a set of deep neural network architectures for semantic edge detection [121, 11, 79, 122] have appeared. One needs to stress that these have set a new state-of-the-art of semantic edge detection.

In broad terms, these methods use similar principles as the model-based ones, i.e., learning filters using convolutional layers and subsequently classifying the corresponding edge pixels by sigmoid or softmax classifiers. Since each convolutional layer represents a level of abstraction of the features in the target images, the initial layers represent 'simple' edges, whereas deeper layers represent more 'complex' features, corresponding to high-level semantics. In that sense, the two steps involved in semantic edge detection, which are semantic-agnostic edge detection followed by edge classification, are conceptually far from each other. Therefore, there is no straightforward way of jointly learning how to extract and classify the edges. This limitation in semantic edge detection is known as *the distracted supervision paradox* [79].

5.3.2.1 A typical deep model for general SED

To introduce previous attempts for using deep supervision in SED, we take the typical deep model as an example, the CASENet [121]. The CASENet architecture is based on the already known Residual Neural Network architecture, also referred to as ResNet (see Figure 5.1) [56]. This architecture has shown tremendous success in different image processing tasks, including image classification. It in fact won the ImageNet challenge in 2015. CASENet receives as input the image and it produces as output a two-dimensional array of the same size with the classified edges represented as pixels with the value given by the corresponding category.

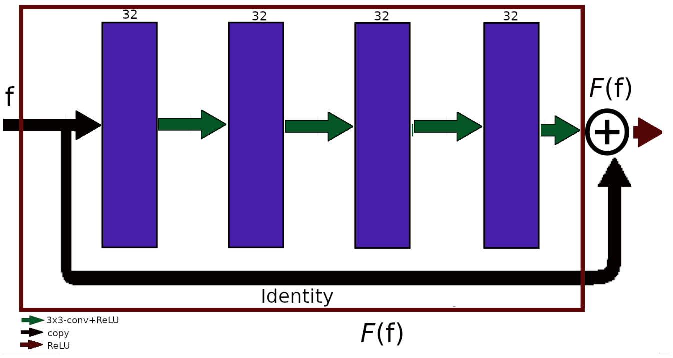

Figure 5.1: Illustration of the principal block in ResNet, namely the skip connection from the input to the output is the main characteristic of this architecture.

We next explain the CASENet architecture, which is displayed in Figure 5.2, in more detail. The input image is connected to a $1-\textit{channel}$ convolutional layer (conv1), which is followed by four stacked ResNet subnetworks; res2c, res3b, res4b22, and res5c correspondingly. Each of those sub-networks is a block of the network ResNet-101, where res(N)(M) represents the M-th layer (represented by the letter "a", "b" and "c") of the N-th stage of ResNet-101 [56].

The first three stages of CASENet (i.e. conv1, res2c, res3b) produce a single channel feature map $F^{(m)}$, which is used to perform the edge detection part. The last stage, res5c, is connected to a 1×1 convolutional layer to produce a K-channel class activation map $A^{(5)} = \{A_1^{(5)}, A_2^{(5)}, \dots, A_K^{(5)}\}$, where K is the total number of categories. In order to combine the edge information coming from the first stages, at the end of the network, one replicates the bottom features $F^{(m)}$, by concatenating them in each channel of the class activation map at the last stage, namely:

$$A^f = \{F^{(1)}, F^{(2)}, F^{(3)}, A_1^{(5)}, \ldots, F^{(1)}, F^{(2)}, F^{(3)}, A_K^{(5)}\}.$$

At the end, a K-grouped 1×1 convolutional layer is applied to A^f, generating a semantic edge map with K channels, where the k-th channel represents the edge map for the k-th category. Summarizing, the first four stages of CASENet produce category-agnostic edge feature maps with different levels of refinement. This depth becomes necessary to produce edges fine enough to be classified by the last stage.

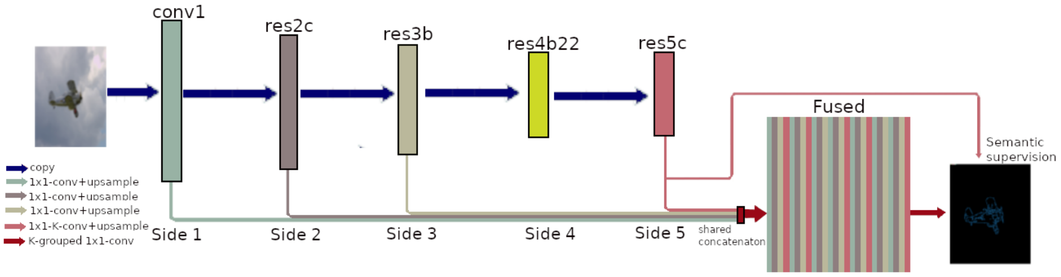

Figure 5.2: Illustration of the CASENet architecture.

After the introduction of CASENet, there have been other methods to perform general SED with deep supervision, e.g. [64, 122, 79], but all with similar network design. These methods have been typically performed on datasets with semantic classes corresponding to the object that the edges belong to. Specialized methods have been also introduced by the authors in [6, 8], where the semantic classes correspond to the orientations of the boundaries at each edge. We will further discuss this in Section 5.4 but before presenting our method, we will shortly discuss a particular challenge when performing semantic edge detection, namely *the distracted supervision paradox*.

5.3.2.2 The distracted supervision paradox

In the early deep supervised SED models ([64, 121, 122]), the authors only imposed supervision on Side 5 (see Figure 5.2) and the final fused activation. In [121] the authors have tried several deeply supervised architectures. They first used all of Side 1 to Side 5 for SED separately, with each side connected with a classification loss. In this situation, the evaluation results were found to be even worse than the basic architecture that directly applies 1×1 convolutions at Side 5 to obtain semantic edges.

It is widely accepted that the lower levels of neural networks contain low-level, less-semantic features such as local edges, which are unsuitable for semantic classification because semantic category recognition needs abstracted high-level features that appear in the top layers of neural networks. Thus, they would obtain poor classification results at the bottom sides. Unsurprisingly, simply connecting each low-level feature layer and high-level feature layer with a classification loss and deep supervision for SED task results in a clear performance drop. In their original work, Yu et al. [121] also attempted to impose deep supervision of binary edges at Side 1 Side 3 in CASENet but observed divergence in the semantic classification at Side 5. With the top supervision of semantic edges, the top layers of the network are supervised to learn abstracted high-level semantics that can summarize different appearances of the target categories.

Since the bottom layers are the bases of the top layers for the representation power of the deep convolutional neural networks, the bottom layers are supervised in order to help the top layers to obtain high-level semantics through backpropagation. Also, with bottom supervision of category-agnostic edges, the bottom layers are taught to focus on the distinction between edges and non-edges, rather than visual representations for semantic classification, which require higher relational abstractions. These two fundamentally distinct tasks cause conflicts in the bottom layers and therefore fail to provide discriminative gradient signals for weight updating.

Note that Side 4 is not used in CASENet. We believe it is a naive way to alleviate the information conflicts by regarding the whole res4 block as a buffer unit between the bottom and top sides, in order to store useful latent information. Indeed, when adding Side 4 to CASENet, the model achieves a 70.9% mean F-measure compared with the 71.4% of the original CASENet (see [79, Section 5.2]). Moreover, the classical 1×1 convolutional layer after each side is too weak to buffer the conflicts. There have been recent approaches to use an architecture similar to CASENet but at the same time avoid the distracted supervision paradox, see [79]. We will discuss this approach in the next subsection. The main drawback of these approaches lies in the complexity of the related deep neural network architectures, which represents an elaborate way to avoid the distracted supervision paradox. Furthermore, the large number of network parameters makes those methods slow and difficult to train.

Later, the authors introduced in [6] a method that uses the powerful singularity representation given by the shearlet transform, already discussed in Chapter 3. This method uses the main backbone of the CASENet, without the buffer block, in order to achieved higher accuracy in SED with fewer parameters. In the next subsection, we will briefly introduce these alternative approaches. The latter method was directly inspired by our proposed digital wavefront set extraction DeNSE, introduced in Section 5.4.3, which performs a specialized SED task while avoiding the distracted supervision paradox.

5.3.2.3 Diverse deep supervision

The fundamental challenge introduced by the distracted supervision paradox has forced researchers who are interested in semantic edge detection to find smart ways to avoid it. The developers of the CASENet architecture later introduced the *Simultaneous Edge Alignment and Learning* (SEAL, [122]), which is a new training approach for the CASENet architecture. It simultaneously aligns the ground truth edges and learns the corresponding classifier, with the downside of being time-consuming due to the necessary CPU usage by the alignment step.

Recently Liu et al. introduced a novel way to train CASENet [79], also known as the deep diverse supervision. This approach makes use of an information converter based on a convolutional residual block (see Figure 5.1), where the output of each stage of

CASENet is fused in a final shared concatenation. Figure 5.3 depicts this architecture, it is worth noticing that in this case, stage four is not anymore a buffer, but it is already used in the supervision.

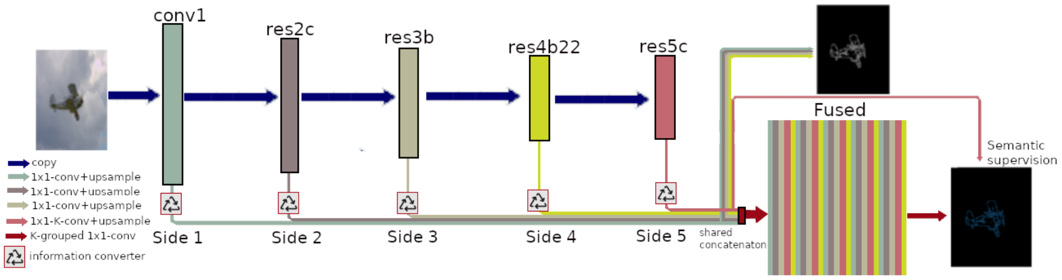

Figure 5.3: Illustration of the classical Diverse Deep Supervision architecture.

The information converters help to assist low-level feature learning (Side 1-4) in order to generate consistent gradient signals from the higher levels (Side 5). This produces a highly discriminative feature map for high-performance semantic edge detection. Having the category-agnostic edge maps obtained from the information converter applied to each of the first four stages, namely $E = E^{(4)} \circ E^{(3)} \circ E^{(2)} \circ E^{(1)}$. The final map is then given by the information conversion of the fifth stage and the shared concatenation, i.e.,

$$E^f = A_K^{(5)} \circ E \circ A_{K-1}^{(5)} \circ E \circ \ldots \circ A_1^{(5)} \circ E.$$

This network is trained with a multi-task loss, meaning, two different losses, corresponding to category-agnostic and category-aware edge detection. These two losses are then optimized jointly. Both losses are based on reweighted sigmoid cross-entropy loss, which is typically used for multi-label classification. For further details, we refer to [79].

5.3.2.4 Shearlet feature extraction for SED

In [6] the authors presented an alternative method inspired in CASENet and DDS, using as input the discrete shearlet coefficients of the target image. In addition, the authors also deleted the buffer block (Side 4) and adapted the channels of the layers to the corresponding shearlet channels. The main motivation behind CASENet and DDS comes from the fact that the classifiers based on convolutional neural networks use convolutional kernels as feature extractors. Those feature extractors transform the input image into a suitable representation system for the particular classification task. We know, as discussed in Chapter 3 that the shearlet transform can represent optimally the oriented singularities of a two-dimensional image. This suggests that it is a suitable representation system for semantic edge detection.

Based on this fact, we propose an alternative to the CASENet and DDS architectures, which take as input the shearlet coefficients. These proposed architectures are able to reduce the number of necessary parameters to achieve better accuracy than the original architectures, mainly due to the heavy lifting performed by the shearlet transform. In

Figures 5.4 and 5.5 we depict the alternative, shearlet-based architectures, shear-CASENet and shear-DDS. Furthermore, in Section 7.4 we present the numerical results obtained with these architectures, with benchmarks for comparison with other relevant methods.

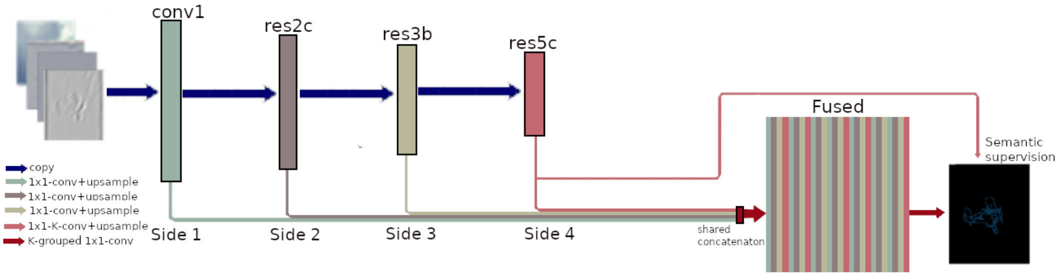

Figure 5.4: Illustration of the Shear-CASENet architecture.

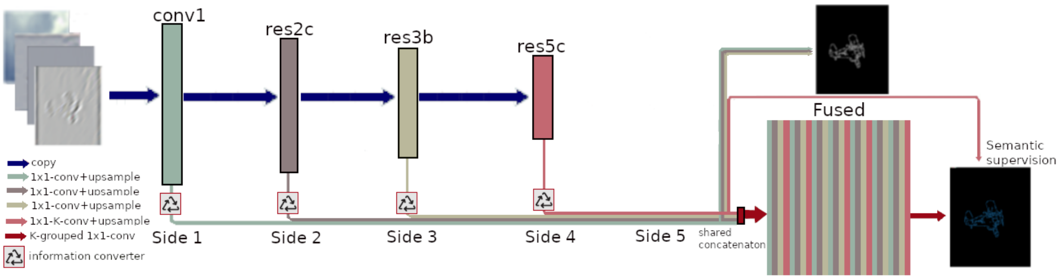

Figure 5.5: Illustration of the shearlet Diverse Deep Supervision architecture.

As we have mentioned before, wavefront set extraction is a particular case of semantic edge detection. Thus, one can use similar ideas to motivate the design of a particular algorithm that performs this task. We cover this approach in Section 5.4 where we introduce the digital wavefront set extraction method performed by deep neural networks on the shearlet domain. This method was introduced by the authors in [8] coined Deep Network Shearlet Edge Extractor, or DeNSE.

5.4 Computing the digital wavefront set with shearlets and deep learning

We will now propose an algorithm that replaces the heuristic approach of the shearlet-based edge detection and classification algorithm of [120] by a data-driven approach. In other words, instead of hand-crafted heuristics, we train a deep neural network using a variety of training data, adapted to the particular classification procedure. Although this might also involve some heuristics, the data-driven approach assumes less conditions on the solutions, making it more general. The neural network takes as input the shearlet coefficients of an image and produces a set of point-direction pairs that are classified as elements of the wavefront set. We will describe the construction of the classifier below and then present the computational realization of our algorithm in Section 5.4.3 at the end of this section. This algorithm was first introduced in [8].

5.4.1 Digital shearlet transform

The classifier that we will construct below is based on the shearlet transform of a digital image. Therefore, we need to work with a digitized shearlet transform, defined on a digital domain of pixel images. The digital shearlet transform was introduced in [75] and is defined as follows:

Definition 5.4.1. *Let $M \in \mathbb{N}$ be the number of pixels, $J \subset \mathbb{N} \setminus \{\infty\}$ be the set of scales, $k_j \subset \mathbb{N}$ be the shearing parameter for all $j \in J$ and $K_j := [-k_j, \dots, k_j]$. We pick $2\sum_{j\in J} K_j + 1$ matrices in $\mathbb{R}^{M\times M}$. We denote these matrices by ϕ^{dig} and $\psi^{dig}_{j,k,\iota}$ for $j \in J, k \in K_j, \iota \in \{-1,1\}$. To make the connection to the classical shearlet transform, we can think of $\psi^{dig}_{j,k,\iota}$ as a digitized version of $\psi_{2^{-j},2^{-j/2}k,0,\iota}$ and of ϕ^{dig} as a digitized version of a low frequency filter. A concrete construction of the matrices ϕ^{dig} and $\psi^{dig}_{j,k,\iota}$ can be found in [75]. Then, we define the* digital shearlet transform *of an image $I \in \mathbb{R}^{M\times M}$ by*

$$\mathrm{DSH}(I)(j,k,m,\iota) := \begin{cases} \langle I, T_m \psi^{dig}_{j,k,\iota} \rangle & \text{if } \iota \in \{-1,1\}, \\ \langle I, T_m \phi^{dig} \rangle & \text{if } \iota = 0, \end{cases}$$

where $j \in J, k \in K_j$, $m \in \{1,\dots,M\}^2$, and $T_m : \mathbb{R}^{M\times M} \to \mathbb{R}^{M\times M}$ circularly shifts the entries of the elements of a matrix by $m \in \mathbb{N}\times\mathbb{N}$, i.e.

$$(T_m I)[i,j] = I[(i+m_1)\%M, (j+m_2)\%M],$$

where $(i+m_1)\%M$ is $i+m_1$ modulo M.

Thus the digital shearlet transform of an image $I \in \mathbb{R}^{M\times M}$ is a stack of $2\sum_{j\in J}(K_j - 1) + 1$ matrices of dimension $M \times M$. In all our numerical experiments presented in Chapter 8, we fixed the number of scales $J = 4$ and the shearing parameter $k_j = 2^{\lceil j/2+1\rceil} + 1$ so $2\sum_{j\in J}(k_j - 1) + 1 = 49$. The computation of the digital shearlet transform is performed by using the Julia implementation of ShearLab [74] (`www.shearlab.org/software`). In the next section we will introduce the digital wavefront set extractor on the shearlet coefficients, defined as a deep convolutional neural network.

5.4.2 Network architecture for wavefront set extraction

In order to compute the digital wavefront set, we have introduced in [8] a deep convolutional neural network. The input of the network are the digital shearlet coefficients of a digital image and the network outputs an estimation to the wavefront set of the image. The network architecture consists of four convolutional layers, with 2 × 2 max pooling, ReLU activation, and batch normalization, followed by a fully-connected layer with

1024 neurons, softmax activation function, and a one-dimensional output. The network architecture is depicted in Figure 5.6. We chose this architecture, since it performed well in a series of tests while being of moderate size. Here we focused on networks with only a few layers because we expect that the shearlet transform already acts as the correct feature extractor of the problem. Therefore, the classifier does not need to learn the correct data representation. Nonetheless, it is conceivable that a deeper and larger neural network architecture could potentially lead to improvements for the classification results below, on account of efficiency.

We train the neural network on spatial patches of the shearlet coefficients of a function. This network produces a prediction of which directions belong to the wavefront set of the function at the position associated with this patch. These patches are of size $21 \times 21 \times 49$ voxels. We pick 180 values $(\theta_i)_{i=1}^{180} \in \{0, 1\}$ that the network will output, representing each direction in the wavefront set. For each θ_i, we then train a network Φ_i with the described architecture by passing to the network patches of shearlet coefficients of images $I \in \mathbb{R}^{M\times M}$ of the form

$$(\mathrm{DSH}(I)(j,k,m,\iota))_{j\in J,k\in K_j,\iota\in\{-1,0,1\},m\in[m_1^*-10,m_1^*+10]\times[m_2^*-10,m_2^*+10]}, \tag{5.4.1}$$

where $m^* \in \{11, \dots, M - 10\}^2$, to the network. The associated label to a patch of (5.4.1) is 1, if and only if the image I has a singularity centered on that patch with direction $i \in \{0, \dots, 180\}$ at m^*, and 0 else. In total, this procedure yields 180 digital classifiers. We train one more network with the same data, but the label is 1 if I has no singularity at m^* and 0 else. This additional classifier is used in test (unseen) cases, where all competing algorithms only perform edge detection and not edge-orientation detection.

The final classifier is constructed by putting all these 181 networks in parallel, producing one large network with 181 outputs. For every $21 \times 21 \times 49$ patch of shearlet coefficients, this classifier generates a vector of length 181, indicating if the underlying function is smooth at the center point of the patch and listing all directions of edges present at the center point.

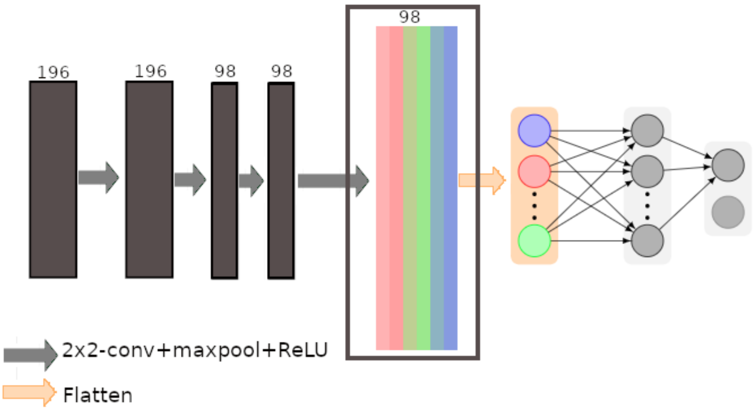

Figure 5.6: Illustration of the network architecture forming the foundation of the classifier. This network consists of four convolutional layers and one fully-connected layer. The colored block in the middle represents a stack of the output of the last convolutional layer. The colors correspond to the different channels.

5.4.3 DeNSE: Deep Network Shearlet Edge Extractor

Now we present our algorithm for extracting the wavefront set of a digital image. For $M \in \mathbb{N}$, and a digital image $I \in \mathbb{R}^{M\times M}$, this algorithm produces, for every $m^* \in [11, M-10]^2$ a prediction of the wavefront set of I at m^*. The algorithm proceeds along with the following three steps.

Algorithm 7: DeNSE algorithm, [8]

```
1: Train the network classifier on a set of labeled training data.
2: For a given test image $I \in \mathbb{R}^{M\times M}$, compute the digital shearlet transform of $I$
   with 49 shearlet generators: the digital shearlet transform of $I$ is given by
   $(\mathrm{DSH}(I)(j,k,m,\iota))_{j\in J,k\in K_j,\iota\in\{-1,0,1\},m\in[1,M]^2}$.
3: For every $m^* = (m_1^*, m_2^*) \in [11, M-10]^2$, pass the patch
```

$$\big(\mathrm{DSH}(I)(j,k,m,\iota)\big)_{j\in J,k\in K_j,\iota\in\{-1,0,1\},m\in[m_1^*-10,m_1^*+10]\times[m_2^*-10,m_2^*+10]} \tag{5.4.2}$$

The network in Step 1 of Algorithm 7 is then trained to classify a set of labeled training data. We refer to the above method as Deep Network Shearlet Edge Extractor (DeNSE, [8]). In Section 7.3 we present the results obtained on different training sets, including ideal phantoms with analytical wavefront set and natural images with approximate wavefront set. Before we conclude this chapter dedicated to digital wavefront set extraction, we will discuss shortly the *distracted supervision paradox* in the case of the DenSE architecture.

After this discussion we will proceed to the analysis of digital wavefront set propagation in convolutional neural networks,

5.4.3.1 Distracted Supervision Paradox for DeNSEE

As we have discussed in Section 5.3.2, semantic edge detection represents a challenge for deep supervision in form of the distracted supervision paradox. Deep supervised digital wavefront set extraction, being a particular case of SED, also suffers from this challenge, but hopefully our method DeNSE is able to overcome such limitation. Indeed, the key to the Deep Network Shearlet Edge Extractor (DeNSE) lies in the splitting of the multi-label classification task into individual binary classifiers inspired by the performance increment. In addition, DeNSE separates the category-agnostic edge detection and the semantic edge classification, which already avoids the distracted supervision paradox. This is easily shown, since this separation avoids the joint supervised training of the edge detector and the edge classifiers, by training them individually.

In the design process of DeNSE, we have also observed that the joint supervision of all the classes (multi-label classification) did not present a satisfactory performance, and based on that we split the classifiers, at the beginning without noticing that we were avoiding the distracted supervision paradox. The latter is one of the reasons for the high accuracy achieved by our method; we present these results in Section 8.1. Furthermore, in Chapter 6 we will use the notion of digital wavefront set and the theory presented in Chapter 4 to characterize the digital wavefront set propagation across certain class of convolutional neural networks, in particular the ones that have residual structure. These results allow us to characterize the digital wavefront set of the output of the learned primal-dual reconstruction architecture by knowing the wavefront set of the input, and then use it as a-priori information in the context of task-adapted reconstruction. In order to formalize the notion of digital wavefront set and digital wavefront set extractor, we will use the notion of semantic edge detection in the context of statistical estimation theory, a detailed discussion is presented in Section 5.5.

5.5 Wavefront set extraction as statistical estimation

In this section we interpret the wavefront set extraction as a statistical estimation problem. This is a way to formally define a digital wavefront set extraction even in the realm of Theorem 5.2.4, i.e., when there is no clairvoyant wavefront set extraction. Define $\square := (0,1)^2$ and the operator $w : L^2(\square) \to 2^{\square \times \mathbb{S}^1}$ given by

$$w(f) := \mathrm{WF}(f), \tag{5.5.1}$$

where $2^{\square \times \mathbb{S}^1}$ denotes the power set of $\square \times \mathbb{S}^1$ and $\mathrm{WF}(f)$ is the wavefront set of f in the continuous sense. The operator w maps a function to its wavefront set. In the following we introduce some relevant definitions from measure theory.

Definition 5.5.1 (σ-algebra)**.** *Let X be a set. Then a σ-algebra $\mathcal{F}$ is a nonempty collection of subsets of X such that the following hold:*

1. *X is in $\mathcal{F}$.*
2. *If B is in $\mathcal{F}$, then A^c is also in $\mathcal{F}$.*
3. *If $\{B_n\}_{n=1}^{\infty}$ is a sequence of elements of $\mathcal{F}$, then $\bigcup_{n=1}^{\infty} A_n$ is also in $\mathcal{F}$.*

We refer to the tuple $(X, \mathcal{F})$ as sample space*.*

Definition 5.5.2 (Probability measure)**.** *Let $(X, \mathcal{F})$ be a sample space, i.e., $\mathcal{F}$ is a $\sigma-$algebra of X. A* probability measure *is a real valued function on $\mathcal{F}$, namely $\mathbb{P} : \mathcal{F} \to \mathbb{R}$ that satisfies the following three conditions.*

1. *$\mathbb{P}(B) \geq 0$ for every element of $B \in \mathcal{F}$.*
2. *$\mathbb{P}(X) = 1$.*
3. *If $\{B_n\}_{n=1}^{\infty}$ is a sequence of pairwise disjoint elements of $\mathcal{F}$*

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} B_n\right) = \sum_{n=1}^{\infty} \mathbb{P}(B_n).$$

We refer to the tuple $(X, \mathcal{F}, \mathbb{P})$ as measurable space*, or* probability space*.*

Definition 5.5.3. *A set $\mathcal{I} \subset L^2(\square)$ is called an* image class*. For a σ-algebra $\mathcal{F}$ over $\mathcal{I}$ and an associated probability measure $\mathbb{P}$, we call the measurable space $(\mathcal{I}, \mathcal{F}, \mathbb{P})$ an* image model*.*

In the following, we introduce the notion of digitization operator for wavefront sets, which will be used later to digitize the wavefront set extractor.

Definition 5.5.4. *Let $n \in \mathbb{N}$ be a resolution and $\mathbf{I} = Q^n$ be a discrete grid, where $Q \subset \mathbb{R}$ is a discrete set. For an image model $(\mathcal{I}, \mathcal{F}, \mathbb{P})$, we call the map*

$$D : \mathcal{I} \to \mathbf{I} \times \{0,1\}^{181n}$$

for $n \in \mathbb{N}$ a digitization operator for wavefront sets*.*

In Definition 5.5.4 $\mathbf{I} = Q^n$ is a discrete grid on the spatial domain of the image. Therefore, D will map a continuous image in $\mathcal{I}$ to the discrete grid where each pixel has values on $\{0,1\}^{181n}$ representing the wavefront set orientation at the location of the pixel.

According to Definition 5.5.3, a digitization operator induces a distribution $\mathbb{P}_{\mathcal{J}}$ on $\mathbf{I} \times \{0,1\}^{181n}$. Since $\mathbf{I} \times \{0,1\}^{181n}$ is a discrete space, the distribution $\mathbb{P}_{\mathcal{J}}$ is a discrete probability distribution. In addition, such map will be also measurable.

Definition 5.5.5. *Let $(\mathcal{I}, \mathcal{F}, \mathbb{P})$ be an image model. Let $n \in \mathbb{N}$ and $\mathbf{I} = Q^n$ be as in Definition 5.5.4. Further, let D be an associated digitization operator and $\mathbb{P}_{\mathcal{J}}$ the corresponding discrete probability distribution. We define*

$$dwf : \mathbf{I} \to \{0,1\}^{181n} \tag{5.5.2}$$

$$\mathbf{f} \mapsto \operatorname{argmax}_{\mathbf{x} \in \{0,1\}^{181n}} \mathbb{P}_{\mathcal{J}}(\mathbf{x}|\mathbf{f}). \tag{5.5.3}$$

Intuitively, dwf outputs an estimation to a set of directions of the wavefront set for each location of the discrete image.

In statistical terms, dwf is the maximum likelihood estimator for the estimation problem associated to the statistical model $\big(\mathbf{I}, 2^{\mathbf{I}}, (\mathbb{P}(\cdot|\mathbf{f}))_{\mathbf{f}\in(\{0,1\}^{181})^n}\big)$, where $2^{\mathbf{I}}$ is the power set of $\mathbf{I}$. A natural question is if it would make sense to replace (5.5.3) by

$$\mathbf{f} \mapsto \sum_{\mathbf{x}\in\{0,1\}^{181n}} \mathbf{x}\mathbb{P}_{\mathcal{J}}(\mathbf{x}|\mathbf{f}),$$

i.e., the conditional expected value of $\mathbf{x}$ given $\mathbf{f}$.

To fully formalize the task of digital wavefront set extraction, we now need to specify the digital wavefront set extractor D. The most natural choice is when D operates as follows. Let $(\tau_i)_{i=1}^n$ be a partition of $\square$ then we define

$$D(f) = (\mathbf{f}, q) \in \mathbf{I} \times \{0,1\}^{181n}, \tag{5.5.4}$$

where $\mathbf{f}_i$ is the closest point in Q to $\frac{1}{|\tau_i|}\int_{\tau_i} f(x)dx$ and for $r = 0, \ldots, n-1$ and $t = 1, \ldots, 180$ we set $q_{181r+t} = 1$ if the angular part of $wf(f)$ is at some point in τ_{r+1} between $t - 1/2$ and $t + 1/2$ degrees. Otherwise, $q_{181r+t} = 0$. Also $q_{181r} = 0$ if f is smooth in τ_r and 1 if not.

From this section we can conclude that although digital wavefront set extraction cannot be defined in closed-form, one can define it as the solution of an statistical estimation problem. The big question here is whether this concept of digital wavefront set will also have the classical properties of the continuous version, namely, the microcanonical relation. In Chapter 6 we will answer this question by introducing the notion of digital microcanonical relation which describes the propagation of digital wavefront set under digital Fourier integral operators.

6 Digital wavefront set propagation in convolutional neural networks

The main goal of this chapter is to transfer the microlocal behavior of Fourier integral operators and convolutional neural networks in the continuous setting presented in Chapters 2 and 4 to the digital case. In other words, we aim to find the digital microcanonical relation that characterizes the propagation of digital wavefront sets under the application of digital Fourier integral operators and convolutional neural networks.

The first step towards the analysis of the propagation of the digital wavefront set under the application of a particular Fourier integral operator (FIO) is the faithful optimal discretization of such operator. The idea of discretizing a Fourier integral operator has been explored widely in the last years [51, 52, 4, 20, 31], using different approaches, from finite element methods to directional multiscale systems.

In the following we are going to focus our attention on the directional multiscale system approach. In particular, in Section 6.1 we are going to present a faithful discretization of Fourier integral operators using discrete shearlets, in which case error bounds can be established. In this section we will also define the dicrete microcanonical relation as a mapping between shearlet coefficients. The discretization error bounds will be later used in Section 6.2 to digitize Fourier integral operators as well as their action on digital wavefront sets in the form of a digital microcanonical relation. The bound is then written in terms of shearlet parameters. In Section 6.3 we will use this digitization procedure to digitize the microcanonical relation of the conv-ResNet blocks and the Radon transform discussed in Chapter 4. This digitization allows us to propagate the digital wavefront set of a sinogram through the learned primal-dual architecture. In Chapter 7 we will make use of this tool to define a tomographic reconstruction method in the framework of *task-adapted reconstruction* [1].

My contribution: This chapter was developed fully by myself, extending the notion of digital microlocal analysis originally introduced by Maarten de Hoop in [116] for the case of curvelets, to shearlets. The actual writing was done fully by myself, with the corresponding references of the work by Maarten de Hoop. The content of this chapter has not been published in form of a paper yet.

6.1 Discretization of Fourier integral operators via shearlets

Similar to our argument in Section 5.1.1, in this chapter we will work with the discretization of $L^2(\Omega)$ functions instead of tempered distributions. In particular, Remark 5.1.1 allows us to choose $X = L^2(\Omega)$ and $Y = L^2(\Xi)$ as the *image* and *data space*, respectively, where $\Omega, \Xi \subset \mathbb{R}^2$ are open. Next, recall from Chapter 2 that a Fourier integral operator (FIO) $\mathcal{P} : X \longrightarrow Y$, where $X = L^2(\Omega)$ and $Y = L^2(\Xi)$, with $\Omega, \Xi \subset \mathbb{R}^2$ open domains, is an operator of the form

$$\mathcal{P}f(y) = \int_{\xi\in\mathbb{R}^q}\int_{x\in\mathbb{R}^q} e^{i\phi(y,x,\xi)}p(y,x,\xi)f(x)dxd\xi, \quad \text{for every } f \in X \text{ and } y \in \Xi,$$

where ϕ is a phase function (see Definition 2.4.3) and p is the amplitude function (see Definition 2.3.8), also known as *the symbol.* For this chapter we would like to write the Fourier integral operator in the next form also known as *generalized Radon transform* [116]

$$\mathcal{P}f(y) = \int_{\xi\in\mathbb{R}^q} e^{i\phi(y,\xi)}p(y,\xi)\hat{f}(\xi)d\xi, \quad \text{for every } f \in X \text{ and } y \in \Xi. \tag{6.1.1}$$

Assume furthermore that ϕ and p satisfy the following properties.

(i) $\phi \in C^\infty(\Xi \times \mathbb{R}^2 \setminus \{0\})$ is real-valued and positive homogeneous in ξ, i.e., $\phi(y, k\xi) = k\phi(y,\xi)$, for all $k > 0$.

(ii) $p \in C^\infty(\Xi \times \mathbb{R}^2 \setminus \{0\})$ is a standard amplitude of order m as in Definition 2.3.8, that is

$$|\partial_\xi^\alpha \partial_x^\beta p(y,\xi)| \leq C_{\alpha\beta}(1 + |\xi|)^{m-|\alpha|} \tag{6.1.2}$$

for multi-indices $\alpha = (\alpha_1, \alpha_2)$ and $\beta = (\beta_1, \beta_2)$. In addition, we assume that $y \mapsto p(y,\xi)$ has compact support in $\Xi \times \mathbb{R}^2 \setminus \{0\}$.

In the continuous setting, one can then characterize the propagation of singularities under the action of a FIO by the microcanonical relation [70] (Definition 2.4.5). More precisely, one can prove that

$$\mathrm{WF}(\mathcal{P}f) \subset C_\phi \circ \mathrm{WF}(f), \tag{6.1.3}$$

where C_φ is given by (2.4.3). Moreover, we can express the microcanonical relation in (6.1.3) in terms of coordinate transformations as follows.

Remark 6.1.1 (Classical microcanonical relation mapping)**.** *Let $\mathcal{P} : L^2(\Omega) \to L^2(\Xi)$ be a Fourier integral operator of the form* (6.1.1)*, where $\phi \in C^\infty(\Xi \times \mathbb{R}^2 \setminus \{0\})$ is the phase function. The microcanonical relation* (6.1.3) *can be represented by the mapping $\chi : \Xi \times \mathbb{R}^2 \setminus \{0\} \to \Omega \times \mathbb{R}^2 \setminus \{0\}$ given by*

$$\chi(-\partial_\xi\phi(y,\xi), \xi) = (y; \partial_y\phi(y,\xi)), \quad \textit{for all } (y;\xi) \in \Xi \times \mathbb{R}^q \setminus \{0\}. \tag{6.1.4}$$

In other words, we have

$$C_\phi \circ \mathrm{WF}(f) = \Big\{(y;\mu) : (y;\mu) = \chi(x;\lambda) \textit{ for some } (x,\lambda) \in \mathrm{WF}(f)\Big\}.$$

We call the mapping χ the classical microcanonical mapping.

Having the definition of the digital wavefront set (Definition 5.5.5), it is natural to ask if the digital wavefront set satisfies some form of microcanonical relation with respect to a particular digitization of a Fourier integral operator. We here essentially follow the approach taken in [4]. The digitization of a Fourier integral operator requires a discretization step. We would like such discretization to be efficient. In order to achieve that, it is natural to seek expansions of the amplitude and the complex exponential in terms products in the space $\Xi \times \mathbb{R}^2 \setminus \{0\}$, also known as *the phase space*. Notice that elements of the phase space are of the form $(y, \xi) \in \Xi \times \mathbb{R}^2 \setminus \{0\}$, where $y \in \Xi$ represent the spatial coordinate and $\xi \in \Xi$ the frequency coordinate. In the next section, we focus on a specific discretization method based on multiscale systems.

6.1.1 The discrete shearlet system

In the last decades, researchers have tried to use classical discretization methods to discretize FIOs, e.g., differences and finite element methods, however, discretizations based on expansion using multiscale systems on the time-frequency domain have shown better estimates. Moreover, traditional methods for time-frequency and multiscale analysis have proven to be very effective in the study of a large class of operators, such as pseudodifferential operators [47]. However, traditional time-frequency methods do not apply directly to the study of Fourier integral operators. For instance, in general, a Fourier integral operator $\mathcal{P}$ does not have a sparse matrix representation with respect to the frame of wavelets [20].

In 2005, E. Candès and his collaborators showed that the curvelet frames (see Definition 3.2.7) are able to represent Fourier integral operators sparsely [20]. This happens mainly, as discussed in Section 3.2.2, since the frequency tilling generated by the curvelet system efficiently covers the Fourier domain, resulting in an optimal representation with high-frequency behavior (see Theorem 3.2.5). Furthermore, such optimal tilling allows curvelets to be used to compute the wavefront set of a function [22]. This type of frequency tilling, based on a dyadic parabolic decomposition of the phase space, with curvelets supported on a wedge-like tile, is also present in other multiscale directional systems coming from applied harmonic analysis, such as shearlets (Definition 3.3.1).

In this chapter, we focus on the approach presented by Guo and Labate [51, 52], where the shearlet system is used to sparsely represent FIOs. The sparse representation is later used to establish an error bound, similar to the one already done in the curvelet case by Hoop et al. (see [116]). This can be then applied to digitize such discretization with a known precision. In this section we introduce the discrete shearlet transform and present some relevant results.

Remark 6.1.2. *Based on the continuum shearlet system given by* (3.3.3)*, we can define discrete sheralets by discretizing the parameter spaces, namely*

$$\mathcal{SH}_\psi := \{\psi_{j,k,m}(\cdot) = |\det A_j|^{1/2}\psi(S_k A_j \cdot -m) : j, k \in \mathbb{Z}, m \in \mathbb{Z}^2\}, \tag{6.1.5}$$

where the scaling matrix $A_j \in \mathbb{R}^{2\times 2}$ *and the shearing matrix* $S_k \in \mathbb{R}^{2\times 2}$ *are given by*

$$A_j := \begin{pmatrix} 2^j & 0 \\ 0 & 2^{j/2} \end{pmatrix}, \quad S_k := \begin{pmatrix} 1 & k \\ 0 & 1 \end{pmatrix}, \quad \textit{for all } j,k \in \mathbb{Z}. \tag{6.1.6}$$

In addition, $\psi \in L^2(\mathbb{R}^2)$ *in an special case can defined by its Fourier transform as*

$$\hat{\psi}(\xi) = \hat{\psi}(\xi_1, \xi_2) = \hat{\psi}_1(\xi_1)\hat{\psi}_2\left(\frac{\xi_2}{\xi_1}\right) \tag{6.1.7}$$

where $\hat{\psi}_1, \hat{\psi}_2 \in C^\infty(\hat{\mathbb{R}})$*,* $\operatorname{supp}\hat{\psi}_1 \subset [-1/2, -1/16] \cup [1/16, 1/2]$ *and* $\operatorname{supp}\hat{\psi}_2 \subset [-1,1]$.

Similar to the continuum case, under some assumptions we can prove that the discrete shearlet system $\mathcal{SH}_\psi$ defined in (6.1.5) is a tight frame. We present this result in the following proposition.

Proposition 6.1.3. *Let* $\psi \in L^2(\mathbb{R}^2)$ *be a shearlet function given by* (6.1.7)*, where* $\hat{\psi}_1, \hat{\psi}_2 \in C^\infty(\hat{\mathbb{R}})$ *follow the estimates*

$$\sum_{j\geq 0} |\hat{\psi}_1(2^{-j}\omega)|^2 = 1 \quad \textit{for} \quad |\omega| \geq 1/8, \tag{6.1.8}$$

and, for any $j \geq 0$

$$\sum_{k=-2^j}^{2^j} |\hat{\psi}_2(2^j\omega + k)|^2 = 1 \quad \textit{for} \quad |\omega| \leq 1. \tag{6.1.9}$$

Then, the shearlet system

$$\mathcal{SH}_\psi := \{\psi_{j,k,m}(\cdot) = |\det A_j|^{1/2}\psi(S_k A_j \cdot - m) : j,k \in \mathbb{Z}, m \in \mathbb{Z}^2\}, \tag{6.1.10}$$

is a tight frame of $L^2(\mathcal{C}_h)^\vee = \{f \in L^2(\mathbb{R}^2) : \operatorname{supp}\hat{f} \subset \mathcal{C}_1\}$.

Proof. Notice that $(\xi_1,\xi_2)A_j^{-1}S_k^{-1} = (2^{-j}\xi_1, -k2^{-j/2}\xi_1 + 2^{-j/2}\xi_2)$. Thus, in the frequency domain, the elements of the shearlet system $\psi_{j,k,m}$ have the form

$$\begin{aligned}\hat{\psi}_{j,k,m}(\xi) &= |\det A|^{-j/2}\psi(\xi A_j^{-1}S_k^{-1})e^{2\pi i \xi A_j^{-1}S_k^{-1}m} \\ &= 3^{-3j/4}\hat{\psi}_1(2^{-j}\xi_1)|\hat{\psi}_2(2^{j/2}\xi_2/\xi_1 - k)|e^{2\pi i\xi A_j^{-1}S_k^{-1}m},\end{aligned}$$

and, as a result, the elements $\hat{\psi}_{j,k,m}$ are supported in the sets

$$W_{j,k} = \{(\xi_1,\xi_2) : \xi_1 \in [-2^{j-1/2}, -2^{j-2}] \cup [2^{j-2}, 2^{j-1/2}], |\xi_2/\xi_1 - k2^{-j/2}| \leq 2^{-j/2}\}, \tag{6.1.11}$$

where $j \geq 0, |k| \leq 2^{j/2}$. Using Equations (6.1.8) and (6.1.8) we obtain

$$\sum_{j\geq 0}\sum_{k=-2^{j/2}}^{2^{j/2}} |\hat{\psi}(\xi A_j^{-1}S_k^{-1})| = 1 \quad \text{for } (\xi_1,\xi_2) \in \mathcal{C}_h,$$

where $\mathcal{C}_h$ is the horizontal cone given by

$$\mathcal{C}_h := \{(\xi_1, \xi_2) \in \hat{\mathbb{R}}^2 : |\xi_1| \geq 1, |\xi_2/\xi_1| \leq 1\}.$$

This result, together with the fact that $\operatorname{supp} \hat{\psi} \subset [-1/2, 1/2]^2$, implies that the collection

$$\Psi(\psi) := \{\psi_{j,k,m}(y) = 2^{3j/4}\psi(S_k A_j y - k) : j \geq 0, -2^{j/2} \leq k \leq 2^{j/2}, m \in \mathbb{Z}^2\}$$

is a tight frame for $L^2(\mathcal{C}_h)^\vee = \{f \in L^2(\mathbb{R}^2) : \operatorname{supp} \hat{f} \subset \mathcal{C}_1\}$. □

In order to obtain a Parseval frame for $L^2(\mathbb{R}^2)$, one can easily construct a similar system of shearlets $\{\tilde{\psi}_{j,k,m}\}_{j,k,m}$ for the vertical cone

$$\mathcal{C}_v := \{(\xi_1, \xi_2) \in \hat{\mathbb{R}}^2 : |\xi_1| \geq 1, |\xi_1/\xi_2| \leq 1\}.$$

Finally, one can construct a Parseval frame for $L^2([-\frac{1}{2}, \frac{1}{2}]^2)$. Therefore any function $f \in L^2(\mathbb{R}^2)$ can be written as a sum of three components, i.e.,

$$f = P_{\mathcal{C}_h} f + \mathcal{P}_{\mathcal{C}_v} f + \mathcal{P}_{\mathcal{R}} f, \tag{6.1.12}$$

corresponding to the orthogonal projection of f into the three subspaces of $L^2(\mathbb{R}^2)$. The resulted tiling of the frequency plane $\hat{\mathbb{R}}^2$ is depicted in Figure 6.1 (left). As one can see in Figure 6.1, for each j, k, the set $W_{j,k}$ is a symmetric pair of trapezoids centered around $\pm\xi_{j,k}$, where $\xi_{k,m} = 2^{-j}(2^j, k2^{j/2})$. These trapezoids are normally referred to as *wedges*. Each wedge is contained in a box of size $2^j \times 2^{j/2}$ in the frequency domain as shown in Figure 6.1 (right), oriented along the line $\xi_2 = k2^{-j/2}\xi_1$. Thus, the frequency support of the shearlets satisfies a *parabolic scaling*, and becomes increasingly elongated as the scale j increases. This fact is crucial to optimally discretize Fourier integral operators. In order to simplify the notation, from now on we will denote by $\{\psi_\mu : \mu \in \mathcal{M}\}$ and $\{\tilde{\psi}_\mu : \mu \in \mathcal{M}\}$, the horizontal and vertical shearlet systems, where

$$\mathcal{M} := \{\mu = (j, k, m) : j \geq 0, 2^{-j/2} \leq k \leq 2^{j/2}, k \in \mathbb{Z}^2\}. \tag{6.1.13}$$

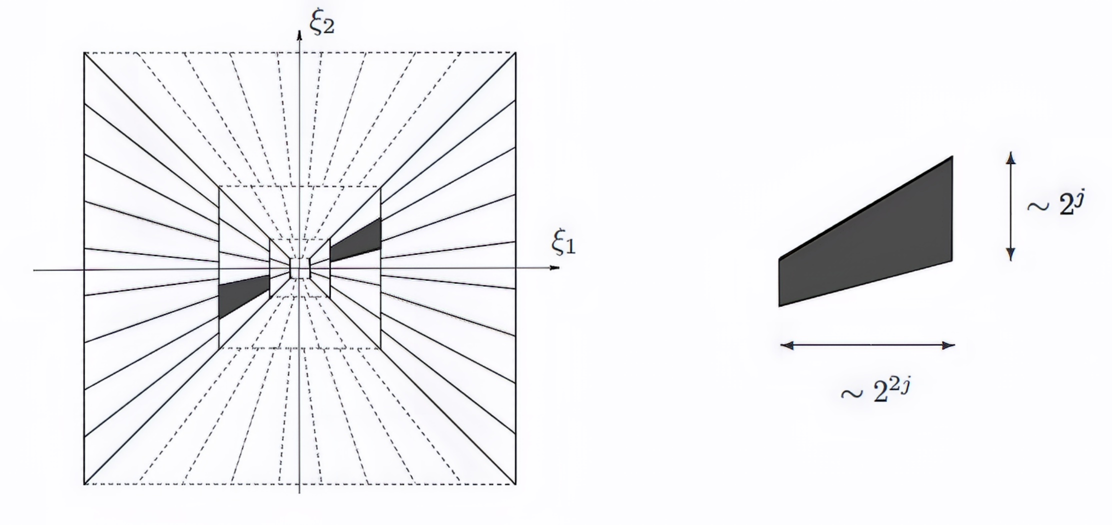

Figure 6.1: Right: Fourier tiling of the shearlets. Left: Frequency support of the shearlet $\psi_{j,k,m}$, for $\xi_1 > 0$.

6.1.2 Sparse shearlet representation of Fourier integral operators

If one has an orthonormal system $\{\phi_\nu\}_{\nu\in\mathcal{N}} \subset L^2(\mathbb{R}^2)$, it is easy to expand a function in terms of the system as a weighted sum by taking the coefficients as the inner products of the function and the elements of the system, i.e.,

$$f = \sum_{\nu\in\mathcal{N}} \langle f, \phi_\nu\rangle \phi_\nu.$$

However, shearlets do not form an orthonormal family. Following the approach of Guo and Labate [51], one can define an appropriate inner product in the parameter space $\mathcal{N}$ such that distinct shearlet parameters are almost orthonormal with respect to the inner product, i.e., the inner product of two distinct shearlets decays almost exponentially asymptotically in the scale. The distance defined by the inner product is also known as *dyadic parabolic pseudo-distance* and will later allow us to bound the elements of the shearlet representation matrix of any Fourier integral operator. We will also prove that this bound decays exponentially outside the diagonal of the representation matrix making the representation *optimal.* Let us now define the dyadic parabolic pseudo-distance as follows.

Definition 6.1.4. *Let $\mu, \mu' \in \mathcal{M}$ two shearlet indices, where $\mu = (j,k,m)$ and $\mu' = (j',k',m')$. We define the* dyadic parabolic pseudo-distance *between two indices μ and μ as the function $\omega : \mathcal{M}^2 \to \mathbb{R}$ given by:*

$$\omega(\mu,\mu') = 2^{|j-j'|/2}\left(1 + 2^{\min(j,j')} d(\mu,\mu')\right),$$

where

$$d(\mu,\mu') = |k2^{-j/2} - k'2^{-j'/2}|^2 + |m_{j,k} - m'_{j',k'}|^2 + |\langle e_\mu, m_{j,k} - m'_{j',k'}\rangle|^2,$$

for $e_\mu = (\cos\theta_\mu, \sin\theta_\mu)$ *and* $\theta_\mu = \arctan(k2^{-j/2})$, *and* $m_{j,k}$ *(*$m'_{j',k'}$ *correspondingly) the spatial centers of the wedges* $W_{j,k}$ *from Equation* (6.1.11)*.*

From Definition 6.1.4 we have that ω increases as the difference between the scales, the shearings, and the positions increases. This definition is an adaptation of the original pseudo-distance in Candès and Demanet [20] to the shearlet realm, provided by Guo and Labate [52]. Before we present the main theorem of this section, we need to further introduce the notion of shearlet representation matrix of a Fourier integral operator.

Definition 6.1.5. *Let* $\mathcal{P} : X \to Y$ *be a Fourier integral operator of the form* (6.1.1)*, where* $X = L^2(\Omega)$*, and* $Y = L^2(\Xi)$ *with* $\Omega, \Xi \subset \mathbb{R}^2$ *open. Moreover, let*

$$\mathcal{SH}_\psi = \{\psi_\mu : \mu \in \mathcal{M}\} \cup \{\tilde{\psi}_\mu : \mu \in \mathcal{M}\}$$

be a discrete shearlet parseval frame from Proposition 6.1.3. The shearlet representation matrix *is given by* $\{\mathcal{P}(\mu,\mu')\}_{\mu,\mu'\in\mathcal{M}}$*, whose elements* $\mathcal{P}(\mu,\mu') \in \mathbb{R}$ *are defined as*

$$\mathcal{P}(\mu,\mu') = \langle \mathcal{P}\psi_\mu, \psi_{\mu'}\rangle \quad \textit{for all } \mu,\mu' \in \mathcal{M}. \tag{6.1.14}$$

The main theorem in [51] states the sparse representation of Fourier integral operators given by shearlets. The precise result reads as follows.

Theorem 6.1.6 ([51, Theorem 3.1])**.** *Let* $\mathcal{P} : X \to Y$ *be a Fourier integral operator of the form* (6.1.1) *of order* $m = 0$ *(see* (6.1.2)*), where* $X = L^2(\Omega)$*, and* $Y = L^2(\Xi)$ *with* $\Omega, \Xi \subset \mathbb{R}^2$ *open. In addition, suppose that the phase of* $\mathcal{P}$*,* $\phi \in C^\infty(\Xi \times \mathbb{R}^2 \setminus \{0\})$ *satisfies the non-degeneracy condition, i.e., there is a constant* $c > 0$ *such that*

$$|\det \partial_y\partial_\xi\phi(y,\xi)| \geq c \quad \textit{for all } (y,\xi) \in \Xi \times \mathbb{R}^2 \setminus \{0\} \textit{ uniformly.}$$

Furthermore, let $P(\mu,\mu') \in \mathbb{R}$ *be an element of the shearlet representation matrix of* $\mathcal{P}$ *(see Definition 6.1.5). Then, for each* $N > 0$*, there is a constant* $C_N > 0$ *such that*

$$|\mathcal{P}(\mu,\mu')| \leq C_N\omega(\mu, h_{\mu'}(\mu'))^{-N} \quad \textit{for all } \mu,\mu' \in \mathcal{M},$$

where $h_\mu : \mathcal{M} \to \mathcal{M}$ *is the index mapping induced by the microcanonical relation of* $\mathcal{P}$ *(the map* h_μ *is constructed within the proof, see* (6.1.30)*).*

In Theorem 6.1.6 the Fourier integral, operator $\mathcal{P}$ is represented in the shearlets frame by the matrix with elements $\mathcal{P}(\mu,\mu')$ (see Definition 6.1.5). This theorem states that the elements can be bounded by an estimate that decays exponentially outside the diagonal $[\mathcal{P}(\mu,\mu)]_{\mu\in\mathcal{M}}$. This result establishes an optimal representation estimate for the shearlet representation of Fourier integral operators. The fact that shearlet optimally represent Fourier integral operators allows us to establish a digitization procedure with fast decaying errors (see Section 6.2). For the sake of completeness, we present the proof of Theorem 6.1.6 from [51]. As the first step towards the proof, we need to introduce a convenient representation of the Fourier integral operator $\mathcal{P}$ with respect to the Parseval frame of shearlets, this construction is based on the curvelet case [20].

Definition 6.1.7. *The* shearing and scaling matrices $S_{j,k} \in \mathbb{R}^{2\times 2}$ *is defined by*

$$S_{j,k} = A_j^{-1} S_k^{-1} A_j = \begin{pmatrix} 1 & -k2^{-j/2} \\ 0 & 1 \end{pmatrix} \quad \textit{for all } j \geq 0 \textit{ and } |k| \leq 2^j.$$

In addition, let $W_j \subset \mathbb{R}^2 \setminus \{0\}$ be the set in the Fourier domain given by

$$\begin{aligned} W_j &= W_{j,0} = W_{j,k} S_{j,k} \\ &= \{(\xi_1, \xi_2) : \xi_1 \in [-2^{j-1/2}, -2^{j-2}] \cup [2^{j-2}, 2^{j-1/2}], |\xi_2/\xi_1| \leq 2^{-j/2}\}. \end{aligned}$$

Remark 6.1.8. *Observe from Definition 6.1.7 that $S_{j,k}$ maps $W_{j,k}$ into another pair of symmetric wedges oriented along the ξ_1-axis. For $\mu \in \mathcal{M}$, let ψ_μ be a shearlet centered at ξ_μ in the Fourier space. Then, using the change of variables $\xi = \eta S_{j,k}^{-1}$, we have*

$$\begin{aligned} \mathcal{P}\psi_\mu(y) &= \int_{\mathbb{R}^2\setminus\{0\}} e^{i\phi(y,\xi)} p(y,\xi) \hat{\psi}_\mu(\xi) d\xi \\ &= 2^{-3j/4} \int_{W_j} e^{i(\phi(y,\eta S_{j,k}^{-1}) - \eta A_j^{-1} m)} p(y, \eta S_{j,k}^{-1}) \hat{\psi}(\eta A_j^{-1}) d\eta. \end{aligned} \tag{6.1.15}$$

Following the form (6.1.15), it is convenient to locally linearize the phase $\phi(y,\xi)$ to separate the nonlinearities in ξ from those in y. This is a standard procedure also done by Hoop in [4] for the case of prolate spheroidal wave functions and is in general a standard approach in the study of Fourier integral operators (see [108, Chapter 9]) as well as Guo and Labate for the shearlet case [51]. In order to do this, let us introduce the following proposition

Proposition 6.1.9. *Let $\mathcal{P} : X \to Y$ be a Fourier integral operator of the form (6.1.1) with amplitude $p \in C^\infty(\Xi \times \mathbb{R}^2 \setminus \{0\})$ and phase $\phi \in C^\infty(\Xi \times \mathbb{R}^2 \setminus \{0\})$. In addition, let W_j^+ and W_j^- be the negative and positive parts of $W_j \subset \mathbb{R}^2 \setminus \{0\}$ given by*

$$W_j^+ := W_j \cap \{(\xi_1, \xi_2) : \xi \geq 0\}, \quad W_j^- := W_j \cap \{(\xi_1, \xi_2) : \xi < 0\}.$$

In addition, for $j \geq 0$ and $|k| \leq 2^j$ let $\delta_{j,k} : \mathbb{R}^n \times W_j \to \mathbb{R}$ be the function given by as

$$\delta_{j,k}(y,\eta) := \begin{cases} \phi(y, \eta S_{j,k}^{-1}) - \eta S_{j,k}^{-1} \cdot \partial_\xi \phi(y, (1,0) S_{j,k}^{-1}), & \textit{for } \eta \in W_j^+, \\ \phi(y, \eta S_{j,k}^{-1}) - \eta S_{j,k}^{-1} \cdot \partial_\xi \phi(y, (-1,0) S_{j,k}^{-1}), & \textit{for } \eta \in W_j^-. \end{cases}$$

For fixed index $\mu \in \mathcal{M}$, we can decompose $\mathcal{P}$ with the operators $\mathcal{P}_\mu^{(1)} : X \to Y$ and $\mathcal{P}^{(2)} : Y \to Y$ as

$$\mathcal{P} = \mathcal{P}_\mu^{(2)} \mathcal{P}_\mu^{(1)},$$

where

$$\mathcal{P}_\mu^{(1)} f(y) = \int_{W_j} e^{i\eta S_{j,k}^{-1} y} \beta_\mu(y,\eta) \hat{f}(\eta S_{j,k}^{-1}) d\eta \quad \textit{for all } y \in \Xi, \tag{6.1.16}$$

with

$$\beta_\mu(y,\eta) = e^{i\delta_{j,k}(\varphi_\mu^{-1}(y),\eta)} p(\varphi_\mu^{-1}(y), \eta S_{j,k}^{-1}) \quad \textit{for all } (y,\eta) \in \Xi \times W_j,$$

and

$$\mathcal{P}_\mu^{(2)} f(y) = f(\varphi_\mu(y)) \quad \textit{for all } y \in \Xi, \tag{6.1.17}$$

for

$$\varphi_\mu(y) = \partial_\xi \phi(y, (1,0)S_{j,k}^{-1}) \quad \textit{for all } y \in \Xi.$$

Proof. Notice that by definition of $\mathcal{P}_\mu^{(2)}$ and $\mathcal{P}^{(1)}$ we have

$$\begin{aligned}\mathcal{P}_\mu^{(2)}\mathcal{P}_\mu^{(1)} f(y) &= \mathcal{P}^{(2)}\left(\int_{W_j} e^{i\eta S_{j,k}^{-1} y}\beta_\mu(y,\eta)\hat{f}(\eta S_{j,k}^{-1})d\eta\right)\\ &= \int_{W_j} e^{i\eta S_{j,k}^{-1}\varphi_\mu(y)}\beta_\mu(\varphi_\mu(y),\eta)\hat{f}(\eta S_{j,k}^{-1})d\eta.\end{aligned}$$

By plugging the explicit form of β_μ we obtain

$$\mathcal{P}_\mu^{(2)}\mathcal{P}_\mu^{(1)} f(y) = \int_{W_j} e^{i\alpha_\mu(y,\eta)} p(y, \eta S_{j,k}^{-1})\hat{f}(\eta S_{j,k}^{-1})d\eta,$$

where the exponent $\alpha_\mu \in C^\infty(\mathbb{R}^n \times W_j)$ is given by

$$\alpha_\mu(y,\eta) = \eta S_{j,k}^{-1}\varphi_\mu(y) + \delta_{j,k}(y,\eta) \quad \text{for all } (y,\eta) \in \Xi \times \{0\}.$$

Now, assuming that $\eta \in W_j^+$ and plugging the explicit form of φ_μ and $\delta_{j,k}$ we get

$$\begin{aligned}\alpha_\mu(y,\eta) &= \eta S_{j,k}^{-1} \cdot \delta_\xi \phi(y,(1,0)S_{j,k}^{-1}) + \phi(y, \eta S_{j,k}^{-1}) - \eta S_{j,k}^{-1} \cdot \delta_\xi \phi(y,(1,0)S_{j,k}^{-1})\\ &= \phi(y, \eta S_{j,k}^{-1}) \quad \text{for all } (y,\eta) \in \Xi \times \mathbb{R}^2 \setminus \{0\}.\end{aligned}$$

Similarly if $\eta \in W_j^-$ we have $\alpha_\mu(y,\eta) = \phi(y, \eta S_{j,k}^{-1})$, obtaining finally

$$\begin{aligned}\mathcal{P}_\mu^{(2)}\mathcal{P}_\mu^{(1)} f(y) &= \int_{W_j} e^{i\alpha_\mu(y,\eta)} p(y,\eta S_{j,k}^{-1})\hat{f}(\eta S_{j,k}^{-1})d\eta\\ &= \int_{W_j} e^{i\phi(y,\eta S_{j,k}^{-1})} p(y,\eta S_{j,k}^{-1})\hat{f}(\eta S_{j,k}^{-1})d\eta\\ &= \mathcal{P}f(y) \quad \text{for all } (y,\eta) \in \Xi \times \mathbb{R}^2 \setminus \{0\}.\end{aligned}$$

□

Following [51], we are going to analyze the operators $\mathcal{P}_\mu^{(1)}$ and $\mathcal{P}_\mu^{(2)}$ in the next two sections.

6.1.2.1 Analysis of operator $\mathcal{P}_\mu^{(1)}$

Let us first notice that the operator $\mathcal{P}_\mu^{(1)} : X \to Y$ defined in (6.1.16), has a linear phase $\tilde{\phi}(y,\eta) = \eta S_{j,k}^{-1} y$. We should also notice that it is not strictly a pseudo-differential operator, since the amplitude $\beta_\mu \in C^\infty(\Xi \times \mathbb{R}^2 \setminus \{0\})$ is not a standard amplitude function in the sense of Definition 2.3.8. In particular, it is an amplitude function multiplied by a complex exponential.

Let ϕ be the phase function of $\mathcal{P}$, defined in (6.1.1). Let us explore the particular case when $\phi(\xi) = |\xi|$, for $\xi = (\xi_1, \xi_2) \in \mathbb{R}^2 \setminus \{0\}$, and $\xi_\mu = 2^{j/2} e_\mu$ for $e_\mu = (\cos\theta_\mu, \sin\theta_\mu)$ and $\theta_\mu = \arctan(k2^{-j/2})$. This means that

$$\partial_\xi \phi(\xi_\mu) = \frac{\xi_\mu}{|\xi_\mu|} = e_\mu,$$

and

$$\delta_{j,k}(y, S_{j,k}\xi) =: \delta_\mu(y,\xi) = \phi(\xi) - \partial_\xi\phi(\xi_\mu)\xi = \phi(\xi) - e_\mu \xi.$$

For $\theta_\mu = 0$, we have that $e_\mu = (1,0)$ and

$$\delta_\mu(y,\xi) = |\xi| - \xi_1 = \sqrt{\xi_1^2 + \xi_2^2} - \xi_1.$$

This implies that the derivatives of $\delta_\mu(\xi)$ are homogeneous of degree 0 in ξ, hence, they present no decay in ξ. Thus, $\beta_\mu(y,\xi) := \beta_\mu(y, S_{j,k}\xi)$ does not satisfy the condition (6.1.2) unless $\delta_\mu(y,\xi) = 0$. In addition, notice that $\delta_{j,k}(y,\xi)$ is generally unbounded since the phase ϕ is unbounded, with the exception when $\xi \in \mathrm{supp}(\hat{\psi}_\mu)$, due to the parabolic scaling nature of shearlets (see [51]). This is a key point on why shearlets and other similar systems are effective in dealing with the operator $\mathcal{P}_\mu^{(1)}$ [4], in other words, they can be used to sparsely represent it.

In the following, based on the technique presented in [51], we prove that the operator $\mathcal{P}_\mu^{(1)}$ maps a shearlet ψ_μ into a shearlet-like function m_μ, with the same phase space location. Such function is coined a *shearlet molecule*, defined as follows.

Definition 6.1.10 (Shearlet molecules, [51, Definition 3.2])**.** *For $\mu = (j,k,m) \in \mathcal{M}$, let $a_\mu \in C^\infty(\Xi)$ be a smooth function. Then, the function $b_\mu \in C^\infty(\Xi)$ given by*

$$b_\mu(y) = 2^{3j/4} a_\mu(S_k A_j y - m) \quad \textit{for all } y \in \Xi \tag{6.1.18}$$

is a horizontal shearlet molecule *with regularity $R \in \mathbb{R}_+$ if a_μ satisfies the following properties:*

(i) For each $\gamma = (\gamma_1, \gamma_2) \in \mathbb{N} \times \mathbb{N}$ and each $N \geq 0$ there is a constant $C_N > 0$ independent of μ such that

$$|\partial_y^\gamma a_\mu(y)| \leq C_N (1 + |y|)^{-N}. \tag{6.1.19}$$

(ii) For each $M \leq R$ and each $N \geq 0$ there is a constant $C_{N,M} > 0$ independent of μ such that

$$|\hat{a}_\mu(\xi)| \leq C_{N,M}(1+|\xi|)^{-N}(2^{-j}+|\xi_1|)^M. \tag{6.1.20}$$

Vertical shearlet molecules *are similarly defined.*

One can associate the second factor in the inequality (6.1.20) with *almost vanishing moments.* This implies, that the frequency support of a shearlet molecule b_μ is mostly concentrated around $|\xi| \approx 2^j$ (see [51]). Coarse-scale molecules are defined as elements of the form

$$\{a_\mu(y-m) : m \in \mathbb{Z}^2\},$$

where a_μ satisfies (6.1.19). Let us explore some implications of Definition 6.1.10.

Remark 6.1.11. *If $m_\mu(y)$ is a horizontal shearlet molecule with regularity R, then from Equation* (6.1.19) *it follows that*

$$|(i\xi)^\gamma \hat{a}_\mu(\xi)| \leq ||\partial^\gamma a_\mu||_{L^1} \leq C_\gamma.$$

Therefore, for all $N \geq 0$ there is a constant C_N such that

$$|\hat{a}_\mu(\xi)| \leq C_N(1+|\xi|)^{-N}.$$

From these results it follows that, for all $N \geq 0$, there is a constant C_N such that

$$|\hat{b}_\mu(\xi)| \leq C_N 2^{-3j/4}(1+|\xi A_j^{-1} S_k^{-1}|)^{-N}. \tag{6.1.21}$$

In addition, from Equation (6.1.20) it follow that, for each $M \leq R$ and $N \geq 0$, there is a constant $C_{N,M} > 0$ such that

$$|\hat{b}_\mu(\xi)| = |\hat{a}_\mu(\xi A_j^{-1} S_k^{-1})| \leq C_{N,M} 2^{-3j/4} \min\{1, 2^{-j}(1+|\xi_1|)\}^M (1+|\xi A_j^{-1} S_k^{-1}|)^{-N}. \tag{6.1.22}$$

We have therefore the following theorem

Theorem 6.1.12 ([51, Thm. 3.3])**.** *Let $\{\psi_\mu : \mu \in \mathcal{M}\}$ be a Parseval frame of shearlets (see* (6.1.10)*). For each index $\mu \in \mathcal{M}$ the operator $\mathcal{P}_\mu^{(1)}$ maps ψ_μ into a shearlet molecule $b_\mu = \mathcal{P}_\mu^{(1)}\psi_\mu$ with arbitrary regularity R, uniformly in μ. That is, the constant in Definition 6.1.10 is independent of μ. The same result holds also for the vertical shearlets $\tilde{\psi}_\mu$.*

Finally, we also have that the shearlet molecules from Definition 6.1.10 form an almost orthogonal family with respect to the dyadic parabolic pseudo-distance ω.

Proposition 6.1.13 ([51, Proposition 3.4]). *Let b_μ and $b_{\mu'}$ be two shearlet molecules (see* (6.1.18)*) with regularity R. Let $j, j' \geq 0$. For every $N \leq C(R)$, there is a constant $C_N > 0$ such that*

$$|\langle b_\mu, b_{\mu'} \rangle| \leq C_N \omega(\mu, \mu')^{-N} \quad \textit{for } \mu, \mu' \in \mathcal{M}.$$

The number $C(R)$ increases with R and goes to infinity as R goes to infinity. This result extends to the case when both b_μ and b'_μ are vertical shearlets. It also extends to the case when one molecule is vertical and the other horizontal.

For the proofs of Theorem 6.1.12 and Proposition 6.1.13 we refer to [52]. Let us now study the operator $\mathcal{P}_\mu^{(2)}$.

6.1.2.2 Analysis of operator $\mathcal{P}_\mu^{(2)}$

As one can observe, in the analysis the operator $\mathcal{P}_\mu^{(1)}$, we made extensive use of the fact that the shearlets $\{\psi_\mu : \mu \in \mathcal{M}\}$ have compact support in the frequency domain (in other words, they are *band-limited*). As the singularities of a function, which are propagated by the Fourier integral operator, have not just a frequency component (the orientation) but a spatial component (the position), we would like to also have a high spatial resolution. The operator $\mathcal{P}_\mu^{(2)}$ allows us to introduce a family of shearlet-like functions with compact support in the spatial domain, known as *shearlet atoms*.

This family of functions let us introduce an atomic decomposition of the form

$$f(y) = \sum_\mu \nu_\mu \rho_\mu(y) \quad \text{for } f \in L^2(\Xi),$$

where the shearlet atoms $\rho_\mu \in L^2(\Xi)$ have compact support and satisfy certain regularity and vanishing moments conditions. In addition, $\nu_\mu \in \mathbb{R}$ are the coefficients.

Following the construction in [51], we introduce the family of shearlet-like functions with compact support of the form

$$\psi_{ast}(y) = |\det A_a|^{-1/2} \psi(A_a^{-1} S_s^{-1}(y - t)),$$

where

$$A_a = \begin{pmatrix} a & 0 \\ 0 & \sqrt{a} \end{pmatrix}, \quad S_s = \begin{pmatrix} 1 & s \\ 0 & 1 \end{pmatrix},$$

and a, s, t are continuous parameters, such that, $0 < a \leq 1$, $|s| \leq 2$ and $t \in \mathbb{R}$. Now, let us introduce the notion of *vanishing moments*.

Definition 6.1.14 (Vanishing moments). *Let $\varphi \in \mathcal{S}(\Xi)$ be a Schwartz function. Such function has $k-$vanishing moments in the y_1-direction if there exists $\tilde{\varphi} \in \mathcal{S}(\Xi)$ such that*

$$\varphi(y) = \partial_{y_1}^k \tilde{\varphi}(y) \quad \textit{for all } y \in \Xi.$$

Remark 6.1.15. *Notice that if φ has a certain number of vanishing moments in the y_1 direction, then*

$$\hat{\varphi}(\xi) = (i\xi_1)^k \widehat{\tilde{\varphi}}(\xi),$$

and therefore, $\hat{\varphi}(0, \xi_2) = 0$. This implies that $\hat{\varphi}(\xi)$ is concentrated along the ξ_1-axis, meaning that $\hat{\varphi}(\xi S_s A_a)$ is concentrated in elongated regions (increasingly elongated as $a \to 0$). Moreover, this regions are symmetric with respect to the origin, along the direction $\xi_1 = s\xi_1$.

Before presenting the prove of Theorem 6.1.6, we introduce two proposition whose proofs can be found in [51].

Proposition 6.1.16 ([51, Proposition 3.5])**.** *Let ψ be a Schwartz function such that $\hat{\psi}(\pm 1, 0) \neq 0$ and having at least one vanishing moment in the y_1 direction. Therefore, there is a function $q(\xi)$ such that we have*

$$q(\xi) \int_{|s| \leq 2} \int_{a \leq 1} a^{3/2} |\hat{\psi}(\xi S_s A_a)|^2 \frac{da}{a^3} ds = 1, \quad \textit{for } \xi \in \Gamma.$$

In addition, $q(\xi)$ is a smooth function satisfying $|\partial^\alpha q(\xi)| \leq C|\xi|^{-|\alpha|/2}$ on Γ, where

$$\Gamma = \{(\xi_1, \xi_2) \in \mathbb{R}^2 : |\xi_1| \geq 1, |\xi_2/\xi_1| \leq 1\}. \tag{6.1.23}$$

Since there are plenty of ways to choose a Schwartz function ψ which satisfies the assumptions of Proposition 6.1.16, we then choose a separable one, i.e., of the form

$$\psi(y_1, y_2) = \psi_1(y_1)\psi_2(y_2),$$

where $\psi_1, \psi_2 \in C_c^\infty(\mathbb{R})$ with $\operatorname{supp}\psi_1, \operatorname{supp}\psi_2 \subset [0, 1]$. We also assume that ψ_1 has vanishing moments up to order R, that is,

$$\int_{\mathbb{R}} \psi_1(y) y^k dy = 0, \quad k = 0, 1, \ldots, R.$$

This lets us obtain the reproducing formula on the next proposition.

Proposition 6.1.17 ([51, Proposition 3.6])**.** *Let Γ be as Equation (6.1.23) and suppose that $\hat{f}$ vanishes outside the set Γ. Then we have the reproducing formula*

$$f(y) = \int_{\mathbb{R}}^2 \int_{|s| \leq 2} \int_{a \leq 1} \langle q(D)f, \psi_{ast} \rangle \psi_{ast}(y) \frac{da}{a^3} ds dt \quad \textit{for } y \in \Xi, \tag{6.1.24}$$

where $q \in L^2(\mathbb{R}^2)$ is defined by $\widehat{(q(D)f)}(\xi) = q(\xi)\hat{f}(\xi)$, and

$$\psi_{ast}(y) = a^{-3/4}\psi(A_a^{-1} S_s^{-1}(y - t)).$$

The reproducing formula (6.1.24) can be written as an *atomic decomposition* where the integral is broken into several components associated with distinct regions. For $\mu = (j, k, m)$, let

$$\begin{aligned} Q_\mu = \{(a, s, t) :& 2^{-(j+1)} \leq a < 2^{-j}, k2^{-j} \leq s < (k+1)2^{-j}, \\ & A_j^{-1} S_k^{-1} t \in [m_1, m_1 + 1) \times [m_2, m_2 + 1)\}. \end{aligned}$$

Notice that each of the regions Q_μ are disjoint and that

$$\bigcup_{j\geq 0} \bigcup_{k=-2^{(j+1)/2}}^{2^{(j+1)/2}-1} \bigcup_{(m_1,m_2)\in\mathbb{Z}^2} Q_\mu = \{(a, s, t) : a \leq 1, |s| \leq 2, t \in \mathbb{R}^2\}.$$

By splitting the integral (6.1.24) into components corresponding to different cells Q_μ, we finally get

$$f(y) = \sum_{j\geq 0} \sum_{k=-2^{(j+1)/2}}^{2^{(j+1)/2}-1} \sum_{(m_1,m_2)\in\mathbb{Z}^2} \nu_\mu \rho_\mu(y), \tag{6.1.25}$$

where

$$\begin{aligned} \rho_\mu(y) &= \frac{1}{\nu_\mu} \int\int\int_{Q_\mu} \langle q(D)f, \psi_{ast}\rangle \psi_{ast}(y) \frac{da}{a^3} dsdt, \\ \nu_\mu &= \left(\int\int\int_{Q_\mu} |\langle q(D)f, \psi_{ast}\rangle|^2 \frac{da}{a^3} dsdt \right)^{1/2}. \end{aligned} \tag{6.1.26}$$

We can now define the handle functions known as *atoms* and *shearlet atoms* as follows

Definition 6.1.18. *Let $\mu \in \mathcal{M}$ a shearlet index and let $\tilde{\psi}_\mu \in L^2(\Omega)$. Let us now define the functions α_μ given by*

$$\alpha_\mu(y) = 2^{-3j/4}\tilde{\psi}_\mu(A_j^{-1}B_k^{-1}(y+m)) = 2^{-3j/4}\tilde{\psi}_\mu(S_{j,k}A_j^{-1}(y+m)), \mu \in \mathcal{M}, y \in \Xi, j \geq 0, |k| \leq 2^{j/2}.$$

We refer to the elements α_m as atoms *with regularity R. Also, the shearlet-like functions ρ_μ, given by*

$$\rho_\mu(y) = 2^{3j/4}\alpha_\mu(S_k A_j y - m), \quad \mu \in \mathcal{M}, y \in \Xi, j \geq 0, \tag{6.1.27}$$

are referred to as shearlet atoms *with regularity R, we also refer to them as* shearlet-like functions*.*

Remark 6.1.19. *Notice that α_μ have the following properties*

(i) ***Compact support:*** $\operatorname{supp} \alpha_\mu \subset C[-1, 1]^2$*, where C is independent of μ and f.*

(ii) ***Regularity:*** *for each $\beta = (\beta_1, \beta_2)$, there is a constant C_β independent of μ and f such that*

$$|\partial_y^\beta \alpha_\mu(y)| \leq C_\beta.$$

(iii) ***Vanishing moments on the*** y_1 ***direction:*** *for all* $n = 0, 1, \dots, R$,

$$\int_{\mathbb{R}} \alpha_\mu(y_1, y_2) y_1^n dy_1 = 0,$$

with $R \in \mathbb{R}$ *as in Proposition 6.1.16.*

Observe that, by definition, both α_μ and ρ_μ are compactly supported in the spatial domain. We introduce a final preliminary result in the form of a theorem, so we can prove Theorem 6.1.6.

Theorem 6.1.20 ([51, Theorem 3.7])**.** *Let* $\{\rho_{\mu'} : \mu' \in \mathcal{M}\}$ *be a family of shearlet atoms with regularity R. For each* $\mu' \in \mathcal{M}$*, the operator* $\mathcal{P}_\mu^{(2)}$ *maps* $\rho_{\mu'}$ *into a shearlet atoms* $m_{h_\mu(\mu')}$ *with the same regularity R, uniformly over* $\mu' \in M$*.*

Now, if $\phi(y, \xi)$ is the phase associated with the Fourier integral operator $\mathcal{P}$, it induces a change of variables in terms of the microcanonical relation (see Definition 2.4.5), also represented given by the classical microcanonical relation mapping $\chi : \Omega \times \mathbb{R}^2 \setminus \{0\} \to \Xi \times \mathbb{R}^2 \setminus \{0\}$ (see Remark 6.1.1). Let $b_{\mu'}$ and $\xi_{\mu'}$ be the space and frequency locations of $\rho_{\mu'}$, and define

$$\phi_\mu(y_{\mu,\mu'}) := m_{\mu'},$$

where $\phi_\mu(y) = \partial_\xi \phi(y, \xi_\mu)$. Since $\phi(y, \xi)$ is homogeneous of degree one in ξ, then $\partial_y \phi(y, \xi) = \xi \partial_y \partial_\xi \phi(y, \xi)$.

In Theorem 6.1.20, likewise in Theorem 6.1.6, one makes use of the bijective mapping h_μ acting on $\mathcal{M}$ induced by the microcanonical relation in terms of the transformation (6.1.4). We refer to this mapping as the *discrete micro-canonical relation* and is formally intorduced in Definition 6.1.21. Using these observations, we have that

$$\begin{aligned} \partial_\xi \phi(m_{\mu'}, \xi_\mu) &= \phi_\mu(y_{\mu,\mu'}) =: m_{\mu'}, \\ \partial_y \phi(m_{\mu'}, \xi_\mu) &= \xi_\mu \partial_y \partial_\xi \phi(m_{\mu'}, \xi_\mu) =: \eta_{\mu,\mu'}. \end{aligned} \tag{6.1.28}$$

These two relations allow us to describe the action of the operator $\mathcal{P}_\mu^{(2)}$ on the phase space coordinates of the shearlet atoms $\rho_{\mu'}$. This results in a change of shearlet parameters described by the *discrete microcanonical relation.* We define this mapping precisely in the following.

Definition 6.1.21 (Discrete microcanonical relation)**.** *Let* $\phi(y, \xi)$ *be the phase function of the Fourier integral operator* $\mathcal{P}$ *with an associated classical microcanonical relation* χ *(see Remark 6.1.1). Let* $m_{\mu'}$ *and* $\xi_{\mu'}$ *be the space and frequency locations of* $\rho_{\mu'}$*, defined by Equation* (6.1.26)*. Moreover, by Equation* (6.1.25) *the action of the operator* $\mathcal{P}^{(2)}$ *on the phase space coordinates of the shearlet atom* $\rho_{\mu'}$ *is given by the mapping* $\chi_\mu : \Omega \times \mathbb{R}^2 \setminus \{0\} \to \Xi \times \mathbb{R}^2 \setminus \{0\}$ *defined as*

$$\chi_\mu(m_{\mu'}, \xi_{\mu'}) = (y_{\mu,\mu'}; \eta_{\mu,\mu'}), \quad \mu, \mu' \in \mathcal{M}, \tag{6.1.29}$$

where $\eta_{\mu,\mu'}$ *and* $y_{\mu,\mu'}$ *are given by Equation* (6.1.28). *This induces the index mapping* $h_\mu : \mathcal{M} \to \mathcal{M}$, *referred to as the* discrete microcanonical relation, *given by*

$$h_\mu(\mu') = h_\mu(j_{\mu'}, k_{\mu'}, m_{\mu'}) := (j_{\mu,\mu'}, k_{\mu,\mu'}, m_{\mu,\mu'}) \tag{6.1.30}$$

induced by the microcanonical relation (6.1.29), *where* $j_{\mu,\mu'}, k_{\mu,\mu'}, m_{\mu,\mu'})$ *are the shearlet indices corresponding to the shearlet centered at* $(y_{\mu,\mu'}; \nu_{\mu,\mu'})$ *in the phase space.*

Before the proof of Theorem 6.1.6, we need to introduce one last proposition with some useful properties of the pseudo-distance ω.

Proposition 6.1.22 ([51, Proposition 3.8])**.** *Let* $\mu, \mu', \mu'', \mu_0 \in \mathcal{M}$ *be shearlet parameters. The dyadic parabolic pseudo-distance* ω *(Definition 6.1.4) satisfies the following properties:*

(i) ***Symmetry:*** $\omega(\mu, \mu') \sim \omega(\mu', \mu)$.

(ii) ***Triangle inequality:*** *there is a constant* $C > 0$ *such that* $d(\mu, \mu') \leq Cd(\mu, \mu') + d(\mu'', \mu')$.

(iii) ***Composition:*** *for every* $N > 0$, *there is a constant* $C_N > 0$ *such that*

$$\sum_{\mu''} \omega(\mu, \mu'')^{-N} \omega(\mu'', \mu')^{-N} \leq C_N \omega(\mu, \mu')^{-N+1}.$$

(iv) ***Invariance under the bijective index mappping*** h_{μ_0} ***induced by the microcanonical relation:*** $\omega(\mu, \mu') \sim \omega(h_{\mu_0}(\mu), h_{\mu_0}(\mu'))$, *uniformly over* $\mu_0 \in \mathcal{M}$, *where* h_μ *is given by* (6.1.30).

Now we are ready to prove Theorem 6.1.6 taken from [51], this proof is an adaptation of the proof in [51] to the shearlets realm.

Proof. Let ψ_{μ_0} and ψ_{μ_1} be two fixed shearlets, and for simplicity, let us assume that both are horizontal. Using Theorem 6.1.12, we have that $b_{\mu_0} = \mathcal{P}^{(1)}_{\mu_0}\psi_{\mu_0}$ is a shearlet molecule. In addition, by using the atomic decomposition (6.1.29), we can expand the shearlet ψ_{μ_1} as the linear combination of the shearlet atoms $\rho_{\mu'}$:

$$\psi_{\mu_1} = \sum_{\mu'} c_{\mu',\mu_1} \rho_{\mu'},$$

where

$$c_{\mu',\mu_1} = \left(\int\int\int_{Q_{\mu'}} |\langle q(D)\psi_{\mu_1}, \psi_{ast}\rangle|^2 \frac{da}{a^3} dsdt \right)^{1/2}. \tag{6.1.31}$$

Thus, using these observations and Equation (6.1.21), we have that

$$\begin{aligned} \langle \psi_{\mu_1}, \mathcal{P}\psi_{\mu_0}\rangle = \langle \psi_{\mu_1}, \mathcal{P}^{(2)}_{\mu_0}\mathcal{P}^{(1)}_{\mu_0}\psi_{\mu_0}\rangle &= \langle (\mathcal{P}^{(2)}_{\mu_0})^* \psi_{\mu_1}, \mathcal{P}^{(1)}_{\mu_0}\psi_{\mu_0}\rangle \\ &= \sum_{\mu'} c_{\mu',\mu_1} \langle (\mathcal{P}^{(2)}_{\mu_0})^* \rho_{\mu'}, b_{\mu_0}\rangle \\ &= \sum_{\mu'} c_{\mu',\mu_1} \langle b_{\tilde{h}_{\mu_0}(\mu')}, b_{\mu_0}\rangle, \end{aligned}$$

where $b_{\tilde{h}_{\mu_0}(\mu')}$ is a shearlet molecule and $\tilde{h}_{\mu_0} = h_{\mu_0}^{-1}$ is the inverse of the discrete canonical relation mapping h_{μ_0}.

Now, let us observe that for every $N > 0$, there is a constant C_N such that

$$|c_{\mu',\mu_1}| \leq C_N \omega(\mu', \mu_1)^{-N}.$$

This can be shown by discretizing the integral (6.1.31) and using Proposition 6.1.13. By using Propositions 6.1.13 and 6.1.22, we finally get that for every $N > 0$, there is a constant C_N such that:

$$\begin{aligned} |\langle \psi_{\mu_1}, \mathcal{P}\psi_{\mu_0}\rangle| &\leq \sum_{\mu'} |c_{\mu',\mu_1}| |\langle b_{\tilde{h}_{\mu_0}(\mu')}, b_{\mu_0}\rangle| \\ &\leq C_N \sum_{\mu'} \omega(\mu', \mu_1)^{-N} \omega(\tilde{h}_{\mu_0}(\mu'), \mu_0)^{-N} \\ &\leq C_N \sum_{\mu'} \omega(\tilde{h}_{\mu_0}(\mu'), \tilde{h}_{\mu_0}(\mu_1))^{-N} \omega(\tilde{h}_{\mu_0}(\mu'), \mu_0)^{-N} \\ &\leq C_N \sum_{\mu'} \omega(\tilde{h}_{\mu_0}(\mu_1), \tilde{h}_{\mu_0}(\mu'))^{-N} \omega(\tilde{h}_{\mu_0}(\mu'), \mu_0)^{-N} \\ &\leq C_N \omega(\tilde{h}_{\mu_0}(\mu_1), \mu_0)^{-N+1} \\ &\sim C_N \omega(\mu_1, h_{\mu_0}(\mu_0))^{-N+1}. \end{aligned} \tag{6.1.32}$$

□

Since we have proved that the Fourier integral operators are sparsely represented by discrete shearlets, it is natural to think that such discretization has error bounds with fast decay along the scale. In the next section, we use Theorem 6.1.6 to digitize Fourier integral operators and their action on singularities of function.

6.2 Digitization step

The last section introduced the theory behind the sparse shearlet representation of general Fourier integral operators. Although by using the discrete shearlet coefficients, we can consider the shearlet approach as faithful, i.e., it have similar properties as the continuum counterpart, we would like to establish error bounds on this approximation. Having approximation bounds allows us to know the precision of our discretization. By having such error bounds, we are able to know the extent of precision of our digital shearlet coefficients, which should increase with the number of discrete scales.

Such approximation error results were established for the curvelet transform in [4, 116]. In the case of curvelets, V. de Hoop et al. [4] used the sparsity of FIOs represented by curvelets and the parabolic dyadic nature of such decomposition to define a discretization with rapidly decaying error bound. Having that shearlets possess these two properties (sparse representation and parabolic dyadic nature), we can adapt the curvelets results to our framework. Following V. de Hoop's procedure we are going to first compute error bounds for the case of a pseudodifferential operator, to later extend such results to all Fourier integral operators.

6.2.1 Infinite matrices and operators

In order to proceed with the digitization step, we are going to introduce some matrix classes and operators related to an alternative definition of the dyadic parabolic pseudo-distance. This is known as the *weighted dyadic pseudo-distance.*

Definition 6.2.1. *For weight $\delta > 0$ and shearlet parameters $\mu, \mu' \in \mathcal{M}$ we define the weighted dyadic pseudo-distance $\omega_\delta : \mathcal{M} \times \mathcal{M} \to \mathbb{R}$ by:*

$$\omega_\delta(\mu, \mu') = 2^{|j-j'|(1+\delta)}(1 + 2^{(2+\delta)\min(j,j')}d(\mu, \mu')),$$

where d is the distance function defined in Section 6.1.2.

Let $\delta, r > 0$ be constants and $h_\mu : \mathcal{M} \to \mathcal{M}$ an index mapping for $\mu \in \mathcal{M}$, e.g., the discrete microcanonical relation of a FIO defined in (6.1.30). Following [106], we define the matrix class $M^r_\delta(\chi)$ as follows.

Definition 6.2.2. *Let $\delta, r > 0$ be constants and $h_\mu : \mathcal{M} \to \mathcal{M}$ an index mapping for some $\mu \in \mathcal{M}$. The infinite matrix M whose elements are $M_{\mu\mu'}$, where $\mu, \mu' \in \mathcal{M}$, is an element of the class $\mathcal{M}^r_\delta(h_\mu)$ if and only if there is a constant C_δ such that*

$$|M_{\mu\mu'}| \leq C_\delta 2^r \omega_\delta(\mu', h_\mu(\mu')), \quad \textit{for every } \mu, \mu' \in \mathcal{M}.$$

where ω_δ is the weighted dyadic pseudo-distance from Definition 6.2.1. Moreover, we define

$$\mathcal{M}^r = \cap_{\delta>0}\mathcal{M}^r_\delta(h_\mu). \tag{6.2.1}$$

Finally, we introduce the matrix notation $\mathcal{SH}$ for the shearlet transform (analysis operator).

Definition 6.2.3. *Let $\mathcal{SH}_\psi \subset L^2(\Omega)$ be a discrete shearlet system with generating functions $\psi \in L^2(\Omega)$ defined in (6.1.5). We know by Proposition 6.1.3 that $\mathcal{SH}_\psi$ is a Parseval frame. In addition, let ℓ^2_μ be the space of sequences $\{c_\mu\}_{\mu\in\mathcal{M}} \subset \mathbb{R}$ such that*

$$\sum_{\mu\in\mathcal{M}} |c_\mu|^2 < \infty.$$

The shearlet analysis operator $\mathcal{SH} : L^2(\Omega) \to \ell^2_\mu$ *is given by*

$$\mathcal{SH}(f) = \{c_\mu\}_{\mu\in\mathcal{M}} := \{\langle f, \psi_\mu\rangle\}_{\mu\in\mathcal{M}} \quad \textit{for } f \in L^2(\Omega),$$

which maps $L^2(\Omega)$ functions to its shearlet coefficients c_μ. Moreover, let $\mathcal{SH}^{-1} : \ell^2_\mu \to L^2(\Omega)$ be the shearlet synthesis operator *given by*

$$\mathcal{SH}^{-1}\{c_\mu\}_{\mu\in\mathcal{M}} = \sum_{\mu\in\mathcal{M}} c_\mu\psi_\mu \quad \textit{for all } \{c_\mu\}_{\mu\in\mathcal{M}} \in \ell^2_\mu$$

which maps shearlet coefficients to its shearlet rteconstructions.

We observe that $\mathcal{SH}^{-1}\mathcal{SH} = I$ (the identity operator on $L^2(\mathbb{R}^2)$), and that $\mathcal{SH}\mathcal{SH}^{-1} =:$ Π is an orthogonal projection operator of ℓ^2_μ onto the range of the analysis operator $\mathcal{SH}$. Furthermore, observe that as a matrix operator on ℓ^2_μ the elements of Π are given by

$$\Pi_{\mu'\mu} = \langle \psi_{\mu'}, \psi_\mu \rangle, \quad \text{for every } \mu, \mu' \in \mathcal{M}.$$

Next we state a useful remark.

Remark 6.2.4. *If $\mathcal{A} : L^2(\Omega) \to L^2(\Omega)$, then the matrix $[A] := \mathcal{SH}A\mathcal{SH}^{-1}$ has the same range of $\mathcal{SH}$, since $\mathcal{SH}^{-1}\Pi = \mathcal{SH}^{-1}$, and $\Pi\mathcal{SH} = \mathcal{SH}$. In particular, $[A]\Pi = \Pi[A] = [A]$. Here, and when convenient, we identify bounded operators on ℓ^2_μ with matrices. Finally, let us note that the projection map Π belongs to $\mathcal{M}^0$ (see [106, Lemma 2.4]).*

6.2.2 Pseudodifferential operators, diagonalization, and symbol classes

Now, let us remember from Definition 2.3.9 that a pseudodifferential operator (ΨDO) acting on $L^2(\mathbb{R}^2)$ is an operator $\mathcal{A} : L^2(\Omega) \to L^2(\Omega)$ of the form

$$\mathcal{A}f(x) = \frac{1}{4\pi^2} \int_{\xi \in \mathbb{R}^2 \setminus \{0\}} \int_{y \in \Omega} e^{i(x-y)\cdot\xi} p(x, y, \xi) f(y) dy d\xi, \text{for every } f \in L^2(\Omega) \text{ and } x \in \Omega,$$

where p is the amplitude function (see Definition 2.3.8), also known as the symbol. As we did for the Fourier integral operator, in this chapter we will use the alternative form given by

$$\mathcal{A}f(x) = p(x, D)f(x) = \int_{\xi \in \mathbb{R}^2 \setminus \{0\}} e^{i\langle x, \xi \rangle} p(x, \xi) \hat{f}(\xi) d\xi, \quad \text{for every } f \in L^2(\Omega) \text{ and } x \in \Omega.$$

In addition, we assume that the symbol or amplitude p has order $l \in \mathbb{N}$, i.e.,

$$|\langle \xi, \partial_\xi \rangle^l \partial^\alpha_\xi \partial^\beta_x p(x, \xi)| \leq C_{l,\alpha,\beta}(1 + ||\xi||)^{\frac{|\beta| - |\alpha|}{2}}, \tag{6.2.2}$$

for every multi-indices α and β, where $\langle \xi, \partial_\xi \rangle$ is the differential operator

$$\langle \xi, \partial_\xi \rangle(\cdot) = \xi_1 \partial_{\xi_1}(\cdot) + \xi_2 \partial_{\xi_2}(\cdot).$$

In addition, we have that

$$\langle \xi, \partial_\xi \rangle^l \partial^\alpha_\xi \partial^\beta_x p(x, \xi) := \xi^l \partial^\alpha_\xi \partial^\beta_x p(x, \xi) + \partial^l_\xi \partial^\alpha_\xi \partial^\beta_x p(x, \xi),$$

for every multi-indices α and β. Depending on the exact estimate that a symbol follows, we can define different symbol classes. These classes are characterized by the order of regularity in space and frequency.

Definition 6.2.5 (Symbol classes, [61])**.** *Let $r, \delta, \rho \in \mathbb{R}_+$ be positive constants and $\Omega \subset \mathbb{R}^2$ be a domain. The* symbol class $S^r_{\delta,\rho} \subset C^\infty(\Omega \times \mathbb{R}^2 \setminus \{0\})$ *is defined by the set*

$$\begin{aligned} S^r_{\rho,\delta} := S^r_{\rho,\delta}(\Omega) = \{ & p \in C^\infty(\Omega \times \mathbb{R}^2) : \forall \alpha, \beta, \exists C_{\alpha,\beta} > 0 \text{ } s.t. \\ & |\partial^\alpha_x \partial^\beta_\xi p(x, \xi)| < C_{\alpha,\beta}(1 + ||\xi||)^{r - \delta|\alpha| - \rho|\beta|} \}. \end{aligned}$$

The simplest symbol classes are solely defined by their regularity r, with $\delta = 0$ and $\rho = 1$,

$$S^r := S^r_{1,0}(\Omega) = \{p \in C^\infty(\Omega \times \mathbb{R}^2) : \forall \alpha, \beta, \exists C_{\alpha,\beta} > 0 \textit{ s.t.} \\ |\partial_x^\alpha \partial_\xi^\beta p(x,\xi)| < C_{\alpha,\beta}(1 + ||\xi||)^{r-|\beta|}\}.$$

In addition, we define the symbol class $S^r_{1/2,rad} \subset C^\infty(\Omega \times \mathbb{R}^2)$ as

$$S^r_{1/2,rad} := S^r_{1/2,rad}(\Omega) = \{p \in C^\infty(\Omega \times \mathbb{R}^2) : \langle \xi, \partial_\xi \rangle^l p \in S^r_{1/2,1/2} \textit{ for all } l \in \mathbb{N}\}.$$

The symbol class $S^r_{1/2,rad}$ contains symbols of *generalized Radon transforms* (see [116]), and is relevant for this chapter.

Remark 6.2.6. *From Definition 6.2.5 we can see that a symbol that follows the estimate (6.2.2) can be regarded as an element of the symbol class $S^0_{1/2,rad}$. Thus, we have that $p \in S^0_{1/2,rad}$ when $\langle \xi, \partial_\xi \rangle^l p \in S^0_{1/2,1/2}$ for all $l \in \mathbb{N}$. Therefore, $p \in S^r_{1/2,rad}$ when $\langle \xi, \partial_\xi \rangle^l p \in S^r_{1/2,1/2}$ for all $l \in \mathbb{N}$.*

Pseudodifferential operators of order r, are the most important example of operators with representation matrices of class $\mathcal{M}^r$ (see (6.2.1)). A stationary phase analysis from [116] shows that for a shearlet ψ_μ we have that

$$\mathcal{A}\psi_\mu = 2^{jr} f_\mu,$$

where $f_\mu \in L^2(\Omega)$ is given by

$$\hat{f}_\mu(\xi) = 2^{-3j/4} \hat{g}_{j,k}(\xi) e^{-i\langle x_m^{j,k}, \xi \rangle}, \tag{6.2.3}$$

and there exists a constant $C_{l,\alpha,N} > 0$ such that $\hat{g}_{j,k}$ satisfies

$$|\langle \nu_{j,k}, \partial_\xi \rangle^l \partial_\xi^\alpha \hat{g}_{j,k}| \leq C_{l,\alpha,N} 2^{-j(l+\frac{|\alpha|}{2})}(1+2^{-j}|\langle \nu_{j,k}, \xi \rangle| + 2^{-j/2}||\xi - W_{j,k}||)^{-N} \quad \text{for all } N \in \mathbb{N}. \tag{6.2.4}$$

Here, $||\xi - W_{j,k}||$ denotes the distance of ξ to the wedge $W_{j,k}$ supporting $\hat{\psi}_{j,k}(\xi) = \hat{\psi}(\xi A_j^{-1} S_k^{-1})$, and $\nu_{j,k}$ is the center point of the wedge $W_{j,k}$ in the frequency domain. Following (6.1.27), such f_μ is a "shearlet-like function" centered at μ.

Since a pseudodifferential operator is a special case of Fourier integral operators, using Theorem 6.1.6 we also have a sparse representation of $\mathcal{A}$ using shearlets leading to the estimate

$$|\langle \psi_{\mu'}, f_\mu \rangle| \leq C(\delta)\omega_\delta(\mu', \mu) \quad \text{for all } \delta > 0 \text{ s.t. } \langle \psi_{\mu'}, f_\mu \rangle \in \mathcal{M}^0(I),$$

where $M^0(I)$ is the infinite matrix class of order $r = 0$ given by Definition 6.2.2.

Notice that, if the symbol p is elliptic of order m, then by definition the degree of the principal symbol, namely p_0 (see Definition 2.3.8), is m. Next, if p_0 is homogeneous of order 0, then $p_0(x,\xi) = p_0(x, \xi/||\xi||)$. Therefore, we have the following diagonalization result, which is an adaptation of the phase-linearization of V. de Hoop [116, Lemma 3.1] to the shearlet case.

Lemma 6.2.7 ([116, Lemma 3.1])**.** *Suppose that $\mathcal{A} : L^2(\Omega) \to L^2(\Omega)$ is a pseudodifferential operator with homogeneous principle symbol $p_0(x, \xi)$ of order 0. Then we can express $\mathcal{A}$ as*

$$\mathcal{A}\psi_\mu = p_0(x_m^{j,k}, \nu_{j,k})\psi_\mu + 2^{-j/2} f_\mu \quad \textit{for all } \mu \in \mathcal{M}, \tag{6.2.5}$$

where f_μ is a shearlet-like function *centered at $(x_m^{j,k}, \nu_{j,k})$ (see Definition 6.1.18).*

The proof of Lemma 6.2.7 can be easily adapted to shearlets using [116].

Remark 6.2.8. *If we write in Equation* (6.2.5) *$r_\mu = 2^{-j/2} f_\mu$, by taking inner products with ψ'_μ we obtain that the elements of the matrix representation of $\mathcal{A}$ are given by*

$$[\mathcal{A}]_{\mu'\mu} = p_0(x_m^{j,k}, \nu_{j,k})\Pi_{\mu'\mu} + \langle \psi_{\mu'}, r_\mu \rangle. \tag{6.2.6}$$

In addition, if $\mathcal{A}$ is elliptic (see Definition 2.3.9), we can obtain uniform upper and lower bounds for the symbol $p_0(x, \xi)$. This means $C^{-1} \leq |p_0(x, \xi)| \leq C$ for some positive constant C. By (6.2.6)*, we have*

$$p_0(x_m^{j,k}, \nu_{j,k})^{-1}[\mathcal{A}]_{\mu'\mu} - \Pi_{\mu'\mu} \in \mathcal{M}^{-1/2}(I). \tag{6.2.7}$$

Next, also by (6.2.6) *we obtain*

$$|p_0(x_m^{j,k}, \nu_{j,k}) - \langle \psi_\mu, \psi_\mu \rangle^{-1} [\mathcal{A}]_{\mu\mu}| \leq C 2^{-j/2}.$$

It follows that (6.2.7) *holds with $p_0(x_m^{j,k}, \nu_{j,k})$ replaced by the normalized diagonal*

$$D_\mu = \Pi_{\mu\mu}^{-1}[\mathcal{A}]_{\mu\mu},$$

Finally, by Theorem 6.1.6 we have that the elements of $[\mathcal{A}]$ outside the diagonal decays fast for large scale j. This diagonalization property due to the sparse representation is very useful for inverse problems. In fact, it allows us to have optimal digital representations of such operators, in the sense that their error bounds decay fast. In addition, it also allows us to invert the operator matrix $[\mathcal{A}]$ on the range of $\mathcal{SH}$ restricted to a finite scale j sufficiently large.

In the next section we extend this approximation theoretical results to general Fourier integral operators. In addition, in Section 6.2.4 we will use the decay estimates of the shearlet discretization to develop a faithful digitization method with fast decaying error bounds. Our approach is again strongly based on the theory presented by Hoop et al. in [116].

6.2.3 Matrix approximation of Fourier integral operators

Let $\mathcal{P} : L^2(\Omega) \to L^2(\Xi)$ be a Fourier integral operator given by (6.1.1). In addition, let $\phi(y, \xi), p \in C^\infty(\Xi \times \mathbb{R}^2 \setminus \{0\})$ be the phase and symbol of $\mathcal{P}$. Given the phase ψ we know that the microcanonical relation of $\mathcal{P}$ is given by the mapping χ defined in (6.1.4).

As explored previously on Section 6.1, the action of $\mathcal{P}$ in the shearlet ψ_μ is given by

$$\mathcal{P}\psi_\mu(y) = 2^{-3j/4}\int p(y,\xi)\hat{\psi}_{j,k}(\xi)e^{i(\phi(y,\xi)-\langle\xi,x_m^{j,k}\rangle)}d\xi. \tag{6.2.8}$$

Therefore, we can associate a kernel $\mathcal{P}_{j,k}$ defined by

$$\mathcal{P}_{j,k}(y,x_m^{j,k}) := (\mathcal{P}\psi_\mu)(y). \tag{6.2.9}$$

Using this kernel we can write the infinite Fourier integral operator matrix as

$$[\mathcal{P}]_{\mu'\mu} = \int \overline{\psi'_\mu(y)}\mathcal{P}\psi_\mu(y)dy = \int \overline{\psi'_\mu(y)}\mathcal{P}_{j,k}(y,x_m^{j,k})dy. \tag{6.2.10}$$

Now, by Remark 6.2.4 we have that $\mathcal{SHPSH}^{-1}$ is represented by the matrix operator $[\mathcal{P}]$. By abuse of notation, we simply denote $\mathcal{P} = \mathcal{SH}^{-1}[\mathcal{P}]\mathcal{SH}$. Following the approach in [116], we now aim for an approximation of $\mathcal{P}\psi_\mu$ via expansions of the phase $\phi(y,\xi)$ and the symbol $p(y,\xi)$ near the microlocal support of ψ_μ, i.e., near the wavefront set. Before we present the main result of this section we will introduce the definition of the *spatial microcanonical relations* of a Fourier integral operator associated to a shearlet index (j,k).

Definition 6.2.9 (Spatial microcanonical relation)**.** *Let $\mathcal{P} : L^2(\Omega) \to L^2(\Xi)$ be a Fourier integral operator with phase ϕ and $j,k \in \mathbb{Z}$. The* spatial microcanonical relation *of $\mathcal{P}$ associated to j,k is the mapping $T_{j,k} : \Xi \to \Omega$ given by*

$$T_{j,k}(y) = -\frac{\partial\phi}{\partial\xi}(y,\nu_{j,k}) \quad \textit{for all } y \in \Xi \textit{ and } j,k \in \mathbb{Z}, \tag{6.2.11}$$

where $\nu_{j,k}$ is the center of the wedge $W_{j,k}$ in the frequency domain.

The next theorem presents the main result on optimal shearlet representation of pseudodifferential operators.

Theorem 6.2.10. *Let $\mathcal{P} : L^2(\Omega) \to L^2(\Xi)$ be a Fourier integral operator with phase ϕ, and let $T_{j,k}$ be given by (6.2.11), where $j,k \in \mathbb{Z}$. Moreover, let $\mathbf{1}_{j,k}$ be a smooth cutoff function to the wedge $W_{j,k}$, centered in $\nu_{j,k}$, supporting $\hat{\psi}_{j,k}$. Then, there exist functions $\alpha_{j,k}^{(r)} : \Xi \to \mathbb{R}$ and $\hat{\vartheta}_{j,k}^{(r)} : \mathbb{R}^2 \setminus \{0\} \to \mathbb{R}$ with*

$$e^{i\frac{1}{2}\xi_1^{-1}\xi_2^2\partial_\xi^2\phi(y,\nu_{j,k})}\mathbf{1}_{j,k}(\xi) = \sum_{r=1}^{R}\alpha_{j,k}^{(r)}(y)\hat{\vartheta}_{j,k}^{(r)}(\xi) \quad \textit{for all } (y,\xi) \in \Xi\times\mathbb{R}^2\setminus\{0\} \textit{ and } j,k \in \mathbb{Z}, \tag{6.2.12}$$

so that for a shearlet ψ_μ, one can express $\mathcal{P}\psi_\mu$ as

$$(\mathcal{P}\psi_\mu)(y) = p(y,\nu_{j,k})\sum_{r=1}^{R}\alpha_{j,k}^{(r)}(y)(\hat{\vartheta}_{j,k}^{(r)} * \psi_\mu)(T_{j,k}(y)) + 2^{-j/2}f_\mu \quad \textit{for all } y \in \Xi \textit{ and } \mu \in \mathcal{M}, \tag{6.2.13}$$

with $R \sim \mathcal{O}(j/\log j)$, where f_μ is a shearlet-like function centered at $(y,\nu_{j,k})$ in the phase space (see Definition 6.1.18).

Proof. Let $\nu_{j,k}$ be the center of the shearlet wedge $W_{j,k}$ in the Fourier domain. Then ρ_μ for $\mu \in \mathcal{M}$ given by (6.1.27) is a shearlet-like function centered in the phase space at $(x_m^{j,k}, \nu_{j,k})$. By homogeneity in ξ the first-order Taylor expansion of $\phi(y,\xi)$ along the $\nu_{j,k}$ axis, is given by

$$\phi(y,\xi) - \langle \xi, x_m^{j,k} \rangle \approx \langle \xi, \frac{\partial \phi}{\partial \xi}(y, \nu_{j,k}) - x_m^{j,k} \rangle + h_2(y,\xi), \tag{6.2.14}$$

where the error $h_2(y,\xi)$ follows the estimate (6.2.2) on $W_{j,k}$. A consequence of this is that $e^{ih_2(y,\xi)}$ is a symbol of class $S^0_{1/2,rad}$ if ξ is localized to $W_{j,k}$ (see Definition 6.2.5).

Let $T_{j,k} : \Xi \to \Omega$ be given by (6.2.11). If $b_{j,k}(y,\xi)$ is the symbol of order 0, i.e.,

$$b_{j,k}(y,\xi) = (p(y,\xi) e^{ih_2(y,\xi)})|_{y = T^{-1}_{j,k}(x)}, \tag{6.2.15}$$

then, by Equation (6.2.14) we can write $\mathcal{P}\psi_\mu(y)$ as

$$\mathcal{P}\psi_\mu(y) = [b_{j,k}(y, D)\psi_\mu]_{x = T_{j,k}(y)}. \tag{6.2.16}$$

Such decomposition rewrites the Fourier integral operator $\mathcal{P}$ depending on the shearlet parameters $j,k \in \mathbb{Z}$, followed by a change of coordinates. Next, the decomposition of $\mathcal{P}$ by shearlets in (6.2.16) can be used to show that the matrix $[\mathcal{P}]$, given by (6.2.10), belongs to the class $\mathcal{M}^0(\chi)$, where χ is given by (6.1.4).

Now, we approximate the matrix elements $[\mathcal{P}]_{\mu'\mu}$ with an expansion of the symbol and phase. This approximation has an error of at most $2^{-j/2}$. More precisely, the matrix errors are of class $\mathcal{M}^{-1/2}(\chi)$. Also, the principal part of the symbol $p(y,\xi)$, namely $p_0(y,\xi)$, is homogeneous of order 0. Following Lemma 6.2.7, we can replace $p_0(y,\xi)$ by either $p_0(y, \nu_{j,k})$ or $p_0(y_m^{j,k})$, where $y_m^{j,k} \in \Xi$ is given by

$$y_m^{j,k} = T^{-1}_{j,k}(x_m^{j,k}).$$

This allows us to modify the Fourier integral operator matrix by a matrix of class $\mathcal{M}^{-1/2}(\chi)$. Also, the symbol $h_2(y,\xi)$ is homogeneous of order 1 and class $S^0_{1/2,rad}$ on the support of $\hat{\psi}_{j,k}(\xi)$. This means that we need to take into account the second-order terms in its Taylor expansion to obtain an approximation of order $-1/2$. Similar to [116] we do Taylor expansion in the $\xi-$component in direction perpendicular to $\nu_{j,k}$. This has the advantage of preserving the homogeneity of order 1 in the radial direction. The latter is mainly due to the non-isotropic nature of the parabolic decomposition, resulting on the wedge tilling of the Fourier domain. To simplify the notation, we consider the case when the vector $\nu_{j,k}$ is aligned with the ξ_1-axis. This results on the expansion

$$\phi(y,(\xi_1,\xi_2)) = \xi_1 \phi(y,(1,\xi_2/\xi_1)) = \xi \cdot \frac{\partial \phi}{\partial \xi}(y,\nu_{j,k}) + \frac{1}{2}\partial \xi_2^2 \xi_1 \cdot \frac{\partial^2 \phi}{\partial \xi_2^2}(y,\nu_{j,k}) + h_3(y,\xi). \tag{6.2.17}$$

We also have that if ξ lies on the support of the shearlet $\hat{\psi}_{j,k}$ then $h_3(y,\xi) \in S^{-1/2}_{1/2,rad}$. This allows us to replace $e^{ih_3(y,\xi)}$ by 1 changing the matrix $[\mathcal{P}]$ by terms in the class $\mathcal{M}^{-1/2}(\chi)$

(see proof of Lemma 3.1 in [116]). Therefore we can replace the symbol $p(y,\xi)e^{ih_2(y,\xi)}$ on $W_{j,k}$, up to errors of order $-1/2$, by

$$p(y,\nu_{j,k})e^{i\frac{1}{2}\xi_1^{-1}\xi_2^2\cdot\partial_\xi^2\phi(y,\nu_{j,k})}\mathbf{1}_{j,k}(\xi), \tag{6.2.18}$$

where $\mathbf{1}_{j,k}$ is a smooth cutoff function to the wedge $W_{j,k}$ supporting $\hat{\psi}_{j,k}$. Now, one can see that the exponent in (6.2.18) separates the variables y and ξ. In addition, by (6.2.17), it is bounded by a constant C which does not depend on the parameters (j,k). As in [116], we have that this decomposition of the complex exponential with arguments uniformly bounded by a polynomial allows us to have the tensor-product representation of the symbol as follows:

$$e^{i\frac{1}{2}\xi_1^{-1}\xi_2^2\partial_\xi^2\phi(y,\nu_{j,k})}\mathbf{1}_{j,k}(\xi) \approx \sum_{r=1}^{R} \alpha_{j,k}^{(r)}(y)\hat{\vartheta}_{j,k}^{(r)}(\xi). \tag{6.2.19}$$

In order to obtain an error of size $2^{-j/2}$ we require that $C^R/R! \leq 2^{-j/2}$, or $R \sim \mathcal{O}(j/\log j)$. □

In Theorem 6.2.10 we approximate the Fourier integral operator in (6.1.1) with a sum of R modified shearlets with symbols $\tilde{\psi}_{r;\mu}(x) = (\vartheta_{j,k}^{(r)} * \psi_\mu)(x)$ to order $p(y,\nu_{j,k})\alpha_{j,k}^{(r)}(y)$. As we can notice in (6.2.13) this approximation has an error of order $O(2^{-j/2})$ as $j \to \infty$. This means that the error decreases exponentially with the number of scales that we used. In this approximation we have also made use of the coordinate transforms $\{T_{j,k}\}_{j,k\in\mathbb{Z}}$ coming from the microcanonical relation of $\mathcal{P}$. Now, the approximation in Theorem 6.2.10 allows us to discretize and later digitize the application of the operator $\mathcal{P}$, where the evaluations are done in a discrete set of directions in the Fourier domain $\{\nu_{j,k}\}_{j,k}$. In the next section, similar to the curvelet case [4], we take a step forward and also evaluate such approximations on discrete grids of the domain and co-domain of $\mathcal{P}$. Such further discretization is needed for the implementation of our algorithms.

6.2.3.1 Further approximations of Fourier integral operators

For the digitization of $\mathcal{P}$ we will first introduce a specific notion of discrete grid associated to a shearlet system $\{\psi_\mu\}_{\mu\in\mathcal{M}}$.

Definition 6.2.11. *Let $\mathcal{SH}_\psi = \{\psi_\mu\}_{\mu\in\mathcal{M}}$ be a discrete shearlet system defined by* (6.1.10)*. Moreover, let $\mathcal{P} : L^2(\Omega) \to L^2(\Xi)$ be a pseudodifferential operator of the form 6.1.1. The* $\Omega-$grid *for the shearlet based discretization is given by $\{x_m^{j,k}\}_{(j,k,m)=\mu\in\mathcal{M}}$, where $x_m^{j,k}$ is the center point of* $\operatorname{supp}(\psi_{\mu=(j,k,m)})$ *in the spatial domain. In addition, let $T_{j,k} : \Xi \to \Omega$ be the mapping defined by* (6.2.11)*. The* $\Xi-$grid *for the sharlet based discritization is defined by $\{y_m^{j,k}\}_{(j,k,m)=\mu\in\mathcal{M}}$, where $y_m^{j,k} := T_{j,k}^{-1}(x_m^{j,k})$.*

As one can observe in Definition 6.2.11, the discrete points $\{y_m^{j,k}\}_{(j,k,m)\in\mathcal{M}}$ are obtained by mapping the grid $\{x_m^{j,k}\}_{(j,k,m)\in\mathcal{M}}$ under the spatial component microcanonical relation

$T_{j,k}^{-1}$. Under these assumptions, we are able to approximate the functions $p(y, \nu_{j,k})$, $\frac{\partial \phi}{\partial \xi}(y, \nu_{j,k})$, and $\frac{\partial^2 \phi}{\partial \xi_2^2}(y, \nu_{j,k})$, in (6.2.17) by $p(y_m^{j,k}, \nu_{j,k})$, $\frac{\partial \phi}{\partial \xi}(y_m^{j,k}, \nu_{j,k})$, and $\frac{\partial^2 \phi}{\partial \xi_2^2}(y_m^{j,k}, \nu_{j,k})$, respectively. Also, by Theorem 6.2.10, such approximations lead to an error of order $O(2^{-j/2})$ as $j \to \infty$.

Now, by using Theorem 6.2.10 and the aforementioned approximations we get the form

$$(\mathcal{P}\psi_\mu)(y) = p(y_m^{j,k}, \nu_{j,k})(\vartheta_\mu * \psi_\mu)(T_{j,k}(y)) + 2^{-j/2} f_\mu, \tag{6.2.20}$$

where f_μ is a shearlet-like function centered at $(y_m^{j,k}, \nu_{j,k})$ (see Definition 6.1.21), and $\vartheta_\mu : \mathbb{R}^2 \to \mathbb{R}^2$ is defined by

$$\hat{\vartheta}_\mu(\xi) = e^{i\frac{1}{2}\xi_1^{-1}\xi_2^2\partial_\xi^2\phi(y_m^{j,k}, \nu_{j,k})}\mathbf{1}_{j,k}. \tag{6.2.21}$$

Furthermore, we can approximate the change of coordinates $T_{j,k}$ by a Taylor expansion of $\phi(y, \nu_{j,k})$ around $(y_m^{j,k}, \nu_{j,k})$, obtaining

$$(\mathcal{P}\psi_\mu)(y) = p(y_m^{j,k}, \nu_{j,k})(\vartheta_\mu * \psi_\mu)(DT_\mu(y - y_m^{j,k}) + M_\mu \cdot (y - y_m^{j,k})^2) + 2^{-j/2} f_\mu, \tag{6.2.22}$$

where

$$\begin{aligned} DT_\mu &= \frac{\partial T_{j,k}}{\partial y}(y_m^{j,k}) = \frac{\partial^2 \phi}{\partial_\xi \partial y}(y_m^{j,k}, \nu_{j,k}), \\ M_\mu &= \frac{1}{2}\frac{\partial^2 \phi}{\partial y^2}(y_m^{j,k}, \nu_{j,k})\nu_{j,k}, \end{aligned}$$

On the one hand, following the interpretation in [4], in this approximation, M_μ describes the curvature of a localized plane wave attached to ψ_μ under the corresponding microcanonical relation. On the other hand, DT_μ describes a rigid motion, shear along the wavefront and dilations along and perpendicular to the wavefront. Now that we have a shearlet-based approximation of the Fourier integral operator $\mathcal{P}$ with controlled error bounds, and since such bounds decay exponentially with the scale, we are able to faithfully digitize the action of such operator, as well as the microcanonical relation. We are exploring this situation in the next section.

6.2.4 Digital approximation of Fourier integral operators

The final digital approximation is based on the expansion presented in Theorem 6.2.10. This theorem is used for the evaluation of the approximate action of $\mathcal{P}$ on a function f discretized by the spatial and frequency sample points y_n and ξ_l, respectively. Following the approach of V. de Hoop et al in [4], the digitization is chosen to match the structure of the digital shearlet transform. This enables us to switch from the coefficients of the shearlet transform to data in the frequency domain efficiently through standard *Fast Fourier Transform* (FFT). Here, we assume in our analysis that the partial derivatives $\frac{\partial^2 \phi}{\partial \xi_2^2}(y, \nu_{j,k})$ and the functions $T_{j,k}(y)$ and $T_{j,k}^{-1}(x)$ are known a-priori. Let us first introduce a new notion of discrete grids for the further digitization of $\mathcal{P}$.

Definition 6.2.12. *Let $\mathcal{SH}_\psi = \{\psi_\mu\}_{\mu\in\mathcal{M}}$ be a discrete shearlet system defined by* (6.1.10)*. Moreover, let $\mathcal{P} : L^2(\Omega) \to L^2(\Xi)$ be a pseudodifferential operator of the form 6.1.1. In addition, let $N \in \mathbb{N}$ and $\{x_i\}_{i\in\mathcal{I}} \subset \Omega$ be a grid in Ω, where $x_i = N^{-1}2\pi i$ for $i \in \mathcal{I}$ where $I \subset N(2\pi)^{-1}\Omega$ discrete. In addition, for $j \in \mathbb{Z}$ let $\Xi^j \subset \mathbb{Z}^2$ be given by*

$$\tilde{\Xi}^j = \left\{ l \in \mathbb{Z}^2 \middle| -\frac{N'_j}{2} \le l_1 < \frac{N'_j}{2}, -\frac{N''_j}{2} \le l_2 < \frac{N''_j}{2} \right\}. \tag{6.2.23}$$

Next, let us denote the points in $\tilde{\Xi}^j$ by $\tilde{\Xi}^j_l$ and let $N_j, N''_j \in \mathbb{N}$ be even natural numbers with $N'_j > 2^j$ and $N''_j > 2^{j/2}$. The frequency grid $\{\xi^{j,k}_l\}_{(j,k,l)\in\mathcal{S}} \subset \mathbb{R}^2 \setminus \{0\}$, for $\mathcal{S} \subset \mathbb{Z} \times \mathbb{Z} \times \cup_{j\in\mathbb{Z}}\Xi^j$, is defined by

$$\xi^{j,k}_l = S^{-1}_{j,k}(A_j D_j^{-1}\tilde{\Xi}^j_l + 2^{j-2}e_1), \tag{6.2.24}$$

where $S_{j,k}$ was defined in Definition 6.1.7 and $D_j = \operatorname{diag}(N'_j, N''_j)$. Finally let $\{y_i\}_{i\in\mathcal{I}}$ a discrete $\Xi-$grid, where $y_i = T^{-1}_{j,k}(x_i)$ and $T_{j,k}$ is the mapping defined by (6.2.11)*.*

Let us start the digitization of Theorem 6.2.10 from the discrete adjoint shearlet transform. For that we start by writing the convolutions $(\vartheta^{(r)}_{j,k} * \psi_\mu)(T_{j,k}(y))$ in (6.2.22) in the Fourier domain, namely

$$\tilde{\psi}_\mu(y) = (\mathcal{P}\psi_\mu)(y) \approx p(y, \nu_{j,k})2^{-3j/4}\sum_{r=1}^{R_{j,k}} \alpha^{(r)}_{j,k}(y) \sum_{\xi\in\mathbf{1}_{j,k}} e^{i\langle T_{j,k}(y),\xi\rangle}\hat{\vartheta}^{(r)}_{j,k}(\xi)\hat{\psi}_{j,k}(\xi). \tag{6.2.25}$$

Therefore, since

$$(\mathcal{P}f)(y) = \sum_\mu c_\mu(\mathcal{P}\psi_\mu)(y),$$

where c_μ are the discrete shearlet coefficients of f, we obtain the action of such expansion on a function $f \in L^2(\mathbb{R}^2)$:

$$\begin{aligned}(\mathcal{P}f)(y) &\approx \sum_\mu c_\mu\tilde{\psi}_\mu(y)\\ &= \sum_{j,k} p(y, \nu_{j,k})\sum_{r=1}^{R_{j,k}} \alpha^{(r)}_{j,k}(y) \sum_{\xi\in\mathbf{1}_{j,k}} e^{i\langle T_{j,k}(y),\xi\rangle}\hat{f}(\xi)\hat{\psi}^2_{j,k}(\xi)\hat{\vartheta}^{(r)}_{j,k}(\xi).\end{aligned} \tag{6.2.26}$$

In the following the amplitudes $p(y, \nu_{j,k})$ are absorbed by the functions $\alpha^{(r)}_{j,k}(y)$. Notice that the form of (6.2.26) is similar to the adjoint shearlet transform,

$$f(x) = \sum_\mu c_\mu\psi_\mu(x) = \sum_\xi\sum_{j,k} e^{i\langle x,\xi\rangle}\hat{f}(\xi)\hat{\psi}^2_{j,k}(\xi). \tag{6.2.27}$$

The next theorem introduces the digitization of the representation presented in (6.2.26) and (6.2.27).

Theorem 6.2.13. *Let $\mathcal{SH}_\psi = \{\psi_\mu\}_{\mu\in\mathcal{M}}$ be a discrete shearlet system defined by* (6.1.10)*. Moreover, let $\mathcal{P} : L^2(\Omega) \to L^2(\Xi)$ be a pseudodifferential operator of the form 6.1.1. Let $\{x_i\}_{i\in\mathcal{I}}$, $\{y_i\}_{i\in\mathcal{I}}$ and $\{\xi_l^{j,k}\}_{(j,k,l)\in\mathcal{S}}$ be the discrete grids introduced in Definition 6.2.12. Then, the complete the digitization of* (6.2.27) *with the discrete adjoint transform, given by*

$$f(x_i) = \sum_{j,k}\sum_{l\in\Xi^j} e^{i\langle x_i,\xi_l^{j,k}\rangle}\hat{f}(\xi_l^{j,k})\hat{\psi}^2_{j,k}(\xi_l^{j,k}). \tag{6.2.28}$$

Finally, the digitization of (6.2.26) *is given by*

$$(\mathcal{P}f)(y_i) = \sum_{j,k}\sum_{r=1}^{R_{j,k}}\alpha^{(r)}_{j,k}(y_i)\sum_{l\in\Xi^j} e^{2\pi i\langle x_i,\xi_l^{j,k}\rangle}\hat{f}(\xi_l^{j,k})\hat{\psi}^2_{j,k}(\xi_l^{j,k})\hat{\vartheta}^{(r)}_{j,k}(\xi_l^{j,k}). \tag{6.2.29}$$

Proof. Let us assume first that $\{x_m^{j,k}\}_{(j,k,m)=\mu\in\mathcal{M}} \subset \Omega$ is the shearlet $\Omega-$grid introduced in Definition 6.2.11. In addition, let $\{\xi_l^{j,k}\}_{(j,k,l)\in\mathcal{S}} \subset \mathbb{R}^2\setminus\{0\}$ be the frequency grid from Definition 6.2.12. Then, the digitization of the forward transform is given by

$$\tilde{f}_{j,k,m} = \frac{2^{-3j/8}}{(2\pi)^2\sigma'_j\sigma''_j}\sum_l \hat{f}(\xi_l^{j,k})\hat{\psi}_{j,k}(\xi_l^{j,k})e^{i\langle x_m^{j,k},\xi_l^{j,k}\rangle} \approx f_\mu. \tag{6.2.30}$$

Also, the discretization of the adjoint transform $\hat{f}(\xi)\hat{\psi}^2_{j,k}(\xi) = \sum_{\mu':j'=j,k'=k} f_{\mu'}\hat{\psi}_{j,k}(\xi)$ results on

$$\hat{f}(\xi_l^{j,k})\hat{\psi}^2_{j,k}(\xi_l^{j,k}) = 2^{-3j/8}\left(\sum_m \tilde{f}_{j,k,m}e^{-i\langle x_m^{j,k},\xi_l^{j,k}\rangle}\right)\hat{\psi}_{j,k}(\xi_l^{j,k}). \tag{6.2.31}$$

Let us notice that by construction, the inner product in the phase of the complex exponential in (6.2.31) becomes

$$\langle x_m^{j,k},\xi_l^{j,k}\rangle = (A_jD_j^{-1}\Xi_l^j + 2^{j-2}e_1)^\intercal A_j^{-1}m = \frac{\pi m_1}{2} + 2\pi\left(\frac{m_1l_1}{2^j}+\frac{m_2l_2}{2^{j/2}}\right). \tag{6.2.32}$$

This implies that the specific choice of frequency points $\xi_l^{j,k}$ can be fast evaluated by $\hat{f}(\xi_l^{j,k})\hat{\psi}_{j,k}(\xi_l^{j,k})$ from the data shearlet coefficients $\tilde{f}_{j,k,m}$ for $l\in\Xi^j$,

$$\hat{f}(\xi_l^{j,k})\hat{\psi}_{j,k}(\xi_l^{j,k})e^{\pi i m_1/2} = 2^{-3j/8}N'_jN''_j\sum_m \tilde{f}_{j,k,m}e^{-1\langle x_m^{j,k},\xi_l\rangle}, \tag{6.2.33}$$

where $\xi_l = l$ and $x_m^{j,k} = D_j^{-1}m$ with $m\in\Xi^j$, and $N'_jN''_j = (2\pi)^2\det D_j$. If the values of $\tilde{f}_{j,k,m}$ are known we can make us of the 2−dimensional FFT for the evaluation of $\hat{f}(\xi_l^{j,k})$ and $\hat{\psi}_{j,k}(\xi_l^{j,k})$ in (6.2.33).

Now, we complete the digitization of (6.2.27) with the discrete adjoint transform, given by:

$$f(x_i) \approx \sum_{j,k}\sum_{l\in\Xi^j} e^{i\langle x_i,\xi_l^{j,k}\rangle}\hat{f}(\xi_l^{j,k})\hat{\psi}^2_{j,k}(\xi_l^{j,k}).$$

Next, since $y_i = T^{-1}_{j,k}(x_i)$, the dot product in the phase of the complex exponential in (6.2.26) results in the equality

$$\langle T_{j,k}(y_i), \xi^{j,k}_l \rangle = \langle x_i, \xi^{j,k}_l \rangle \tag{6.2.34}$$

and we obtain the digitization of (6.2.26) given by

$$(\mathcal{P}f)(y_i) \approx \sum_{j,k} \sum_{r=1}^{R_{j,k}} \alpha^{(r)}_{j,k}(y_i) \sum_{l \in \Xi^j} e^{2\pi i \langle x_i, \xi^{j,k}_l \rangle} \hat{f}(\xi^{j,k}_l) \hat{\psi}^2_{j,k}(\xi^{j,k}_l) \hat{\vartheta}^{(r)}_{j,k}(\xi^{j,k}_l).$$

□

Remark 6.2.14. *Since $\mathcal{S}(\Omega)$ is a dense subspace of $L^2(\Omega)$, Theorem 6.2.13 also holds for Fourier integral operators acting on Schwartz functions, i.e., $\mathcal{P} : \mathcal{S}(\Omega) \to \mathcal{S}(\Xi)$. This means, we can also apply such results to the continuum convolutional operator $\mathcal{K}_{\boldsymbol{\theta}} : \mathcal{S}(\Omega) \to \mathcal{S}(\Omega)$ in order to digitize it as well as its microcanonical relation.*

Notice that for fast implementation, the terms in (6.2.28) can be evaluated USFFT (unequally spaced fast Fourier transform) [37] from the irregularly spaced set of points $\xi^{j,k}_l$ to x_i. In a similar fashion, in order to evaluate the terms in (6.2.29), one can make use of the 2−dimension FFT in the fast evaluation of $\hat{f}(\xi^{j,k}_l)$ and $\hat{\psi}_{j,k}(\xi^{j,k}_l)$ from the shearlet transform of the data. In contrast to (6.2.28), the transform USFFT $\xi^{j,k}_l \to x_i$ has to be evaluated for each wedge $W_{j,k}$ separately. This happens, since the functions $T_{j,k}(y)$, $\alpha^{(r)}_{j,k}(y)$ are different for each wedge. Let $f_{j,k}$ be the data component corresponding to the wedge $W_{j,k}$ given by

$$f_{j,k}(x_i) = \sum_{\mu' : j'=j, k'=k} f_{\mu'} \psi_{\mu'}(x_i) \tag{6.2.35}$$

This form shows the organization by wedges of (6.2.29), $(\mathcal{P}f)(y_i) \approx \sum_{j,k} (\mathcal{P}f_{j,k})(y_i)$. This finalizes the digitization of Fourier integral operators by shearlet-based discretization. We are now ready to introduce in the next section the digital microcanonical relation.

6.2.5 Digital microcanonical relation

In our analysis we have a parametrized change of coordinates induced by the microcanonical relation $T_{j,k}(y) = \frac{\partial \phi}{\partial \xi}(y, \nu_{j,k})$. Furthermore, we also showed how it is easy to derive this notion on the digital realm, where the grid points x_i are mapped to points y_i by $T^{-1}_{j.k}$. One should notice, however, that the complete microcanonical relation maps elements between phase spaces, namely (x, ξ) to (y, λ). In this setting, we can define the frequency component of the microcanonical relation by

$$\xi \to \tilde{T}_i(\xi) = \frac{\partial \phi}{\partial y}(y_i, \xi).$$

This function maps digital frequency points $\xi_l^{j,k}$ to digital frequency points $\lambda_l^{j,k} = \tilde{T}_i(\xi_l^{j,k})$, defining the frequency part of the digital microcanonical relation. Finally, the complete digital microcanonical relation is defined as follows.

Definition 6.2.15 (Digitial microcanonical relation)**.** *Let $\mathcal{P}$ a Fourier integral operator with phase ϕ, and $\{x_i\}$ and $\{\xi_l^{j,k}\}$ discrete samples of the spatial and frequency domain as described in Section 6.2.4. Let $T_{j,k}$ be the spatial component of the discrete microcanonical relation given by*

$$T_{j,k}(y) = \frac{\partial \phi}{\partial \xi}(y, \nu_{j,k}).$$

Furthermore, let $y_i = T_{j,k}^{-1}(x_i)$ and $\tilde{T}_i$ be the frequency part given by

$$\tilde{T}_i(\xi) = \frac{\partial \phi}{\partial y}(y_i, \xi).$$

Let $\lambda_{i,l}^{j,k} = \tilde{T}_i(\xi_l^{j,k})$ be the image of the frequency $\xi_l^{j,k}$ under $\tilde{T}_i(\xi_l^{j,k})$. The digital microcanonical relation *associated with the operator $\mathcal{P}$ is given by the map*

$$\chi^d : (x_i; \xi_l^{j,k}) \to (y_i; \lambda_{i,l}^{j,k}). \tag{6.2.36}$$

This completes our digitization of the Fourier integral operator $\mathcal{P}$ and its microcanonical relation χ. This digital microcanonical relation acts as a coordinate transformation on the digital grid of the phase space where the digital wavefront set is defined. With that in mind, we are now able to translate the results on the microlocal behavior of the Radon transform and the residual convolutional neural networks presented in Chapters 2 and 4, from the continuous case to the digital case.

6.3 Digital microlocal analysis of conv-ResNets and the learned primal-dual

In this section, we explore the microlocal behavior of the convolutional residual neural networks (Definition 4.1.1) and the learned primal-dual architecture (Algorithm 5) in the context of the digital microcanonical relation of Definition 6.2.15. We study this particular architecture since we aim to apply our results to the learned primal-dual architecture, but our method can be also easily extended to other architectures and Fourier integral operators. We also assume this architecture to be a discretization of the continuum counterparts from Definition 4.1.15 and Algorithm 6.

In the case of the Radon transform and the convolutional operator, being Fourier integral operators, we compute the digital microcanonical relation following the shearlet-based digitization of FIOs (Theorem 6.2.13). In the case of the ReLU activation nonlinearity, and the residual layer, since they are applied pointwise, the microcanonical relation can be directly digitized from the continuous case to the corresponding digital grid.

Following the approach presented in Section 4.1, we study the basic elements of the conv-ResNets, meaning, the convolutional layer, the ReLU activation function, and the residual layer. In addition, for the learned primal-dual architecture we also study the microlocal behavior of the Radon transform.

6.3.1 Convolutional layers

Following Remark 4.1.3, we know that the continuum convolutional operator $\mathcal{K}_{\boldsymbol{\theta}}$ given by (4.1.6) is a pseudodifferential operator. In particular, $\mathcal{K}_{\boldsymbol{\theta}}$ is also a Fourier integral operator with phase $\phi(y,\xi) := 1$. By using the pseudo-local property (Theorem 2.3.10) we have that for $f \in L^2(\Omega)$ the wavefront set $\mathrm{WF}(\mathcal{K}_{\boldsymbol{\theta}}(f))$ follows

$$\mathrm{WF}(\mathcal{K}_{\boldsymbol{\theta}}(f)) \subset \mathrm{WF}(f).$$

Meaning, $\mathcal{K}_{\boldsymbol{\theta}}$ does not introduce new singularities to f. Moreover, if the coefficients $\beta_{n,m}$ are such that the amplitude function $p_{\boldsymbol{\theta}}$ given by (4.1.7) follows

$$0 < |p_{\boldsymbol{\theta}}(\xi)| \quad \text{for all} \quad ||\xi|| \neq 0,$$

then the operator operator $\mathcal{K}_{\boldsymbol{\theta}}$ is elliptic and preserve the singularities, i.e.,

$$\mathrm{WF}(\mathcal{K}_{\boldsymbol{\theta}}(f)) = \mathrm{WF}(f).$$

Following the shearlet approximation (6.2.29) we can digitize $\mathcal{K}_{\boldsymbol{\theta}}$ by the digital operator $\mathcal{K}^d_{\boldsymbol{\theta}} : \ell^2(\Omega^d) \to \ell^2(\Omega^d)$ given by

$$(\mathcal{K}_{\boldsymbol{\theta}} f)(y_i) = (\mathcal{K}^d_{\boldsymbol{\theta}} f^d)(y_i) := \sum_{j,k} \sum_{r=1}^{R_{j,k}} \alpha^{(r)}_{j,k}(y_i) \sum_{l \in \Xi^j} e^{2\pi i \langle x_i, \xi^{j,k}_l \rangle} \hat{f}(\xi^{j,k}_l) \hat{\psi}^2_{j,k}(\xi^{j,k}_l) \hat{\vartheta}^{(r)}_{j,k}(\xi^{j,k}_l)$$

where Ω^d is the digital grids on Ω,$y_i \in \Omega^d$, and $\xi^{j,k}_l$ are the frequency points in the discrete wedge tiling from Definiton 6.2.12. In addition, $f^d \in \ell^2(\Omega^d)$ is the discretization of $f \in L^2(\Omega)$.

In addition, we have that the digital wavefront set $\mathrm{WF}(\mathcal{K}_{\boldsymbol{\theta}}(f^d))$ also satisfies

$$\mathrm{WF}^d(\mathcal{K}^d_{\boldsymbol{\theta}}(f^d)) \subset \mathrm{WF}^d(f^d).$$

Moreover, if he coefficients $\beta_{n,m}$ are such that

$$0 < |p_{\boldsymbol{\theta}}(\xi^{j,k}_l)| \quad \text{for all} \quad ||\xi^{j,k}_l|| \neq 0,$$

then the digital operator $\mathcal{K}^d_{\boldsymbol{\theta}}$ preserve the singularities, i.e.,

$$\mathrm{WF}^d(\mathcal{K}^d_{\boldsymbol{\theta}}(f^d)) = \mathrm{WF}^d(f^d).$$

This implies that the digital microcanonical relation is given by the identity mapping

$$\chi^d : (x_i; \xi^{j,k}_l) \mapsto (x_i; \xi^{j,k}_l) \quad \text{for } i,j,k,l \in \mathbb{Z} \tag{6.3.1}$$

for spatial points in the digital spatial grid x_i and frequency points in the discrete wedge tiling $\xi_l^{j,k}$.

We present the numerical results of this digital microlocal behavior in Section 8.3, presenting how the convolutional layers in the learned primal-dual architecture propagate digital singularities on simulated data. In this section, we also present an implementation of an “ellipticity” measure, which indicates “how elliptic” the particular convolution is by analyzing the values of the $p_{\boldsymbol{\theta}}$ evaluated on the digital frequency grid points. This allows us to find out the singularities that are more likely to be preserved or not.

6.3.2 The ReLU activation function

Let $\mathsf{ReLU}_{\kappa,\phi_\kappa} : L^2(\Omega) \to L^2(\Omega)$ given by (4.1.10) be the ReLU operator in the continuum setting. Following the strategy of Section 4.2.4 we chose a fixed $\kappa > 0$ and $\psi_\kappa \in \mathcal{S}(\mathbb{R}^2)$ that integrates to 1, and let us rename $\mathsf{ReLU} := \mathsf{ReLU}_{\kappa,\phi_\kappa}$ for such fixed parameter .

Since in this case the ReLU operator acts pointwise on the input function $f \in L^2(\Omega)$, then the digitization becomes simpler, since it also acts pointwise on the points of the digital grid. The digital $\mathsf{ReLU}^d : \ell^2(\Omega^d) \to \ell^2(\Omega^d)$ operator is defined then as

$$(\mathsf{ReLU}^d f^d)(x_i) := (\mathsf{ReLU}\, f)(x_i)$$

for x_i points of the digital spatial grid Ω^d (Definition 6.2.12) and $f^d \ell^2(\Omega^d)$ is the digitization of f. Since the digitization is defined pointwise, we can compute the digital wavefront set of $\mathrm{WF}^d(\mathsf{ReLU}^d(f^d))$ from the known $\mathrm{WF}^d(f^d)$ by rewritting Algorithm 1 in the digital grid.

Algorithm 8: Digital wavefront set classifier of $\mathsf{ReLU}^d(f^d)$.

```
Input: Digital image $f^d \in \ell^2(\Omega^d)$, $\mathrm{WF}^d(f^d)$, $x_i \in \Omega^d$.
Output: Estimate $\mathrm{WF}^d(\mathsf{ReLU}^d(f))_{x_i} \subset \Omega^d$.
initialisation;
if $x_i \in \mathrm{supp}_+(f^d)^{\mathrm{o}}$ then
    return $\Lambda_x^d = \mathrm{WF}(f^d)_{x_i}$;
end
if $x_i \in R_{f^d}$ then
    return $\Lambda_{x_i}^d = \{\pm\nabla_{x_i}^d(f^d)/\|\nabla_{x_i}^d(f^d)\|\}$;
end
if $x_i \in C_{f^d} \cup S_{f^d}$ then
    return $\Lambda_{x_i}^d = \Omega^d$;
end
```

In Algoritm 8, ∇^d is the digital gradient operator given by finite differences and $\mathrm{supp}_+(f^d)^{\mathrm{o}}$ is the projection of $\mathrm{supp}_+(f)^{\mathrm{o}}$ to the digital grid Ω^d. In addition, the set R_{f^d}, C_{f^d} and S_{f^d} are the projection of the corresponding sets in Proposition 4.2.2 to the digital grid Ω^d. The final numerical results are depicted in Section 8.3.

6.3.3 The residual layer

Similar to the digital ReLU, since the continuum residual operator $+ : L^2(\Omega) \times L^2(\Omega) \to L^2(\Omega)$ is also applied pointwise, we can simply digitize it with the operator $+^d : \ell^2(\Omega^d) \times \ell^2(\Omega^d) \to \ell^2(\Omega^d)$, which is defined then as

$$(f^d +^d g^d)(x_i) := (f+g)(x_i)$$

for x_i points of the digital spatial grid Ω^d (Definition 6.2.12) and $f^d, g^d \in \ell^2(\Omega^d)$ the digitization of f and g.

We can also digitze the computation of the digital wavefront set $\mathrm{WF}^d(\mathcal{P}^d f^d)$ by rewriting Algorithm 2 in the digital grid.

Algorithm 9: Digital wavefront set classifier of $f^d + g^d$.

```
Input: Digital images $f^d, g^d \in \ell^2(\Omega^d)$, $\mathrm{WF}^d(f^d), \mathrm{WF}^d(g^d)$, $x_i \in \Omega^d$.
Output: Estimate $\mathrm{WF}^d(f^d +^d g^d)_{x_i} \subset \Omega^d$.
initialisation;
if $x_i \in \mathrm{WF}^d(f^d) \cap \mathrm{WF}^d(g^d)^c$ then
    return $\Lambda^d_{x_i} = \mathrm{WF}^d(f^d)_{x_i}$;
end
if $x_i \in \mathrm{WF}^d(f^d)^c \cap \mathrm{WF}^d(g^d)$ then
    return $\Lambda^d_{x_i} = \mathrm{WF}^d(g^d)_{x_i}$;
end
if $x_i \in \mathrm{WF}^d(f^d) \cap \mathrm{WF}^d(g^d)$ then
    return $\Lambda^d_{x_i} = \mathrm{WF}^d(g^d)_{x_i} \cup \mathrm{WF}^d(f^d)_{x_i}$;
end
```

6.3.4 The Radon transform

In Section 2.5 we explored the microlocal analysis of the Radon transform in the continuous setting $\mathcal{R} : L^2(\Omega) \to L^2(\Xi)$ given by

$$\mathcal{R}f(s,\varphi) = \frac{1}{2\pi} \int_{\xi \in \mathbb{R}} \int_{x \in \mathbb{R}^2} e^{i(s-(x\cdot\omega(\varphi)))\xi} f(x) dx d\xi. \tag{6.3.2}$$

Although (6.3.2) is the classical form to study the Radon transform, in order to digitize its canonical relation, following Section 6.2.4 and 6.2.5, we need to rewrite it in the oscillatory integral form (6.1.1). The next proposition takes care of this.

Proposition 6.3.1. *Let $\mathcal{R} : L^2(\Omega) \to L^2(\Xi)$ be the Radon transform given by* (6.3.2)*, then it can also be written in the oscillatory integral form* (6.1.1) *given by*

$$(\mathcal{R}f)(s,\varphi) = \frac{1}{(2\pi)^2} \int_{\xi \in \mathbb{R}^2} e^{i(\omega(\varphi)\cdot\xi)s} \hat{f}(\xi) d\xi \quad \textit{for all } (s,\varphi) \in \Xi \textit{ and } f \in L^2(\Omega). \tag{6.3.3}$$

Proof. Following [116], in general, a Fourier integral operator can be written as

$$(\mathcal{P}f)(y) = \int_{x\in\mathbb{R}^2} A(y,x)f(x)dx, \tag{6.3.4}$$

where the kernel $A(y,x)$ admits an oscillatory integral representation

$$A(y,x) = \int_{\xi\in\mathbb{R}^2} p(y,\xi)e^{i\phi(y,x,\xi)}d\xi \tag{6.3.5}$$

with the non-degenerate phase function,

$$\phi(y,x,\xi) = x\cdot\xi - S(y,\xi), \tag{6.3.6}$$

and amplitude $p = p(y,\xi)$ an standard symbols of order 0, with principle part homogeneous in ξ of order 0. By applying (6.3.6) to (6.3.4)-(6.3.5), we obtain the original representation

$$(\mathcal{P}f)(y) = \int_{\xi\in\mathbb{R}^2} e^{iS(y,\xi)}\hat{f}(\xi)d\xi.$$

Next, let us first notice that we can rewrite (6.3.2) as

$$(\mathcal{R}f)(s,\varphi) = \int_{x\in\mathbb{R}^2} f(x)\delta(s - x\cdot\omega(\varphi))dx,$$

where δ is the Dirac delta distribution and $\omega(\varphi) = (\cos\varphi, \sin\varphi)$. Following the integral form in (6.3.4), we get that $\mathcal{R}$ is represented by the kernel

$$A((s,\varphi),x) = \delta(s - x\cdot\omega(\varphi)).$$

Taking the Fourier transform with respect to x of A we get

$$\begin{aligned}\mathcal{F}_x(A((s,\varphi),x))(\xi) &= \frac{1}{2\pi}\int_{x\in\mathbb{R}^2} e^{-ix\cdot\xi}\delta(s - x\cdot\omega(\varphi))dx \\ &= \frac{1}{2\pi}e^{-i(\xi\cdot\omega(\varphi))s}.\end{aligned} \tag{6.3.7}$$

Now, computing the inverse Fourier transform with respect to ξ of (6.3.7) we get

$$\begin{aligned}A((s,\varphi),x) &= \mathcal{F}_x^{-1}\left(\mathcal{F}_x(A((s,\varphi),x))\right) \\ &= \frac{1}{2\pi}\int_{\xi\in\mathbb{R}^2} e^{ix\cdot\xi}\mathcal{F}_x(A((s,\varphi),x))(\xi)d\xi \\ &= \frac{1}{(2\pi)^2}\int_{\xi\in\mathbb{R}^2} e^{ix\cdot\xi}e^{-i(\xi\cdot\omega(\varphi))s}d\xi \\ &= \int_{\xi\in\mathbb{R}^2} p((s,\varphi),\xi)e^{i\phi((s,\varphi),x,\xi)}d\xi,\end{aligned} \tag{6.3.8}$$

where the amplitude is $p((s,\varphi),\xi) = \frac{1}{(2\pi)^2}$ and the phase is given by

$$\phi((s,\varphi),x,\xi) = (\xi\cdot\omega(\varphi))s - x\cdot\xi.$$

Therefore, the Radon transform $\mathcal{R}$ can be written in the oscillatory integral form as

$$(\mathcal{R}f)(s,\varphi) = \frac{1}{(2\pi)^2}\int_{\xi\in\mathbb{R}^2} e^{i(\omega(\varphi)\cdot\xi)s}\hat{f}(\xi)d\xi.$$

□

Having the integral oscillatory form (6.3.3) allows us to use Theorem 6.2.13 to digitize the Radon transform $\mathcal{R}$. We can also use Remark 6.1.1 to obtain the microcanonical relation mapping of the Radon transform as well as its digital counterpart.

Proposition 6.3.2. *Let $\mathcal{R}: L^2(\Omega)\to L^2(\Xi)$ be the Radon transform given by* (6.3.3). *Let $\mathcal{SH}_\psi = \{\psi_\mu\}_{\mu\in\mathcal{M}}$ be a discrete shearlet system defined by* (6.1.10). *Moreover, let $\{y_i\}_{i\in\mathcal{I}} = \{(s_i,\varphi_i)\}_{i\in\mathcal{I}} \subset L^2(\Xi)$ and $\{\xi_l^{j,k}\}_{(j,k,l)\in\mathcal{S}} \subset \mathbb{R}^2\setminus\{0\} \subset \mathbb{R}^2\setminus\{0\}$ be digital grids as in Proposition 6.2.13. Then, the digital Radon transform $\mathcal{R}^d$ is given by*

$$(\mathcal{R}f)(y_i) = (\mathcal{R}^d f^d)(y_i) := \sum_{j,k}\sum_{r=1}^{R_{j,k}} \alpha_{j,k}^{(r)}(y_i)\sum_{l\in\Xi^j} e^{2\pi i\langle x_i,\xi_l^{j,k}\rangle}\hat{f}(\xi_l^{j,k})\hat{\psi}_{j,k}^2(\xi_l^{j,k})\hat{\vartheta}_{j,k}^{(r)}(\xi_l^{j,k}), \tag{6.3.9}$$

where $\hat{\theta}_{j,k}^{(r)}$ and $\alpha_{j,k}^{(r)}$ are given by (6.2.12) *and $\psi_{j,k}$ is the shearlet filter for the scale j and shearing k. In addition, the spatial component of the digital microcanonical relation of $\mathcal{R}^d$, $T_{j,k}:\Xi\to\Omega$, is given by*

$$T_{j,k}(y_i) = T_{j,k}((s_i,\varphi_i)) = \frac{\partial S}{\partial\xi}(y_i,\nu_{j,k}) = \omega(\varphi_i)s_i =: x_i \quad \textit{for all } i\in\mathcal{I}. \tag{6.3.10}$$

Also, the frequency component of the digital microcanonical relation *of $\mathcal{R}^d$, $\tilde{T}_i:\mathbb{R}^2\setminus\{0\}\to\mathbb{R}^2\setminus\{0\}$ is given by*

$$\lambda_{i,l}^{j,k} := \tilde{T}_i(\xi_l^{j,k}) = \frac{\partial S}{\partial y}(y_i,\xi) = \frac{\partial S}{\partial(s,\varphi)}((s_i,\varphi_i),\xi) = (\omega(\varphi_i)\cdot\xi,(\omega^{\perp}(\varphi_i)\cdot\xi)s_i). \tag{6.3.11}$$

Given (6.3.10) *and* (6.3.10) *the* digital microcanonical transform mapping *of $\mathcal{R}^d$, χ^d is defined by*

$$\chi^d(x_i;\xi_l^{j,k}) := (T_{j,k}^{-1}(x_i);\tilde{T}_i(\xi_l^{j,k}) = (y_i,\lambda_{l,i}^{j,k}). \tag{6.3.12}$$

Proof. We know by Remark 6.1.1 that the microcanonical mapping of the Radon transform, $\chi:\Omega\times\mathbb{R}^2\setminus\{0\}\to\Xi\times\mathbb{R}^2\setminus\{0\}$, is given by

$$\chi:\left(\frac{\partial S}{\partial\xi};\xi\right)\mapsto\left(y;\frac{\partial S}{\partial y}\right),$$

where $S(y,\xi) = S((s,\varphi),\xi) = (\omega(\varphi)\cdot\xi)s$ (see (6.3.3)) and $y = (s,\varphi)$. Then

$$\begin{aligned}\partial_{(s,\varphi)}S((s,\varphi),\xi) &= (\omega(\varphi)\cdot\xi, (\omega^{\perp}(\varphi)\cdot\xi)s),\\ \partial_{\xi}S((s,\varphi),\xi) &= \omega(\varphi)s.\end{aligned}$$

With the oscillatory matrix form of $\mathcal{R}$ in (6.3.3) we are able to digitize the operator and its microcanonical relation χ with shearlets in the fashion of Section 6.2. Namely, given $\{y_i\}_{i\in\mathcal{I}} = \{(s_i,\varphi_i)\}_{i\in\mathcal{I}} \subset L^2(\Xi)$ and $\{\xi_l^{j,k}\}_{(j,k,l)\in\mathcal{S}} \subset \mathbb{R}^2\setminus\{0\} \subset \mathbb{R}^2\setminus\{0\}$ the digital grids from Proposition 6.2.13, the digital Radon transform $\mathcal{R}^d$ is given by (6.2.29)

$$(\mathcal{R}f)(y_i) \approx (\mathcal{R}^d f)(y_i) := \sum_{j,k}\sum_{r=1}^{R_{j,k}} \alpha_{j,k}^{(r)}(y_i) \sum_{l\in\Xi^j} e^{2\pi i\langle x_i,\xi_l^{j,k}\rangle}\hat{f}(\xi_l^{j,k})\hat{\psi}_{j,k}^2(\xi_l^{j,k})\hat{\vartheta}_{j,k}^{(r)}(\xi_l^{j,k}),$$

where $\hat{\theta}_{j,k}^{(r)}$ and $\alpha_{j,k}^{(r)}$ are given by (6.2.12) and $\psi_{j,k}$ is the shearlet filter for the scale j and shearing k. In this case, by construction $y_i = T_{j,k}^{-1}(x_i)$ (see Proposition 6.2.13), where $T_{j,k}$ is the spatial component of the digital microcanonical relation of $\mathcal{R}^d$, given by

$$T_{j,k}(y_i) = T_{j,k}((s_i,\varphi_i)) = \frac{\partial S}{\partial\xi}(y_i,\nu_{j,k}) = \omega(\varphi_i)s_i =: x_i \quad \text{for all } i\in\mathcal{I}.$$

where $\nu_{j,k} = (\cos\theta_{j,k}, \sin\theta_{j,k})$, $\theta_{j,k} = \arctan(-k/2^{j-1})$.

Finally, following (6.2.36) we have that $\lambda_{i,l}^{j,k} = \tilde{T}_i(\xi_l^{j,k})$ with $\tilde{T}_i$ is the *frequency part of the digital microcanonical relation* of $\mathcal{R}^d$ given by

$$\tilde{T}_i = \frac{\partial S}{\partial y}(y_i,\xi) = \frac{\partial S}{\partial(s,\varphi)}((s_i,\varphi_i),\xi) = (\omega(\varphi_i)\cdot\xi, (\omega^{\perp}(\varphi_i)\cdot\xi)s_i).$$

This implies that the *digital microcanonical transform mapping* of $\mathcal{R}^d$, χ^d is given by

$$\chi^d(x_i;\xi_l^{j,k}) := (T_{j,k}^{-1}(x_i); \tilde{T}_i(\xi_l^{j,k}) = (y_i, \lambda_{l,i}^{j,k}).$$

□

Notice that the same procedure can be done for the back-projection operator $\mathcal{R}^*$ given by (2.3.2), whose digital microcanonical relation is the inverse of the digital microcanonical relation of $\mathcal{R}^d$ given by (6.3.12).

The digital microcanonical relation of $\mathcal{R}^d$ can be used to compute the digital wavefront set of the sinogram $\mathrm{WF}^d(\mathcal{R}^d f)$ by the application of χ^d to the digital wavefront set of the image $\mathrm{WF}^d(f)$. These results also hold for the adjoint Radon transform $\mathcal{R}^*$ where the microcanonical relation in the continuous setting χ is the inverse mapping of the microcanonical relation of $\mathcal{R}$. We have used this approach to compute the microcanonical relation for the digital Radon transform to analyze the digital microlocal behavior of the learned primal-dual. The results are presented in Section 8.3.

Remark 6.3.3. *The digital microcanonical relation of the convolutional ResNet (Definition 4.1.1) and the learned primal-dual algorithm (Algorithm 5) is the iterative application of the digital microcanonical relation of each of their components. In particular, the digital microcanonical relation of the convolutional operator in* (6.3.1)*, the* ReLU *non-linearity in Algorithm 8, the residual layer in Algorithm 9, the Radon transform and its adjoint in 6.3.12.*

In the next chapter, we will explore a particular application of this analysis, namely, the task-adapted reconstruction. In this application, we use the digital wavefront set propagation under the action of the learned primal-dual architecture as a strong prior to improve tomographic reconstruction in the low-dose and limited angle cases.

7 Application of digital wavefront sets to task-adapted reconstruction

In this chapter, we explore a particular application of task-adapted reconstruction using digital wavefront sets and their propagation via the learned primal-dual architecture. This chapter will also be an opportunity to analyze the impact of microlocal analysis on inverse problems and formalize the notion of wavefront set extraction in the context of statistical decision theory. We will also present a task of particular interest, namely, *wavefront set inpainting*. The goal here is to recover oriented singularities on the unknown part of the low-dose and limited-angle tomography (the invisible part, see [17]) from the known part (the visible part) using deep neural networks. Finally, in this chapter, we adapt our approach of digital microlocal analysis of the learned primal-dual architecture to the task-adatped reconstruction introduced by Adler et al. [1]. The goal is to improve the reconstruction procedure using a-priori information provided by the digital microcanonical relation.

My own contribution: As in Chapter 6 this chapter results from numerous discussions with my supervisor, Gitta Kutyniok, and my collaborators Ozan Öktem and Philipp Petersen which was later published as [9]. The main ideas in this case were mine inspired by the work of Adler et.al. [1] on task-adapted reconstruction. This work together with Chapter 6 was the final result of my research throughout my PhD studies. The actual writing was mostly done by myself.

7.1 Motivation

As discussed extensively throughout this thesis, inverse problems play a fundamental role in various real-world applications. This is due to the fact that most problems in physics and medicine which involve measurements of signal parameters (e.g. X-ray tomography, seismic imaging, MRI) can be stated as an inverse problem. As we know, in an inverse problem, we aim to recover (reconstruct) model parameters characterizing a system under investigation from measurements. Formally, we can formulate an inverse problem as follows.

Definition 7.1.1 (Classical inverse problems)**.** *Let X and Y be separable Banach spaces, known as* model parameter space *and* data space *respectively. In addition, let $A : X \to Y$ be a* forward operator*. An* inverse problem *aims to recover a ground truth unknown signal $f^* \in X$ from noisy measurements described by the equation*

$$g = \mathcal{A}(f^*) + \delta g, \tag{7.1.1}$$

where $\delta g \in Y$ is the noise, defined as a single sample of a $Y-$valued random variable. In this setting, the forward operator *$\mathcal{A}$ models how the data is measured from signal in the absence of noise.*

The most prominent example of an inverse problem for this thesis is the problem of *tomographic reconstruction*, which involves the Radon transform (Definition 2.3.1) as the forward operator.

Definition 7.1.2. *Let $\Omega \subset \mathbb{R}^2$ and $\Xi \subset \mathbb{R} \times (0, \pi)$ be open. The* tomographic reconstruction problem *aims to recover an unknown ground truth image $f^* \in L^2(\Omega)$ from noisy measurements $g \in L^2(\Xi)$ modeled by*

$$g = \mathcal{R}(f^*) + \delta g,$$

where $\mathcal{R} : L^2(\Omega) \to L^2(\Xi)$ is the Radon transform (Definition 2.3.1) and δg is the noise in the data space $L^2(\Xi)$.

As discussed in Section 1.3, there are different approaches to solving inverse problems. Reconstruction methods for ill-posed inverse problems can be classified in three groups:

- **Model-based methods:** use first principles to incorporate a-priori information and find solutions adequate to the physical problem (e.g. iterative methods [25], variational regularization[112]).
- **Data-driven methods:** use parameterized models to learn solutions from training data (e.g. deep convolutional neural networks [13]).
- **Hybrid methods:** incorporate model-based first principles to parametrize models, in order to find the solution from training data, and reduce the parameter space by the use of a-priori information (e.g. learned primal-dual reconstruction [3]).

In Section 4.3 we discussed the learned primal-dual architecture, a hybrid method originally introduced by Adler and Öktem [3]. Such a method makes use of a primal-dual iterative scheme to design a deep convolutional neural network capable to perform image reconstruction in imaging inverse problems. This method uses as a-priori information the forward operator and its adjoint. This allows convolutional residual neural networks to learn the local properties of the solution, while the operators handle the global properties. Moreover, in Sections 1.1 and 3.1 we have discussed the important role that microlocal analysis and wavefront sets play in inverse problems. Since oriented edges contain a large amount of information on an image, they can be used as a-priori in regularization methods. In addition, microlocal analysis allows us to understand how oriented singularities are propagated under Fourier integral operators, via the microcanonical relation. A significant number of inverse problems in real-world applications are modeled by forward operators that are in fact, Fourier integral, such as the Radon transform [70]. Based on this fact, the microcanonical relation allows us to obtain a subset of the wavefront set of the solution f^* using the wavefront set of the data g without the need for any reconstruction.

In [17] Bubba et al. introduced a method for limited-angle tomographic reconstruction that uses the microcanonical relation to extract the singularities that can be faithfully reconstructed from the data. Such singularities are obtained by propagating the wavefront set of the measured sinogram through the adjoint of the radon transform, via the microcanonical relation. These singularities, also known as the visible part, are then used to reconstruct the rest of the singularities (the invisible part). Bubba et al. used a deep neural network architecture to find the invisible part and the shearlet transform to perform orientation separation in the reconstructed images. In Section 4.3 we have presented a microlocal analysis of the learned primal-dual architecture in the continuous setting, where we are able to describe the propagation of singularities from the input data through the architecture. This allowed us to obtain singularities of the output reconstruction. In addition, Chapter 6 introduced the digital counterpart of the microlocal analysis discussed in Chapter 4, using the concept of digital wavefront set introduced in Chaptar 5 in addition to a discretization based on shearlets.

Being able to describe singularities and their propagation in deep neural networks for inverse problems is a powerful tool, in particular, it allows us to use the wavefront sets as a-priori information for the reconstruction. In that sense, we would like to develop a reconstruction method that is able to well approximate the ground truth wavefront set. An appropriate framework for this is the framework of *task-adapted reconstruction*. In the next sections, we will present a novel method that jointly performs tomographic reconstruction and adapts it so that its wavefront set is close to the wavefront set of the ground truth, the last being the task of interest. We will present two different tasks, namely wavefront set reconstruction and wavefront set inpainting. For this purpose, we need to formulate the tomographic reconstruction and the tasks in the same statistical framework –the statistical decision theory–. This is the purpose of the next two sections.

7.2 Inverse problems as statistical estimations

In this section, we study the statistical setting of inverse problems, also known as statistical inverse problems. This is not a new contribution, since it has been extensively discussed in the past, for example in [40]. The approach presented in this section, and throughout most of this chapter, is highly based on the theory presented by Adler et al. in [1]. For this purpose, we are going to revisit shortly the basic concepts of probability theory also presented in Section 5.5.

Let us assume we have an inverse problem described by the forward operator $\mathcal{A} : X \to Y$ as in (7.1.1). In [40] Evans et al. introduced the notion of a *statistical inverse problem*, which reads as follows.

Definition 7.2.1. *Let X and Y be separable Banach spaces. In addition, let $\mathfrak{S}_X$ and $\mathfrak{S}_Y$ be $\sigma-$algebras of X and Y, respectively (Definition 5.5.1). Thus, $(X, \mathfrak{S}_X)$ and $(Y, \mathfrak{S}_Y)$ are* measurable spaces. *Furthermore, let $\mathscr{P}_X$ and $\mathscr{P}_Y$ be probability measure spaces on X and Y (Definition 5.5.2). A* statistical inverse problem *aims to reconstruct $f^* \in X$*

from measurements $g \in Y$ drawn from a $Y-$valued random variable $\boldsymbol{g}$, i.e.

$$\boldsymbol{g} \sim \mathcal{M}(f^*), \tag{7.2.1}$$

and $\mathcal{M} : X \to \mathscr{P}_Y$ is a known data model.

This kind of reconstruction is also known as *statistical estimation.* Functions in the model parameter space X are potential reconstructions and data in Y are potential measurements. In tomography, being our main interest, functions in X are defined on a fixed domain $\Omega \subset \mathbb{R}^2$, which represent gray-scale images, see [70]. In addition, measurements in Y are real-valued functions defined on a manifold $\mathbb{M}$, also known as *sinograms.* This manifold is given by the sensor geometry related to the measurements. From now on, we will use bold-face notation to denote random variables, for instance, whereas g is an element in Y, $\boldsymbol{g}$ is a $Y-$valued random variable.

Remark 7.2.2. *As in the case of the classical (non-statistical) setting, it is often the case that statistical inverse problems are ill-posed. In particular, they do not have a unique solution. A commonly used data model include noise in the data*

$$\boldsymbol{g} = \mathcal{A}(f^*) + \delta\boldsymbol{g}. \tag{7.2.2}$$

In this case, $\delta\boldsymbol{g} \sim \mathbb{P}_{noise}$ is the noise for some known $\mathbb{P}_{noise} \in \mathscr{P}_Y$ and $\mathcal{A}$ is the classical forward operator from (7.1.1)*. If the noise $\delta\boldsymbol{g}$ does not depend on the ground-truth f^*, then the inverse problem* (7.2.2) *has the data model*

$$\mathcal{M}(f) = \mathbb{P}_{noise}(\cdot - \mathcal{A}(f)) \quad \forall f \in X.$$

A standard approach to solve the statistical estimation problem (7.2.1) is the so-called *Bayesian inversion.* In this method, in addition to estimating the solution $f^* \in X$, one also aims to take uncertainty into account.

Let us first introduce an $X-$valued random variable $\boldsymbol{f} \sim \pi^*$, where its probability distribution π^* generates f^*. The challenge here is that π^* is assumed to be unknown. Therefore, one can reformulate the inverse problem (7.2.1) as the problem of recovering the probability measure $\pi^* \in \mathscr{P}_X$ while the data $g \in Y$, generated by $\boldsymbol{g}$, is known. In this context $\boldsymbol{g}$ is related to f^* via the data model (7.2.1). In the particular case where we know explicitly how π^* depends on the ground-truth $f^* \in X$, the inverse problem becomes the task of estimating $f^* \in X$.

Remark 7.2.3. *In the Bayesian formulation, we aim to find the posterior distribution of $\boldsymbol{f}$ given $\boldsymbol{g} = g$. The last equality means that the given data g is drawn from the random variable $\boldsymbol{g}$. In such case, following Bayes' theorem, one has that the joint law $(\boldsymbol{f}, \boldsymbol{g}) \sim \mu$ can be written in terms of the conditional probabilities, i.e.*

$$\mu = \pi_0(f^*) \otimes \pi(\boldsymbol{g}|\boldsymbol{f} = f^*) = \pi_0(f^*) \otimes \mathcal{M}(f^*), \tag{7.2.3}$$

where π_0 is a prior. Equation (7.2.3) *is obtained from the definition of the data model as the conditional distribution of $\boldsymbol{g}$ given $\boldsymbol{f} = f^*$.*

The main contribution of the Bayesian setting is the possibility to explore the posterior distribution of $\boldsymbol{f}$ given $\boldsymbol{g} = g$. This can be done when both the prior $x \mapsto \pi_0 \in \mathscr{P}_X$ and the data model $f \mapsto \mathcal{M}(f)$ are known, but the reconstruction $f^* \in X$ is unknown. In addition we also assume that we can define a density $\mathcal{L}$ associated to the data model. Such density is also known as the *data-likelihood*. Generally, the data-likelihood $\mathcal{L}$ is known up to sufficient degree of accuracy, where $d\mathcal{M}(f)(g) = \mathcal{L}(g|f)dy$.

So far, we have introduced the setting for statistical inverse problems based on the approach followed in [1]. In this setting the statistical model $((Y, \mathfrak{S}_Y), \{\mathcal{M}(f)\}_{f \in X})$ is parametrized by the model parameter space X, where $(Y, \mathfrak{S}_Y)$ is a measurable space from Definition 7.2.1. In such context, a reconstruction method can be represented by a mapping $\mathcal{A}^\dagger : Y \to X$, also referred to as *point estimator*. We refer to [60] for a more detailed exploration of statistical estimation theory. Although generally in an inverse problem we aim to compute a reconstruction, the final goal in real-world application is to take a decision on the reconstruction. For example, to decide whether an image coming from a tomographic reconstruction depicts a tumor. The area of statistics that formally studies the way decisions are made is known as *statistical decision theory*, which is the focus of the next section.

7.3 Reconstruction as a statistical decision

There is a natural connection between statistical estimation and statistical decision theory. The purpose of statistical estimation (or inference) is to make a conclusion on the true but unknown distribution $\pi^* \in \mathscr{P}_X$ of $\boldsymbol{f}$ after the experiment has been carried out and the observation f is available. Such a conclusion can be interpreted also as a decision taken over the observation f. In the next section we will introduce the basic concepts and notions of statistical decision theory so we can later proceed with the task-adapted reconstruction framework.

7.3.1 Statistical decision theory

Statistical decision theory is the area of statistics that study how to make *conclusions* based on data. In this context, we choose a separable Banach space D known as the *decision space*. As in the case of the model parameters and data spaces, we assume that D is equipped with a $\sigma-$algebra $\mathfrak{S}_D$. The simplest case of a decision-making procedure is to choose a point $d \in D$ after $f \in X$ has been observed. This decision-making procedure is known as *non-randomized decision*. Formally, a non-randomized decision is a measurable mapping $\mathcal{T} : X \to D$, where $\mathcal{T}(f)$ is a decision made after f is observed. We denote the space of measurable mappings from X into D by $\mathbb{D}$. This chapter focuses on *non-randomized decisions*, although for some problems this approach is not sufficient, with the need of the so-called *randomized decisions* [78], and we will also present this notion for sake of completeness. An appropriate mathematical structure for representing randomized decisions is the notion of *stochastic kernel*.

Definition 7.3.1. *Let X and D be a separable Banach spaces and $\mathfrak{S}_D$ be a $\sigma-$algebra of D. A* stochastic kernel *is a mapping $\mathcal{K} : \mathfrak{S}_D \times X \to [0,1]$ which has the following two properties.*

(i) For every fixed $f \in X$, the object $\mathcal{K}(\cdot|f)$ is a probability distribution on the measurable space $(D, \mathfrak{S}_D)$ (Definition 7.2.1).

(ii) For every fixed $\boldsymbol{d} \in \mathfrak{S}_D$, the mapping $f \mapsto \mathcal{K}(\boldsymbol{d}|f)$ from X into $[0,1]$ is measurable.

We refer to every stochastic kernel $\mathcal{K} : \mathfrak{S}_D \times X \to [0,1]$ as a randomized decision *rule or simply a* decision*.*

The interpretation of $\mathcal{K}(\boldsymbol{d}|f)$ on Definition 7.3.1 is that after $f \in X$ has been observed $\mathcal{K}(\boldsymbol{d}|f)$ is the probability of a point in $\boldsymbol{d} \in \mathfrak{S}_D$ to be chosen. Next, non-randomized decision rules are a particular case of the randomized decision rules, where the stochastic kernel is defined by the Dirac delta $\mathcal{K}(\cdot|f) = \delta_{\mathcal{T}(f)}(\cdot)$ for every decision operator $\mathcal{T} : X \to D$. This gives us two representations of a non-randomized decision rules, given by either $\mathcal{K}$ or $\mathcal{T}$. For convenience, we keep using $\mathcal{T}$ throughout this work. In the following, we introduce some key definitions coming from statistical decision theory.

Definition 7.3.2 ([78, Definition 3.1])**.** *Let $((X, \mathfrak{S}_X), \{\mathbb{P}_z\}_{z\in\Delta})$ be a statistical model where the model parameter space $(X, \mathfrak{S}_X)$ is a measurable space and $\{\mathbb{P}_z\}_{z\in\Delta} \subset \mathscr{P}_X$ is some family of probability measures on X parametrized by the set Δ. Moreover, let $(D, \mathfrak{S}_D)$ be a decision space and $\mathcal{K}$ a stochastic kernel $\mathcal{K} : D \times X \to [0,1]$. The class of all decisions $\mathcal{K}$ is denoted by $\mathbb{D}$. A decision $\mathcal{K}$ is called a* non-randomized decision *if $\mathcal{K}(\cdot|f) = \delta_{\mathcal{T}(f)}$ for some $\mathcal{T} : X \to D$.*

Now, we know that by construction the solutions of statistical inverse problems result in a point estimator. In the context of statistical decision theory, this can be seen as a non-randomized decision for a statistical decision problem. Here the model parameter space X not only parametrizes the underlying statistical model $((Y, \mathfrak{S}_Y), \{\mathcal{M}(x)\}_{x\in X})$ but also at the same time the decision space $(D := X)$.

If the inverse problem is ill-posed, in particular, if it does not have a unique solution, there will be many possible reconstruction methods. Therefore, we need a way to find an optimal decision. In statistical decision theory, the notion of optimal decision comes in the form of the *risk to a decision rule*. In a general sense, the risk measures how good a particular reconstruction method performs. In order to introduce the concept of risk, we need to first discuss the notion of the loss function.

Remark 7.3.3. *The process of first observing the data and then making a decision can also be described by means of a random vector $(\boldsymbol{d}, \boldsymbol{f})$ that is defined on some probability space $((\Omega, \mathfrak{S}_\Omega), \{\mathbb{P}_z\}_{z\in\Delta})$, $z \in \Delta$. Here the random variable $\boldsymbol{f} : \Omega \to X$ is the observation, and $\boldsymbol{d} : \Omega \to D$ is the statistician's action after observing X. From now on, we will focus on the non-randomized decisions case.*

In the non-randomized setting, it is clear that the decision operator $\mathcal{T}$ depends on the outcome of $\boldsymbol{f} = f$, where $\boldsymbol{f}$ is a random variable, and is defined as the decision $\mathcal{K}(\cdot|f)$, $f \in X$. More precisely, $D(\cdot|f)$ is the conditional distribution of $\mathcal{T}$ given $\boldsymbol{f} = f$, and $\mathcal{L}(z, \boldsymbol{f}) := \mathbb{P}_z \circ \boldsymbol{f}^{-1} =: P_z$, $z \in \Delta$, is the marginal distribution of $\boldsymbol{f}$. This means that by the definition of the conditional distribution, for every set $C \in \mathfrak{S}_D \otimes \mathfrak{S}_X$ it holds

$$\begin{aligned} &\mathcal{L}(z, \boldsymbol{d}, \boldsymbol{f}) := \mathbb{P}_z \circ (\boldsymbol{d}, \boldsymbol{f})^{-1}, \\ &(\mathcal{K} \otimes P_z)(C) = \iint \chi_C(d, f) \mathcal{T}(\boldsymbol{d}|f) P_z(df), \quad C \in \mathfrak{S}_D \otimes \mathfrak{S}_X, \end{aligned} \tag{7.3.1}$$

where χ_C is the indicator function over C and $\mathcal{L}$ is the likelihood function. For a more detailed explanation on this, we refer to [78].

Remark 7.3.4. *Following the standard extension technique (see [78, Section 3.2]), we can use a linear combinations of indicator functions and the approximation of nonnegative measurable functions by increasing sequences of such linear combinations [40], obtaining*

$$\mathbb{E}_z h(\boldsymbol{d}, \boldsymbol{f}) = \iint h(d, f) \mathcal{K}(\boldsymbol{d}|f) P_z(df), \tag{7.3.2}$$

for every $h : D \times X \to \mathbb{R}_+$.

In order to introduce the notion of *risk* (or Bayesian risk) we need the definition of a loss function, which measures the quality of the decision. We assume that the loss of a decision is given by some values $L_D(z, d)$, $z \in \Delta$ and $d \in D$. Here $L(z, d)$ is the loss when a decision is made in favor of d and the true parameter is z.

Definition 7.3.5 ([78, Definition 3.2], Decision loss)**.** *A* decision loss *function L_D is a function $L_D : \Delta \times D \to \mathbb{R}$ such that for every fixed $z \in \Delta$ the function $L_D(z, \cdot)$ is measurable and it holds*

$$-\infty < \inf_{d \in D} L_D(z, d), \quad z \in \Delta. \tag{7.3.3}$$

Remark 7.3.6. *The condition in Equation* (7.3.3) *ensures that for any probability measure μ the integral $\int L_D(z, d)\mu(\boldsymbol{d})$ is well defined. If we assume there exists a mapping $\tau : \Delta \to D$ and a distance in D, $\ell_D : D \times D \to \mathbb{R}$ (decision distance), the loss function $L_D : \Delta \times D \to \mathbb{R}$ given by*

$$L_D(z, d) := \ell_D(\tau(z), d) \tag{7.3.4}$$

is a well-defined decision loss function.

From now on we will refer to Δ as the *feature space.* Under a decision $\mathcal{T}$, after $\boldsymbol{f}$ has been observed, the statistician's action in order to take a final decision is the random variable $\boldsymbol{d}$, where the joint distribution of $\boldsymbol{f}$ and $\boldsymbol{d}$ is given by (7.3.1). Thus the loss under $\mathcal{T}$ is a random variable $L_D(\boldsymbol{z}, \boldsymbol{d})$. This allows us to define the risk of a decision as its expected loss.

Definition 7.3.7 ([78, Definition 3.3], Decision risk)**.** *Let $((X,\mathfrak{S}_X),\{\mathbb{P}_z\}_{z\in\Delta})$ be a statistical model and $L_D:\Delta\times D\to\mathbb{R}$ a decision loss function. The* risk function *(or risk) of a decision $\mathcal{T}\in\mathbb{D}$ is given by the expected value*

$$\mathcal{R}_z(\mathcal{T})=\mathbb{E}_z\left[L_D(z,\boldsymbol{d})\right],\quad z\in\Delta, \tag{7.3.5}$$

where $\mathcal{L}((\boldsymbol{d},\boldsymbol{f})|\mathbb{P}_z)=\mathcal{T}\otimes\mathbb{P}_z$, $z\in\Delta$.

With all these ingredients we can define a *statistical decision problem* as the triple $(\mathcal{M},(D,\mathfrak{S}_D),L_D)$, consisting of a statistical model $\mathcal{M}$, a decision space $(D,\mathfrak{S}_D)$, and a decision loss function L_D. An optimal non-randomized decision rule is a decision $\mathcal{T}$ that minimizes the risk $\mathcal{R}_z(\mathcal{T})$ from (7.3.5). Following the motivation of the Bayesian inversion, we would like to be able to account for uncertainty on the decision problem. For this, we can assume that we have a probability distribution $\eta_0\in\mathscr{P}_\Delta$ and a non-randomize decision rule $\mathcal{T}\in\mathbb{D}$. We can then define the *Bayes decision risk* as follows.

Definition 7.3.8 (Bayes decision risk, [40])**.** *Let us assume we that have a statistical decision problem $(\mathcal{M},(D,\mathfrak{S}_D),L_D)$, where the statistical model is parametrized by the model parameter space, $\mathcal{M}=((X,\mathfrak{S}_X),\{\mathbb{P}_z\}_{z\in\Delta})$ with $\{\mathbb{P}_z\}_{z\in\Delta}\subset\mathscr{P}_\Delta$. In addition we assume that the decision loss function $L:\Delta\times D\to\mathbb{R}$ is given by*

$$L_D(z,d)=\ell_D(\tau(z),d),\quad \textit{for every } z\in\Delta \textit{ and } d\in D,$$

where the feature mapping *$\tau:\Delta\to D$ maps features $z\in\Delta$ to decisions $d\in D$, and $\ell_D:D\times D\to\mathbb{R}$ is the* decision distance*. Given a fixed probability measure $\eta_0\in\mathscr{P}_\Delta$, the* Bayes decision risk *associated to the non-randomized decision $\mathcal{T}\in\mathbb{D}$ ($\mathcal{T}:X\to D$) is defined by the expected loss*

$$\mathcal{R}_{\eta_0}(\mathcal{T}):=\mathbb{E}_{\eta\otimes\mathbb{P}_z}\left[\ell_D(\tau(\boldsymbol{z}),\mathcal{T}(\boldsymbol{f}))\right]\quad \textit{where}\quad (\boldsymbol{z},\boldsymbol{f})\sim\eta_0\otimes\mathbb{P}_z. \tag{7.3.6}$$

An optimal Bayes non-randomized decision rule *is one that optimizes* (7.3.6) *and is called* decision operator*.*

As one can observe, the framework of statistical decision theory allows us to connect the decision-making procedures with the statistical estimation involved on inverse problems. This connection will be discussed in detail in the following.

7.3.2 Bayesian inversion as optimal decision rule

The concept of Bayes decision risk can be extended to the framework of Bayesian inversion. Following [1], we can interpret a statistical inversion problem (Definition 7.2.1) as a statistical decision problem (Definition 7.3.8) as follows.

Definition 7.3.9 (Reconstruction Bayes risk,[1])**.** *Let $((Y,\mathfrak{S}_Y),\{\mathcal{M}(f)\}_{f\in X})$ be a statistical model (Definition 7.2.1) and $D:=X$ be a decision space. In addition let $\ell_X: X\times X\to\mathbb{R}$ be the corresponding decision distance, given by the distance function in X.*

Given a prior π_0 *as in* (7.2.3) *and following* (7.3.6)*, we can compute the* reconstruction Bayes risk *of a potential reconstruction operator* $\mathcal{A}^\dagger : Y \to X$ *by*

$$\mathcal{R}_{\pi_0}(\mathcal{A}^\dagger) = \mathbb{E}_{\pi_0 \otimes \mathcal{M}(f)} \left[\ell_X(\boldsymbol{f}, \mathcal{A}^\dagger(\boldsymbol{g})) \right]. \tag{7.3.7}$$

Following the theory presented in Section 7.3.1, similar to the notion of "optimal decision", we have that a "good" reconstruction method (estimator) is given by a reconstruction operator $\mathcal{A}^\dagger$ that minimizes the Bayes risk (7.3.7). As presented in [1], if we are working on finite dimensions and the loss function is given by the L^2-distance, minimizing the Bayes risk is equivalent to estimating the conditional mean.

We recall that in Bayesian inversion, the ground truth function $f^* \in X$ and the measured data $g \in Y$ are generated by random variables, namely $\boldsymbol{f}$ and $\boldsymbol{g}$. In this case, one aims to recover the conditional probability of the model parameter f^* while the data g and the prior π_0 are given. In other words, one aims to recover the posterior distribution. In contrast, in the classical case, an inverse problem is given by the equation (7.1.1). In this equation, just the data g is generated by a random variable $\boldsymbol{g}$, and the data is generated by a random variable, while the model parameters are simply functions in X, also known as *deterministic approach.* This limits the statistical understanding of the problem, resulting in the lack of the possibility for uncertainty quantification. Thus, we will not be able to compute relevant statistical parameters like mean or standard deviation, which become handy when assessing our results. Therefore, this approach represents a better analysis in comparison with the deterministic case since the posterior $\boldsymbol{f}$ contains all the possible solutions. In this context, one can obtain different reconstructions from different estimators. Also, as discussed in [1], small changes in the data will result in small changes in the posterior distribution [29, Theorem 16]. This can be interpreted as that this approach stabilizes an ill-posed inverse problem when one chooses a correct prior π_0.

Although the Bayesian approach has a lot of upsides, it also presents two important challenges. The first challenge is the fact that normally the posterior is very hard to be written in closed form. Thus, most of the modern approaches deal with the Bayesian inversion without an explicit form of the posterior. We then need to consider the design of a "good" prior $\pi_0 \in \mathscr{P}_X$. The second challenge is related to the computational feasibility of the exploration for the posterior. In the following, we discuss how to address these two challenges.

A "good" prior π_0 should be able to make use of first principles to capture the relevant a-priori information. It should also lead to Bayesian inference methods with desirable asymptotic properties. These properties are usually regarded as consistency and good contraction rates [40]. As seen in [1], there are examples of handcrafted priors coming from Bayesian non-parametric theory, such as [46, 66, 29]. The main issue with handcrafted priors is that they can only describe a fraction of the a-priori information that is available. This fraction corresponds to the part of the physics that can be written in closed-form. A remarkable example in biomedical imaging comes from the fact that the human body, being the object on the images, is impossible to be analytically encoded as a prior.

In the introduction of this section, we discussed the different approaches in classical inverse problems, namely, model-based, data-driven, and hybrid. In Bayesian inverse problems, the handcrafted prior corresponds to the model-based approaches. Similar to the data-driven and hybrid approaches, one can also consider a prior that is learned from training examples in X. In the following sections, we will present, on the one hand, an example of this approach for computed tomography. On the other hand, the exploration of the posterior, when minimizing the Bayes risk, involves sampling from a high-dimensional probability distribution. As one would expect, this sampling is impossible to be implemented, therefore, one needs to discretize the model parameters in finite-dimensional components. This scenario is computationally exhausting for large-scale problems, such as tomographic reconstruction and almost any imaging problem.

In order to tackle this challenge, researchers have made us of Markov chain Monte Carlo methods techniques (see [29, Section 5] and [2]). Unfortunately, these approaches have two main downsides. Some of them need access to the closed-form handcrafted prior resulting in poor performance. The second challenge is that these approaches are hardly scalable for large-scale inverse problems, involving integration over the whole model parameter space, which as we discussed is computationally unfeasible.

7.3.3 Learned iterative methods for Bayesian inversion

In Section 1.1 we discussed the particular case of data-driven/hybrid approaches known as learned iterative methods. These methods combine techniques coming from classical iterative reconstruction approaches with deep neural networks. They can overcome the challenges associated with Bayesian inversion (selecting a good prior and making it computationally feasible). The idea behind the effectiveness of learned iterative methods is the flexibility of highly parameterized non-linear models which can be adapted at hand to specific loss criteria, like (7.3.7), when trained against data.

In particular, learned iterative methods use a deep neural network (Section 1.2) to define a set of reconstruction methods, parametrized by weights. In the context of Bayesian inversion, the training of learned iterative methods against data results on an estimator. This estimator minimizes the Bayes risk and at the same time takes into account a-prior information about how data is generated.

Remark 7.3.10. *In order to illustrate the last statement, let us consider the joint law* $\mu = \pi_0 \otimes \mathcal{M}(x)$ *in* (7.3.7) *which is used to define the Bayes risk. Although in most of the times, this joint law is not explicitly known, one may have access to the corresponding empirical measure defined by the training data* $\{(x_i, y_i)\}_{i=1}^m \subset X \times Y$ *generated by* $(\boldsymbol{f}, \boldsymbol{g}) \sim \mu$*. Therefore, it will not be necessary to introduce any handcrafted (model-based) prior* $\pi_0 \in \mathscr{P}_X$*.*

As we know, the search over all non-randomized decisions is computationally exhaustive. By using a learned iterative method, we restrict our attention to the decisions given by a deep neural network architecture, which is computationally feasible and has large capacity. In this setting, we have a family of reconstruction operators $\mathcal{A}_\theta^\dagger : Y \to X$ parametrized by a parameter set Θ given by weights of a deep neural network, which are also finite-dimensional. The optimal operator is given by training the network against data, i.e.,

$$\theta^* \in \underset{\theta\in\Theta}{\arg\min} \left\{ \frac{1}{m} \sum_{i=1}^{m} \ell_X(f_i, \mathcal{A}_\theta^\dagger(g_i)) \right\}. \tag{7.3.8}$$

Since in this approach, neither the prior nor the data model are handcrafted, we can regard it as being fully data-driven. All this information is learned from the training data, without the knowledge of how the data is generated (data model).

The absence of knowledge on the data generation is an issue when the number of independent samples in the training data set are low compared to the number of weights. This happens frequently in imaging problems due to the high dimension of the model parameter space-. Since in many inverse problems the data model $x \mapsto \mathcal{M}(x)$ which describes the data generation is known, one could use this information to alleviate the above-mentioned issue. In this thesis we will follow this principle by using the learned primal-dual architecture (Section 4.3) to parametrize the reconstruction $\mathcal{A}_{\theta^*}$ such as the learned primal-dual architecture does. Such architecture incorporates the information provided by the data model, by using the forward operator $\mathcal{A}$ (and its adjoint $\mathcal{A}^*$) as a layer. We would like to use this method to perform tomographic reconstruction, but at the same time, we would also like to perform a task (decision) on the model parameters. The next section introduces the notion of task-adapted reconstruction based on statistical decision theory.

7.4 Task-adapted reconstruction

As we know, the inverse problem of reconstructing the model parameters from data is typically only one step in real-world applications. In these applications, the final aim is to use the recovered model parameters for decision-making, also known as a task. The reconstructed model parameters are often analyzed by an expert or an algorithm. Such analysis results in task-dependent features that are finally used in decision-making. Based on [1] Figure 7.1 depicts the typical pipeline of real-world problems involving inverse problem reconstruction and decision-making.

Although performing the various parts of the pipeline in Figure 7.1 independently could seem feasible, this approach has several issues. In particular, each individual step introduces approximation errors and uncertainties, which are not taken into account by the subsequent steps. For example, the reconstruction may not consider the final task. Therefore, we need to adapt the reconstruction method for the specific task. Task-adapted reconstruction, introduced by Adler et al. [1], refers to methods that integrate the reconstruction procedure with the decision-making procedure associated with the task. For that, both reconstruction and task need to be compatible, in the sense that they need to be described by the same theoretical framework. In this case, that framework is statistical decision theory.

The introduction of this framework coincides with the recent effort to incorporate signal processing steps associated with the performance of a task into the reconstruction.

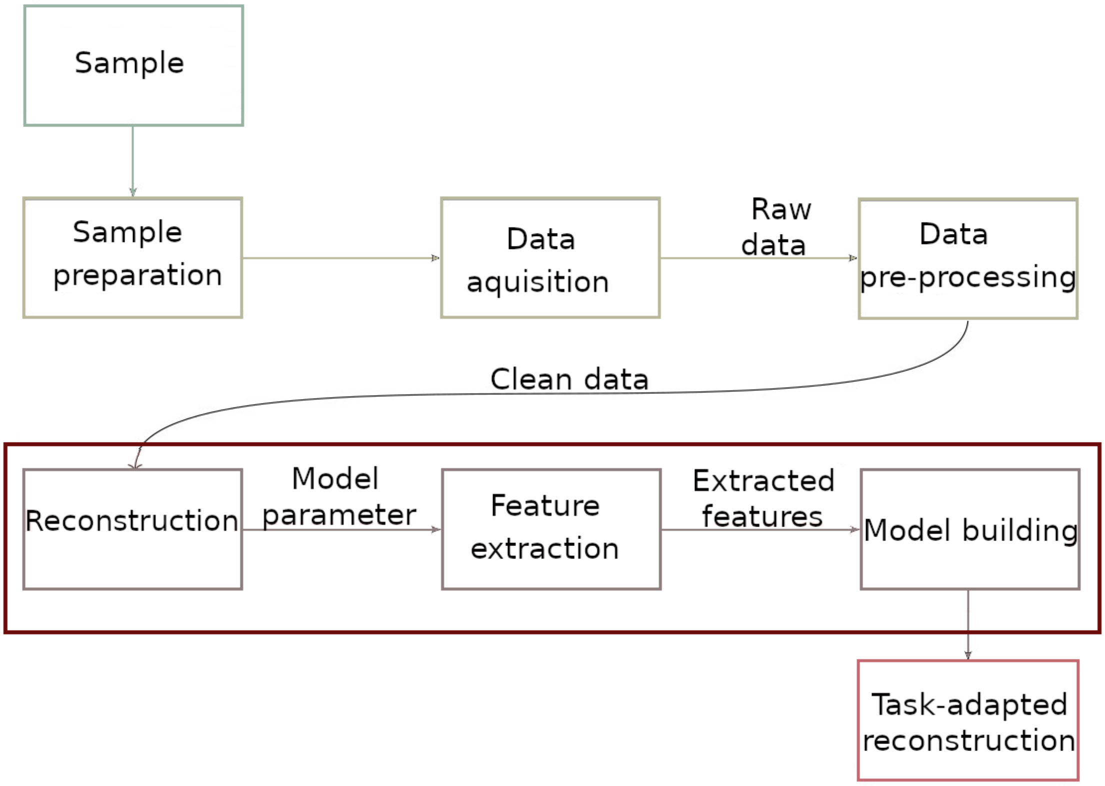

Figure 7.1: Pipeline of decision-making procedure based on inverse problem reconstruction and decision-making.

In computed tomography, the most common tasks correspond to feature extraction. A relevant example of this situation is the extraction of features coming from sparse representations, often done using compressed sensing, for example, the computation of shearlet coefficients for image inpainting [45].

The classical approach to task-adapted reconstruction is similar to the classical approach to inverse problems (Definition 7.1.1), in the sense that it is achieved by solving an operator equation (see (7.2.2)). We introduce a formal definition in the following.

Definition 7.4.1. *Let X, Y, D be separable Banach spaces, refer to as* model parameter*,* data *and* decision space*, respectively. A* task-adapted reconstruction problem *in the classical setting aims to recover a ground truth unknown decision $d^* \in D$ from data $g \in Y$ given the operator equation*

$$g = \mathcal{A}(f^*) + \delta g \quad \textit{and} \quad d^* = \mathcal{T}(f^*), \tag{7.4.1}$$

where $\mathcal{A} : X \to Y$ is the forward operator associated with the inverse problemd, the task operator $\mathcal{T} : X \to D$ represents the decision extraction, both known. As in the classical inverse problem, we have $\delta\boldsymbol{g} \sim \mathbb{P}_{noise}$ for some known $\mathbb{P}_{noise} \in \mathscr{P}_Y$ is the noise.

An illustrative way to represent the task-adapted reconstruction problem from Definition 1.4.1 is shown in Diagram (7.4.2)

$$\begin{tikzcd}
X \arrow[r, "\mathcal{A}"] \arrow[d, "\mathcal{T}"] & Y \arrow[dl, dashed, "?"] \\
D &
\end{tikzcd} \tag{7.4.2}$$

Remark 7.4.2. *It is important to notice that the task operator $\mathcal{T} : X \to D$ is typically highly non-injective. Most likely $\mathcal{T}^{-1}$ is not defined and therefore, also $\mathcal{A} \circ \mathcal{T}^{-1} : D \to Y$. Similarly, real-world problems are ill-posed so normally $\mathcal{T} \circ \mathcal{A}^{-1}$ does not exists.*

In computed tomography, there are different examples of interesting tasks such as edge recovery, segmentation, and image registration. In the next sections, we will introduce new tasks that are of particular interest to this thesis. We are also going to introduce what we consider the most relevant application of our analysis in Chapter 6, the wavefront set extraction and wavefront set inpainting.

A suitable approach for solving (7.4.1) highly depends on the specific task, and its practical application is limited by similar issues as in the Bayesian inversion presented in Section 7.3.2. The first issue is the need for an explicit handcrafted task operator $\mathcal{T} : X \to D$. Access to a closed form of the task operator is extremely difficult for most of the tasks in imaging. The second issue is connected to computational feasibility. Evaluating the task operator is computationally exhaustive and involves the fine-tuning of extra parameters. Moreover, most state-of-the-art model-based methods for solving (7.4.1) make use of variational methods. These methods take high computational complexity in large-scale image problems [98].

More recently, many tasks in imaging have been successfully addressed using deep neural networks [36, 8, 121]. Therefore, it is worth to think whether or not such techniques can be used in the context of task-adapted reconstruction. For that, we first need to explore the tasks on model parameters in its natural framework, the statistical decision theory. Being able to implement the task as a deep neural network training allows us to jointly perform the training with the reconstruction step when suitable training data is available.

7.4.1 Tasks on model parameters

A task in the model parameter space X is an optimal non-randomized decision of the statistical decision problem $(\mathcal{M}, (D, \mathfrak{S}_D), L_D)$, where $\mathcal{M}$ is the model parameter space

$$\mathcal{M} = ((X, \mathfrak{S}_X), \{\mathbb{P}_z\}_{z\in\Delta}) \quad \text{with} \quad \{\mathbb{P}_z\}_{z\in\Delta} \subset \mathscr{P}_\Delta,$$

and L_D is the decision loss function given by (7.3.4). The operator $\mathcal{T} : X \to D$ associated with the optimal decision rule is known as the *task operator.* As we know from Equation (7.3.6), the task operator is found by minimizing the Bayes risk defined by the expected value

$$\mathcal{R}_{\eta_0}(\mathcal{T}) = \mathbb{E}_{\eta_0 \otimes \mathbb{P}_z} \left[\ell_D(\tau(\boldsymbol{z}), \mathcal{T}(\boldsymbol{f})) \right],$$

where $(\boldsymbol{z}, \boldsymbol{f}) \sim \eta_0 \otimes \mathbb{P}_z$. We refer to the probability measure $\eta_0 \in \mathscr{P}_\Delta$ as the *task prior*. As in the case of the Bayesian inversion, finding a "good" task prior is hard. Also, $\mathcal{P}_z \in \mathscr{P}_X$ is normally unknown. Then, the joint law $\eta := \eta_0 \otimes \mathbb{P}_z$ is considered unknown, unlike the reconstruction case, where the joint law is known, or at least its data likelihood.

Remark 7.4.3. *In order to resolve the lack of knowledge on the joint law, and similar to the Bayesian inversion setting, we can replace the measure $\eta_0 \otimes \mathbb{P}_z$ with the empirical measure given by the training data $\{(z_i, \boldsymbol{f}_i)\}_{i=1}^m \subset \Delta \times X$. In this dataset one has i.i.d. (independent and identically distributed) samples generated by a $(\Delta \times X)-$ valued random variable $(\boldsymbol{z}, \boldsymbol{f}) \sim \eta$. At the same time, to ensure the computational feasibility we can parametrize the potential task operators by a family of decision rules $\mathcal{T}_\vartheta : X \to D$, with $\vartheta \in \Xi$ given by a deep neural network architecture.*

The task operator is a decision rule $\mathcal{T}_{\vartheta^*} : X \to D$ parametrized by a finite dimensional parameter in Ξ. Therefore, the optimal task operator $\vartheta^* \in \Xi$ is given by the empirical decision risk minimization

$$\vartheta^* \in \underset{\vartheta \in \Xi}{\arg\min} \left\{ \frac{1}{m} \sum_{i=1}^{m} \ell_D(\tau(z_i), \mathcal{T}_\vartheta(f_i)) \right\}, \tag{7.4.3}$$

where $\tau : \Delta \to D$ is the feature mapping, and $\ell_D : D \times D \to \mathbb{R}$ is the decision distance (see Definition 7.3.5). Let us explore a classical example to illustrate how tasks on model parameters work, the problem of image classification.

Task on model parameter 7.4.4 (Image classification)**.** *Let $X := L^2(\Omega, \mathbb{R})$ be the space of gray-scale images on $\Omega \subset \mathbb{R}^2$ (open). The image classification task is formed by the following elements.*

- ***Task.*** *Classify an image into one of k distinct labels (classes), more precisely, assign to an image a probability distribution over all k labels.*
- ***Model parameter space.*** *The space of gray-scale images defined on a fixed domain $\Omega \subset \mathbb{R}^2$, that is,*
$$X := L^2(\Omega).$$
- ***Feature space.*** *The space defined by the finite cyclic group*
$$\Delta := \mathbb{Z}_k = \mathbb{Z}/k\mathbb{Z},$$
representing the k labels.
- ***Decision space.*** *The space of probability distributions over the k labels, i.e.*
$$D := \mathscr{P}_\Delta.$$

- ***Decision loss.*** *The task-adapted loss function is given by* (7.3.4)*, where*

$$\ell_D(d, d') := -\sum_{z\in\Delta} d(z) \log d'(z) \quad \textit{for} \quad d, d' \in D.$$

- ***Feature map.*** *Dirac delta function for each class*

$$\tau(z) := \delta_z \quad \textit{for} \quad z \in \Delta.$$

- ***Task operator.*** *Given a prior distribution $\eta_0 \in \mathscr{P}_\Delta$, the Bayes risk in* (7.3.6) *associated with a decision rule $\mathcal{T} : X \to D$ becomes*

$$\mathcal{R}_{\eta_0}(\mathcal{T}) := \mathbb{E}_{\eta_0\otimes\mathbb{P}_z}\left[\ell_D(\tau(\boldsymbol{z}), \mathcal{T}(\boldsymbol{f}))\right] = \int_X \int_\Delta \left[-\log[\mathcal{T}(x)(z)]\right] d\eta_0(z) d\mathbb{P}_z(x).$$

The task operator is then parametrized by a continuous neural network $\mathcal{T}_\vartheta : X \to D$ for $\vartheta \in \Xi$. The optimal task operator $\mathcal{T}_{\vartheta^*} : X \to D$ is obtained from the corresponding empirical risk minimization in (7.4.3)*, namely,*

$$\vartheta^* \in \arg\min_{\vartheta\in\Xi} \left\{ \frac{1}{m} \sum_{i=1}^{m} \left[-\log[\mathcal{T}_\vartheta(d_i)(z_i)]\right] \right\} \quad \textit{for training data} \quad (z_i, f_i) \in \Delta \times X. \tag{7.4.4}$$

There are already several approaches using deep neural network architectures for the set of decision rules $\mathscr{D} = \{\mathcal{T}_\vartheta\}_{\vartheta\in\Xi}$, and solving (7.4.4) corresponds to training a classifier. EfficientNet [110] and ResNet [56] are the most relevant examples of deep learning-based image classifiers. This example illustrates the main idea of tasks on model parameters for a real-world application. We are now ready to explain how to adapt reconstruction coming from an inverse problem to a particular task on its model parameters.

7.4.2 Abstract setting of task-adapted reconstruction

We refer to methods that integrate the reconstructions with the decision-making procedure as *task-adapted reconstruction*, since we are in some sense adapting the reconstruction method to a task at hand. Following [1] we introduce a framework for task-adapted reconstruction. This framework is at the same time computationally feasible and adaptable to specific inverse problems and tasks. The key idea in this integration is the formalization of both the reconstruction and tasks as non-randomized decision rules within a statistical estimation problem.

Let us start with an inverse problem, where the data model $\mathcal{M}$ in (7.2.1) is known. As in Section 7.3, the reconstruction can be understood as a decision rule in a statistical estimation problem. This is defined by the statistical model $((Y, \mathfrak{S}_Y), \{\mathcal{M}(f)\}_{f\in X})$, decision space $(X, \mathfrak{S}_X)$, and loss $\ell_X : X \times X \to \mathbb{R}$. If $\pi_0 \in \mathscr{P}_X$ is a prior, one can define a reconstruction method as an optimal non-randomized decision rule obtained by minimizing the Bayes risk corresponding to the prior. This will then be a mapping that solves

$$\hat{\mathcal{A}}^\dagger \in \underset{\mathcal{A}^\dagger \in \mathscr{M}(Y,X)}{\arg\min} \left\{ \mathbb{E}_{\pi_0 \otimes \mathcal{M}(f)} \left[\ell_X(\boldsymbol{f}, \mathcal{A}^\dagger(\boldsymbol{g})) \right] \right\}, \quad \text{where} \quad (\boldsymbol{f}, \boldsymbol{g}) \sim \pi_0 \otimes \mathcal{M}(f), \tag{7.4.5}$$

where $\mathscr{M}(Y, X)$ is the space of measurable mappings from the measure spaces $(X, \mathfrak{S}_X)$ to $(Y, \mathfrak{S}_Y)$. Next, as seen in Section 7.4.1, we know that a task is a decision rule in a statistical estimation problem defined by the statistical model $((X, \mathfrak{S}_X), \{\mathbb{P}_z\}_{z \in \Delta})$, decision space $(D, \mathfrak{S}_D)$, and loss given by (7.3.4) with known feature extraction map $\tau : \Delta \to D$ and decision distance $\ell_D : D \times D \to \mathbb{R}$. Similar to the reconstruction, if $\eta_0 \in \mathscr{P}_\Delta$ is a task prior, we can define the task operator as the non-randomized decision rule that minimizes the Bayes risk corresponding to the prior η_0, i.e.,

$$\hat{\mathcal{T}} \in \underset{\mathcal{T} \in \mathscr{M}(X,D)}{\arg\min} \left\{ \mathbb{E}_{\eta_0 \otimes \mathbb{P}_z} \left[\ell_D(\tau(\boldsymbol{z}, \mathcal{T}(\boldsymbol{f})) \right] \right\}, \quad \text{where} \quad (\boldsymbol{z}, \boldsymbol{f}) \sim \eta_0 \otimes \mathbb{P}_z, \tag{7.4.6}$$

where $\mathfrak{M}(X, D)$ is the set of measurable mappings from $(X, \mathfrak{S}_X)$ to $(D, \mathfrak{S}_D)$. There are three main approaches to combine the involved optimizations.

- **Sequential approach:** in the sequential approach one first computes the reconstruction operator, and then uses it to define the task operator. In this approach, one assumes that the joint laws $(\boldsymbol{z}, \boldsymbol{f}) \sim \eta_0 \otimes \mathbb{P}_z$ and $(\boldsymbol{f}, \boldsymbol{g}) \sim \pi_0 \otimes \mathcal{M}(f)$ are consistent. For example, the $\mathbb{P}_z$ is the push forward of $\mathcal{M}(f)$ through the reconstruction operator. Another possible assumption is that $\pi_0 \in \mathscr{P}_X$ can be obtained by marginalizing the measure $\eta_0 \otimes \mathbb{P}_z$ over Δ using $\eta_0 \in \mathscr{P}_\Delta$. In this approach the task-adapted reconstruction operator is given by

  $$\hat{\mathcal{T}} \circ \hat{\mathcal{A}}^\dagger : Y \to D, \tag{7.4.7}$$

  where $\hat{\mathcal{A}}^\dagger \in \mathscr{M}(Y, X)$ solves (7.4.5) and $\hat{\mathcal{T}} \in \mathscr{M}(X, D)$ solves (7.4.6).

- **End-to-end approach:** this approach ignores the distinction between the reconstruction and the task. Assuming a joint low $(\boldsymbol{z}, \boldsymbol{g}) \sim \nu$ for some measure $\nu \in \mathscr{P}_{\Delta \times Y}$, the task-adapted reconstruction is then given as the operator $\hat{\mathcal{B}} : Y \to D$ that solves

  $$\hat{\mathcal{B}} \in \underset{\mathcal{B} \in \mathscr{M}(Y,D)}{\arg\min} \left\{ \mathbb{E}_{\eta_0 \otimes \mathbb{P}_z} \left[\ell_D(\tau(\boldsymbol{z}), \mathcal{B}(\boldsymbol{g})) \right] \right\} \quad \text{where} \quad (\boldsymbol{z}, \boldsymbol{f}) \sim \nu. \tag{7.4.8}$$

- **Joint approach:** this approach is a mid-way between the sequential and the end-to-end-approaches. In such approach one assumes that there is a joint law $(\boldsymbol{z}, \boldsymbol{f}, \boldsymbol{g}) \sim \sigma$, which by the chain rule in probability can be written in terms of conditional probabilities

  $$d\sigma(z, f, g) = d\pi(g|z, f) d\pi(f|z) d\pi(z).$$

One also assumes that f is a sufficient statistic for g, i.e., $d\pi(g|z,f) = d\pi(g|f)$. This, combined with $(\boldsymbol{z},\boldsymbol{f}) \sim \eta_0 \otimes \mathbb{P}_z$ and $(\boldsymbol{f},\boldsymbol{g}) \sim \pi_0 \otimes \mathcal{M}(f)$, results on

$$d\sigma(z,f,g) = d\mathcal{M}(g)d\mathbb{P}_z(f)d\eta_0(z).$$

We then introduce the *joint loss* that interpolates between the sequential case and the end-to-end approaches. In particular, we let $\ell_{\text{joint}} : (X \times D) \times (X \times D) \to \mathbb{R}$ be given as

$$\ell_{\text{joint}}((f,d),(f',d')) := (1-C)\ell_X(f,f') + C\ell_D(d,d') \quad \text{for fixed} \quad C \in [0,1]. \tag{7.4.9}$$

task-adapted reconstruction is then given by (7.4.7), where the operators jointly solve the equation

$$(\hat{\mathcal{A}}^\dagger, \hat{\mathcal{T}}) \in \underset{\substack{\mathcal{T} \in \mathscr{M}(X,D) \\ \mathcal{A}^\dagger \in \mathscr{M}(Y,X)}}{\arg\min}\ \mathbb{E}_\sigma \left[\ell_{\text{joint}} \Big((\boldsymbol{f}, \tau(\boldsymbol{z})), (\mathcal{A}^\dagger(\boldsymbol{g}), \mathcal{T} \circ \mathcal{A}^\dagger(\boldsymbol{g})) \Big) \right]. \tag{7.4.10}$$

Remark 7.4.5. *Notice that when $C \to 0$, the joint and sequential approach are the same. One may think that it is sufficient to only consider the loss ℓ_D in (7.4.10), i.e., to set $C = 1$ in (7.4.9), which results in the end-to-end approach. The main problem with this approach is that one obtains non-uniqueness since if $(\hat{\mathcal{A}}^\dagger, \hat{\mathcal{T}})$ solves (7.4.10), then $(\mathcal{B}^{-1} \circ \hat{\mathcal{A}}^\dagger, \hat{\mathcal{T}} \circ \mathcal{B})$ solves (7.4.10) for any invertible $\mathcal{B} : X \to X$. This problem does not happen when $C < 1$, hence adding a loss term associated with the reconstruction acts as a regularizer. This also tells us that the limit $C \to 1$ does not necessarily coincide with the case $C = 1$.*

In the next section we will explore the computational implementation of these three approaches.

7.4.3 Computational implementation of task-adapted reconstruction

We have discussed already, on the one hand, the difficulty in finding an appropriate prior $\pi_0 \in \mathscr{P}_X$ for the Bayesian inversion, although the measure $\mathcal{M}(f) \in \mathscr{P}_Y$ is known by the data model. On the other hand, both measures $\eta_0 \in \mathscr{P}_\Delta$ and $\mathbb{P}_z \in \mathscr{P}_X$ are considered unknown for most tasks. Therefore, we consider the joint law $(\boldsymbol{f},\boldsymbol{g},\boldsymbol{z}) \sim \sigma$ as unknown. In order to implement the task-adapted reconstruction, we have to replace such $(\boldsymbol{f},\boldsymbol{g},\boldsymbol{z})$ with their empirical counterparts, where the latter is given by suitable training data.

Another issue that needs to be tackled for the implementation is computational feasibility. In the minimization involved in (7.4.5), (7.4.6), (7.4.8), and (7.4.10) one needs to explore all measurable mappings between relevant spaces. This is clearly computationally unfeasible. As we did before, this problem can be tackled by parametrizing sets of measurable mappings with deep neural networks.

Following [1] we make use of the learned iterative scheme given by the learned primal-dual architecture (Section 5) to parametrize a family of reconstruction methods $\mathcal{A}_\theta^\dagger : Y \to X$, since this parametrization already contained knowledge about the data model.

Similarly, the decision rules associated with the task are given by a parametrized family of mappings $\mathcal{T}_\vartheta : X \to D$. These parametrizations allow us to rewrite (7.4.5), (7.4.6), (7.4.8), and (7.4.10) as a training procedure for adequate training data. This procedure uses stochastic gradient descent for finding appropriate parameters is done by approximately solving the empirical versions of such optimization problems. In order to apply such methods, we require the above parametrizations to be differentiable, which requires a differentiable loss function.

The three approaches of task-adapted reconstruction in the empirical case can be reformulated as follows.

- **Sequential approach:** in this case we have two coupled sets of training data

$$\begin{aligned} &(f_i, g_i) \in X \times Y \quad \text{generated by} \quad (\boldsymbol{f}, \boldsymbol{g}) \sim \pi_0 \otimes \mathcal{M}(f) \quad \text{for} \quad i = 1, \dots, m, \\ &(z_i, f_i) \in \Delta \times X \quad \text{generated by} \quad (\boldsymbol{z}, \boldsymbol{f}) \sim \eta_0 \otimes \mathbb{P}_z \quad \text{for} \quad i = 1, \dots, m. \end{aligned} \tag{7.4.11}$$

  The coupling resides in the fact that f_i's in the second data set are reconstructions obtained from g_i's in the first data set, which ensures consistency with the statistical assumptions of the sequential approach. Therefore, the task-adapted reconstruction operator is given by the mapping

$$\mathcal{T}_{\vartheta^*} \circ \mathcal{A}^\dagger_{\theta^*} : Y \to D, \tag{7.4.12}$$

  where $\theta^* \in \Theta$ solves (7.3.8) and $\vartheta^* \in \tilde{\Xi}$ solves (7.4.3), meaning

$$\begin{aligned} \theta^* &\in \arg\min_{\theta \in \Theta} \left\{ \frac{1}{m} \sum_{i=1}^m \ell_X(f_i, \mathcal{A}^\dagger_\theta(g_i)) \right\}, \quad \text{and} \\ \vartheta^* &\in \arg\min_{\vartheta \in \tilde{\Xi}} \left\{ \frac{1}{m} \sum_{i=1}^m \ell_D(\tau(z_i), \mathcal{T}_\vartheta(f_i)) \right\}. \end{aligned} \tag{7.4.13}$$

  Thus, in the empirical case, the problems (7.4.5) and (7.4.6) are replaced by (7.4.13). The learned task operator related to ϑ^* in (7.4.13) is well defined just for inputs taken from the support of its training data. So, it may fail when applied to data that it has never seen (out of distribution). Hence, it is important to make sure that the range of the reconstruction operator is a subset of the support of the elements $f \in X$ used to train the task. Usually, this can be ensured by letting f_i's in (z_i, f_i) in (7.4.11) be the output of the learned reconstruction operator $\mathcal{A}^\dagger_{\theta^*} : X \to Y$.

- **End-to-end approach:** the training data in this case is of the form

$$(z_i, g_i) \in \Delta \times X \quad \text{generated by} \quad (\boldsymbol{z}, \boldsymbol{g}) \sim \eta_0 \otimes \mathbb{P}_z \quad \text{for} \quad i = 1, \dots, m. \tag{7.4.14}$$

  Then, the task-adapted reconstruction operator is given by $\mathcal{B}_\vartheta : Y \to D$, with $\vartheta^* \in \tilde{\Xi}$ the solution of

$$\vartheta^* \in \arg\min_{\vartheta \in \tilde{\Xi}} \left\{ \frac{1}{m} \sum_{i=1}^m \ell_D(\tau(z_i), \mathcal{B}_\vartheta(g_i)) \right\}. \tag{7.4.15}$$

This replaces (7.4.8) for the empirical case.

- **Joint approach:** in this approach we need training data that jointly contains elements related to the reconstruction and the task

  $$(z_i, f_i, g_i) \in \Delta \times X \times Y \quad \text{generated by} \quad (\boldsymbol{z}, \boldsymbol{f}, \boldsymbol{g}) \sim \sigma \quad \text{for} \quad i = 1, \dots, m. \tag{7.4.16}$$

  Therefore, the corresponding task-adapted reconstruction operator can be defined as in (7.4.12) with $(\theta^*, \vartheta^*) \in \Theta \times \tilde{\Xi}$ solving the following joint empirical loss minimization

  $$(\theta^*, \vartheta^*) \in \underset{(\theta,\vartheta)\in\Theta\times\tilde{\Xi}}{\arg\min} \left\{ \frac{1}{m} \sum_{i=1}^{m} \ell_{\text{joint}}((f_i, \tau(z_i)), (\mathcal{A}_\theta^\dagger(g_i), \mathcal{T}_\vartheta \circ \mathcal{A}_\theta^\dagger(g_i))) \right\}, \tag{7.4.17}$$

  where $\ell_{\text{joint}} : (X \times D) \times (X \times D) \to \mathbb{R}$ is the joint loss in (7.4.9). Therefore, Problem (7.4.17) replaces (7.4.10) in the empirical case.

In the next sections, we will present an application that incorporates the ideas of microlocal analysis of conv-ResNets and the learned primal-dual architecture (Chapter 4) and tomographic reconstruction in the context of task-adapted reconstruction.

7.5 Task-adapted tomographic reconstruction and wavefront sets

In this section, we introduce two particular applications of task-adapted reconstruction (Definition 7.4.1). In these applications, we combine tomographic reconstruction (Definition 7.1.2) with the relevant information provided by the wavefront set. The first application, task-adapted tomographic reconstruction and wavefront set extraction (Section 7.5.1), jointly performs tomographic reconstruction and extracts the wavefront set directly from the data. The second and more relevant applicaton, task-adapted tomographic reconstruction and wavefront set inpainting (Section 7.5.2), makes use of the microcanonical relation of conv-ResNets presented in this thesis (Section 4.2.6) to propagate the undersampled wavefront set from the data to the reconstruction. This information is then used to reconstruct the fully sampled wavefront set. Jointly reconstructing the image and its wavefront set allows us to improve the resolution of the object boundaries, avoiding unusual artifacts. Let us start with the task introduced in Chapter 5, i.e., wavefront set extraction.

7.5.1 Task-adapted tomographic reconstruction and wavefront extraction

In this section we revisit the task of wavefront set extraction of digital images (Chapter 5), in addition, we also reformulate it in the context of statistical decision theory. This allows us to incorporate to the method the framework of task-adapted reconstruction, in the particular case where the inverse problem corresponds to tomographic reconstruction (Definition 7.1.2). Due to the discussed advantages of the joint approach, we focus solely on this setting. For that, let us first recall from Section 7.4 the main assumptions in the task-adapted reconstruction framework.

(i) Both measures $\mathbb{P}_\Delta$ on the feature space Δ and $\mathbb{P}_z$ on the model parameter space X are unknown. Therefore, also the prior $(\boldsymbol{z}, \boldsymbol{f}, \boldsymbol{g}) \sim \sigma$ can be regarded as unknown.

(ii) The minimization of the joint Bayes risk (Equation (7.4.10)) is done over all measurable mappings $\mathcal{T} \in \mathscr{M}(X, D)$ (task) and $\mathcal{A}^\dagger \in \mathscr{M}(Y, X)$ (reconstruction). This is computationally unfeasible.

As presented in Section 7.4.3, we can tackle both problems with tools of supervised learning. The challenge (i) is solved when replacing the original measures with their empirical counterparts given by suitable supervised data

$$\{(z_i, f_i, g_i)\}_{i=1,\dots,N},$$

where $(z_i, f_i, g_i) \in \Delta \times X \times Y$ are generated by $(\boldsymbol{z}, \boldsymbol{f}, \boldsymbol{g}) \sim \sigma$. We can overcome (ii) by parametrizing a collection of task-adapted reconstruction operators $\mathcal{T}_\vartheta \circ \mathcal{A}_\theta^\dagger$ in the search space with deep neural networks

$$\mathcal{D} := \{\mathcal{T}_\vartheta\}_{\vartheta \in \tilde{\Xi}} \text{ and } \tilde{\mathcal{D}} := \{\mathcal{A}_\theta^\dagger\}_{\theta \in \Theta}, \quad \text{where } \mathcal{T}_\vartheta \text{ and } \mathcal{A}_\theta^\dagger \text{ are deep neural networks.}$$

The learned task-adapted reconstruction operator is then given by $\mathcal{T}_{\vartheta^*} \circ \mathcal{A}_{\theta^*}^\dagger : Y \to D$, where $(\theta^*, \vartheta^*) \in \Theta \times \tilde{\Xi}$ is given by

$$(\theta^*, \vartheta^*) \in \underset{(\theta,\vartheta)\in\Theta\times\tilde{\Xi}}{\arg\min} \left\{ \frac{1}{m} \sum_{i=1}^{m} \ell_{\text{joint}}((f_i, \tau(z_i)), (\mathcal{A}_\theta^\dagger(g_i), \mathcal{T}_\vartheta \circ \mathcal{A}_\theta^\dagger(g_i))) \right\},$$

for $\ell_{\text{joint}} : (X \times D) \times (X \times D) \to \mathbb{R}$ a suitable joint loss, see (7.4.9). In this case we focus on computed tomography as the inverse problem, and on wavefront set extraction as the task.

In addition for the reconstruction operator, we use the parametrization of the reconstruction operator $\mathcal{R}_\theta$ given by the learned-primal dual architecture from Algorithm 5. We use the task operator for digital wavefront set extraction given by the DeNSE architecture presented in Algorithm 7.

7.5.1.1 Abstract setting

In the following, we present a detailed description of our problem.

Task adapted reconstruction 7.5.1 (Tomgraphic reconstruction and wavefront set extraction). *Let us assume that Ω^d is an $N \times N$ grid in $\mathbb{R}^2$ as in Definition 6.2.12, also known as the* digital image domain. *The task-adapted tomographic reconstruction and wavefront set extraction is defined as follows.*

- ***Task.*** *Find the point-wise probability distribution for the 180 binary labels associated to each point in an image. Since the probability for each label is a number in* $[0, 1]$*, we may identify this point-wise probability distribution with a 180-channel grey-scale image. The task of extracting the wavefront set is then a mapping that reads the discrete shearlet coefficients of an image and outputs a continuous 180-channel gray-scale image.*

- ***Data.*** *Elements in* Y *are real-valued functions defined on lines representing samples of sinograms coming from noisy measurements of the parallel beam radon transform, defined on* (2.5.1)*. These measurements can come either from a low-dose setting (sparsely sampled angles, equally distanced) or a limited-angle setting (densely sampled angles with an unmeasured wedge).*

- ***Model parameter space.*** *The model parameter is here the space of 2D gray-scale digital images represented by a real-valued* ℓ^2*-function on the image domain* $\Omega^d \subset \mathbb{R}^2$*. Hence, the model parameter space becomes*

$$X := \ell^2(\Omega^d).$$

- ***Model parameter loss.*** *A natural loss on* X *is the channel-wise* ℓ^2*-distance, which is defined as* $\ell_X : X \times X \to \mathbb{R}$ *where*

$$\ell_X(x, x') = \|x - x'\|_2^2 \quad \text{for } x, x' \in X.$$

- ***Feature space.*** *The feature space* Δ *is here the space of* $\{0, 1\}^{180}$*-valued measurable mappings on* Ω*. This space can be identified with the* 180*-product space of* $\{0, 1\}$*-valued measurable mappings on* Ω^d*, i.e.,*

$$\Delta := \mathcal{M}(\Omega^d, \{0, 1\})^{180}$$

- ***Decision space.*** *Elements in* D *represent* digital wavefront sets*. The idea is to encode a digital wavefront set of an image by associating a* 180*-array of binary probability distributions at each point in* Ω*. Hence,* D *is a space of measurable mappings that maps a point in* Ω *to a* 180*-array of probability distributions on* $0, 1$*. Each such binary probability distribution can be identified with a scalar in* $[0, 1]$*, so the decision space can be written as measurable mappings from* Ω *to a* 180*-array of scalars in* $[0, 1]$*, which is*

$$D := \mathcal{M}(\Omega^d, [0, 1])^{180}$$

- ***Decision loss.*** *The decision loss is the sum of cross-entropies (classification loss) corresponding to each category*

$$\ell_D(d, d') := \int_\Omega \Big(-\sum_{i=1}^{180} d_i(t) \log\big(d_i'(t)\big) \Big) dt \quad \text{for } d, d' \in \mathcal{M}(\Omega, [0, 1])^{180}.$$

  Note here that d, d' *represent two probability distributions on* $\{0, 1\}^{180}$ *and* d_i *refers to the entry of* d *corresponding to the i:th class, i.e., it is the marginal distribution corresponding to the i:th class.*

- ***Feature map.*** *The Kronecker (digital) delta function for each class*

$$\tau(z)(t) := \delta_{z(t)} \quad \textit{for } z : \Omega^d \to \{0,1\}^{180} \textit{ and } t \in \Omega^d.$$

- ***Task-adapted reconstruction operator.*** *The task-adapted operator is given by* $\mathcal{T}_{\vartheta^*} \circ \mathcal{R}^\dagger_{\theta^*} : Y \to D$, *where*

$$(\theta^*, \vartheta^*) \in \operatorname*{arg\,min}_{(\theta,\vartheta)\in\Theta\times\tilde{\Xi}} \left\{ \frac{1}{m} \sum_{i=1}^{m} \ell_{joint}((f_i, \tau(z_i)), (\mathcal{R}^\dagger_\theta(g_i), \mathcal{T}_\vartheta \circ \mathcal{R}^\dagger_\theta(g_i))) \right\},$$

where $\{\mathcal{T}_\vartheta\}_{\vartheta\in\Xi} : X \to D$ *is the task operator parametrization given by* ***DeNSE*** *[8] and* $\mathcal{R}^\dagger_\theta : Y \to X$ *is the reconstruction operator parametrization given by the learned primal-dual architecture [3]. In addition,* $\{(z_i, f_i, g_i)\}_i \in \Delta \times X \times Y$ *is a supervised training set of digital wavefront sets, digital images and digital sinograms. The joint loss* $\ell_{joint} : (X \times D) \times (X \times D) \to \mathbb{R}$ *is given by the convex combination of the model-parameter loss and the decision loss, i.e.*

$$\ell_{joint}((f, d), (\tilde{f}, \tilde{d})) := \lambda \ell_x(f, \tilde{f}) + (1 - \lambda)\ell_d(d, \tilde{d}),$$

for all $((f, d), (\tilde{f}, \tilde{d})) \in (X \times D) \times (X \times D)$, *where* $\lambda \in [0, 1]$ *is a constant that modulates the influence of the task and the reconstruction.*

The approach presented above formalizes the concept of task-adapted tomographic reconstruction and digital wavefront set extraction. In Section 8.4.2, we present a collection of numerical experiments for this setting. We also show that this approach performs poorly in general. The poor performance is mainly due to the fundamental limitation imposed by the *distracted supervision paradox* discussed in Section 5.3. The joint training of the tomographic reconstruction and the wavefront set extraction forces us to simultaneously supervise the classification all of the wavefront set orientations, distracting them from converging to good minima. In addition, we are not using the digital microlocal analysis, introduced in Chapter 6, in its full potential. Instead of using the digital microcanonical relation to map the measured singularities in the sinogram to singularities in the image, we are computing the full wavefront set of the image from scratch. In the next section, we will introduce a novel task, the *wavefront set inpainting*, which tackles these two limitations, resulting in a significant improvement in the tomographic reconstruction.

7.5.2 Task-adapted tomographic reconstruction and wavefront set inpainting

Now, we are going to study the task-adapted tomographic reconstruction in the context of a new task, namely *wavefront set* inpainting. Wavefront set inpainting is the task of estimating a densely sampled wavefront set from a sparsely sampled one. The sparse sampling can come either from a low-dose [3] or a limited angle sinogram [17] via the microcanonical relation. As mentioned before, in the low-dose scenario, one measures a sparse set of line integrals along lines with equally spaced orientations. Meanwhile, in the limited-angle case, one has in addition, a wedge of angles that are not measured. Now, unlike wavefront set extraction, in this case, the task does not act directly on the model parameter space, but on the space of sparse digital wavefront sets. The sparse wavefront set on the image domain Ω^d is obtained from the propagation of the wavefront set from the sinogram domain Ξ^d (Definition 6.2.12). This propagation is done due to the digital microcanonical relation of the learned primal-dual architecture (see Remark 6.3.3).

7.5.2.1 Data preprocessing

Before the wavefront set inpainting can be performed, a pre-process on the sinogram is required. For that, we first map the sparsely sampled sinogram to its digital wavefront set, via **DeNSE** (Section 5.4). The latter is regarded as an element of the space $\mathcal{M}(Y, [0,1])^{|K|}$, where K is the set of measured orientations. The digital wavefront set of the sinogram is then mapped to the sparse digital wavefront set of the image, via the digital microcanonical relation, the later mapping singularities in the sinogram to singularities in the image. The sparse digital wavefront set of the image is an element of the space $\mathcal{M}(\Omega^d, [0,1])^{|K|}$. Notice that this pre-processing step is done with a deterministic formula, the composition of the pretrained DeNSE on the sinograms, and the digital microcanonical relation of the learned-primal dual architecture (Remark 6.3.3). This formula also involves the weights of the learned-primal dual architecture, which are used to determine the exact microcanonical relation.

In order to perform wavefront set inpainting, we represent the sparse digital wavefront set of the image, obtained from the preprocessing by another image, where each pixel has as the value the class where this pixel corresponds in $\mathcal{M}(\Omega^d, [0,1])^{|K|}$. For example, if a pixel $[i, j]$ belongs to the class k_i, where k_i is either an angle of the wavefront set, then the value of that pixel will be k_i, or -1 when the pixel is not an edge point. Figure 7.3 depicts the image representation of the sparse wavefront set in both the low-dose and limited angle setting, and the fully sampled wavefront set. In this representation, we perform the task using a standard image-to-image translation architecture, known as U-Net [97]. This architecture, when trained, allows us to approximate the fully sampled wavefront set from its sparsely sampled version. In addition, UNets are well known to be state-of-the-art in classical image processing tasks, such as segmentation [97], denoising [58] and inpainting [68]. The loss used for the U-Net training is the previously presented decision loss on the wavefront set space, given by the classification loss (cross-entropy) on each pixel. Figure 7.3 depicts the U-Net architecture used for this task.

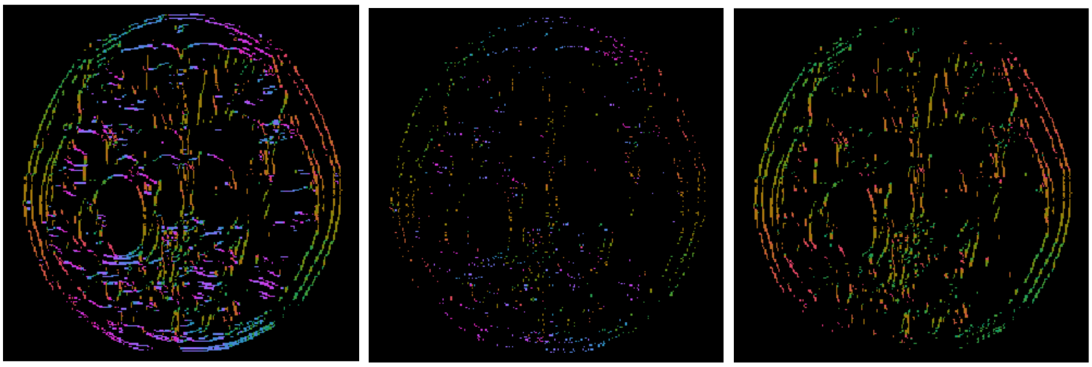

Figure 7.2: Wavefront set coming from brain-CT. Right: Full-dose wavefront set. Middle: Low-dose wavefront set. Right: Limited-angle wavefront set

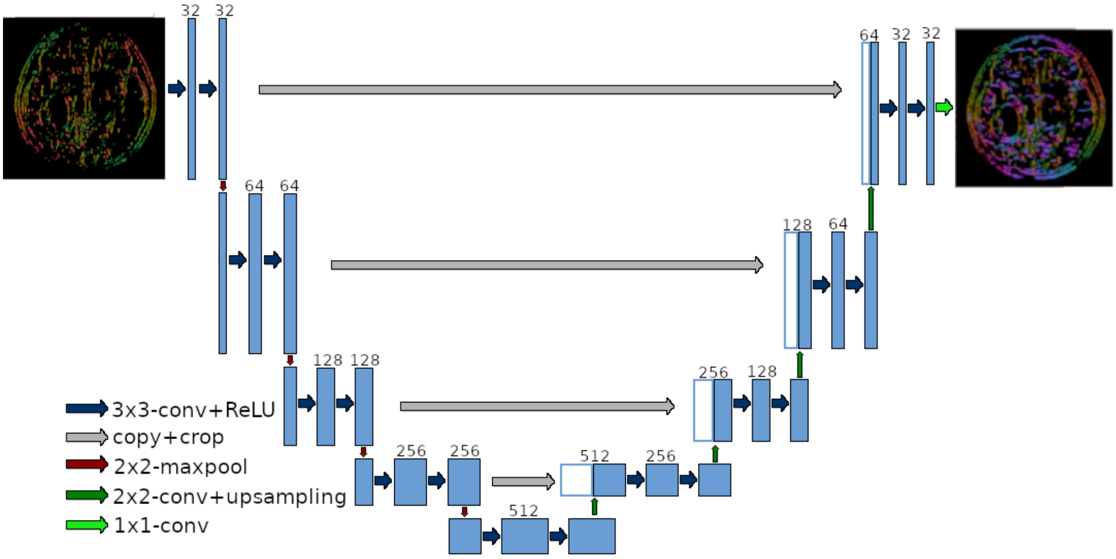

Figure 7.3: U-Net architecture used for wavefront set inpainting.

7.5.2.2 Abstract setting

By jointly training the tomographic reconstruction, via the learned primal-dual architecture, and the wavefront set inpainting, via the U-Net architecture, we force the wavefront set of the reconstruction to closely approximate the wavefront set of the ground truth. Using wavefront set inpainting also solves the challenges of the task-adapted reconstruction and wavefront set extraction presented in Section 7.5.1. On the one hand, we are making use of the a-priori information provided by the wavefront set of the sinogram and the digital microcanonical relation. On the other hand, we are avoiding the distracting supervision paradox, since the wavefront set extraction is trained previously using DeNSE, which already avoids it, as seen in Section 5.4. In the following we present this setting in detail.

Task adapted reconstruction 7.5.2 (Tomographic reconstruction + WF set inpainting)**.** *Let us assume that Ω^d is a $N \times N$ grid in $\mathbb{R}^2$, also known as the* image domain. *The task-adapted tomographic reconstruction and wavefront set inpainting is defined as follows.*

- ***Task.*** *Find the point-wise probability distribution for the 180 binary labels associated to each point in an image, corresponding to the fully sampled wavefront set, from the sparsely sampled wavefront set, corresponding to $|K|$ measured angles.*
- ***Data.*** *Elements in Y are real-valued functions defined on lines representing samples of sinograms coming from noisy measurements of the parallel beam radon transform, defined on* (2.5.1)*. These measurements can come either from a low-dose setting (sparsely sampled angles, equally distanced) or a limited-angle setting (densely sampled angles with an unmeasured wedge).*
- ***Model parameter space.*** *The model parameter is here the space 2D gray-scale digital images, represented by a real-valued ℓ^2-function on the image domain $\Omega^d \subset \mathbb{R}^2$, i.e.*
$$X := \ell^2(\Omega^d).$$
- ***Model parameter loss.*** *A natural loss on X is the L^2-distance, which is defined as $\ell_X : X \times X \to \mathbb{R}$ where*
$$\ell_X(x, x') = \|x - x'\|_2^2 \quad \textit{for } x, x' \in X.$$
- ***Feature space.*** *The feature space, Δ, is here the space of $\{0,1\}^{180}$-valued measurable mappings on Ω. This space can be identified with the* 180*-product space of $\{0,1\}$-valued measurable mappings on Ω^d:*
$$\Delta := \mathcal{M}(\Omega^d, \{0,1\})^{180}$$
- ***Decision space.*** *Elements in D represent* digital wavefront sets*. The idea is to encode a digital wavefront set of an image by associating a* 180*-array of binary probability distributions at each point in Ω. Hence, D is a space of measurable mappings that maps a point in Ω to a* 180*-array of probability distributions on* 0, 1*. Each such binary probability distribution can be identified with a scalar in* [0, 1]*, so the decision space can be written as measurable mappings from Ω to a* 180*-array of scalars in* [0, 1]*:*
$$D := \mathcal{M}(\Omega^d, [0,1])^{180}$$
- ***Decision loss.*** *The decision loss is the sum of cross-entropies (classification loss) corresponding to each category:*
$$\ell_D(d, d') := \int_\Omega \Big(-\sum_{i=1}^{180} d_i(t) \log\big(d'_i(t)\big) \Big) dt \quad \textit{for } d, d' \in \mathcal{M}(\Omega, [0,1])^{180}.$$

Note here that d, d' represent two probability distributions on $\{0,1\}^{180}$ and d_i refers to the entry of d corresponding to the i:th class, i.e., it is the marginal distribution corresponding to the i:th class.

- ***Feature map:*** *The Kronecker (digital) delta function for each class:*

$$\tau(z)(t) := \delta_{z(t)} \quad \text{for } z : \Omega^d \to \{0,1\}^{180} \text{ and } t \in \Omega^d.$$

- ***Task-adapted reconstruction operator:*** *The task-adapted operator is given by $\mathcal{T}_{\vartheta^*} \circ \mathcal{R}^{\dagger}_{\theta^*} : Y \to D$, where*

$$(\theta^*, \vartheta^*) \in \operatorname*{arg\,min}_{(\theta,\vartheta)\in\Theta\times\Xi} \left\{ \frac{1}{m} \sum_{i=1}^{m} \ell_{joint}((f_i, \tau(z_i)), (\mathcal{R}^{\dagger}_{\theta}(g_i), \mathcal{T}_{\vartheta} \circ \mathcal{R}^{\dagger}_{\theta}(g_i))) \right\},$$

where and $\mathcal{R}^{\dagger}_{\theta} : Y \to X$ is the reconstruction operator parametrization given by the learned primal-dual architecture [3]. Also, $\{\mathcal{T}_{\vartheta}\}_{\vartheta\in\Xi} : X \to D$ is the task operator parametrization given by the composition of the U-Net applied to the preprocessed data (see Section 7.5.2.1). In addition, $\{(z_i, f_i, g_i)\}_i \in \Delta \times X \times Y$ is a supervised training set of digital wavefront sets, digital shearlet coefficients and digital sinograms. The joint loss $l_{joint} : (X \times D) \times (X \times D) \to \mathbb{R}$ is given by the convex combination of the model-parameter loss and the decision loss, i.e.:

$$\ell_{joint}((f, d), (\tilde{f}, \tilde{d})) := \lambda \ell_x(f, \tilde{f}) + (1 - \lambda)\ell_d(d, \tilde{d}),$$

for all $((f, d), (\tilde{f}, \tilde{d})) \in (X \times D) \times (X \times D)$, where $\lambda \in [0, 1]$ is a constant that modulates the influence of the task and the reconstruction.

Figure 7.4 depicts the outline of the joint tomographic reconstruction and wavefront set inpainting.

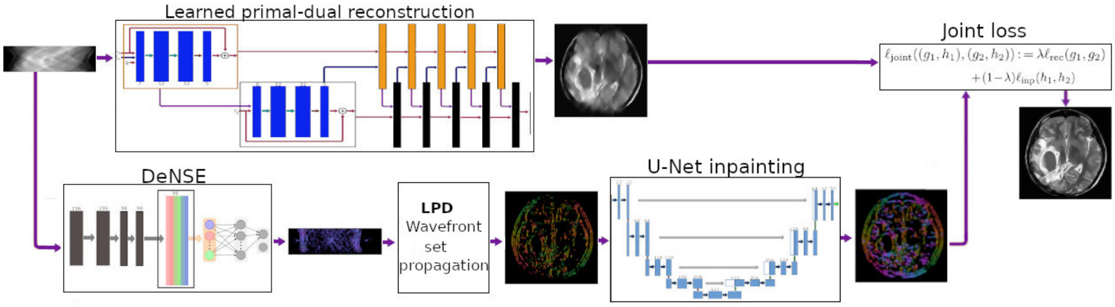
Learned primal-dual reconstruction
Joint loss
$\ell_{\text{joint}}((g_1,h_1),(g_2,h_2)) := \lambda\ell_{\text{rec}}(g_1,g_2) + (1-\lambda)\ell_{\text{inp}}(h_1,h_2)$
DeNSE
LPD
Wavefront
set
propagation
U-Net inpainting

Figure 7.4: Outline of the joint reconstruction and wavefront set inpainting algorithm. The input is partial sinogram data. In the top row first a learned primal-dual architecture is applied. In the bottom row we first apply **DeNSE** to extract the wavefront set, then we apply the canonical relation of the learned primal-dual architecture (Remark 6.3.3). To the output thereof we apply the U-Net for inpainting. This together with the output from the Learned primal-dual is then input into the joint loss function.

Remark 7.5.3. *Notice that setting of the task adapted tomographic wavefront set extraction and wavefront set coincide in almost all the elements, but the task and the task operator. On the one hand, in the case of wavefront set extraction the task operator is parametrized by **DeNSE**. On the other hand, in the case of wavefront set inpainting, the task operator is the composition of **DeNSE** applied on the sinogram and the digital microcanonical relation of the learned primal-dual architecture (Remark 6.3.3).*

Notice that this algorithm combines all of the elements previously presented in this thesis, from digital microlocal analysis of deep neural networks to learned iterative methods for image reconstruction, going through digital wavefront set extraction. In Section 8.4.3, we present a set of numerical experiments, showing that, effectively, the task-adapted tomographic reconstruction and wavefront set inpainting improve the performance of the reconstruction. As expected, the approach reduces the unusual artifacts normally present in low-dose and limited-angle tomographic reconstruction. Finally, in the next chapter, we will present numerical experiments for both cases of task-adapted reconstruction, as well as the other algorithms presented in this thesis.

8 Numerical experiments and further applications

In this chapter, we present various numerical experiments of the theory presented in this thesis, showing its utility in real-world problems. Section 8.1 presents the results on wavefront set extraction on a variety of datasets, based on the material introduced in Chapters 3 and 5. Moreover, Section 8.2 presents the results on general semantic edge detection based in the algorithms shear-CASENet and shear-DDS presented in Section 5.3.

In Section 8.3, we present the results on the wavefront propagation via the convolutional ResNets, as well as the learned primal-dual architecture. This is an implementation of the theory presented in Chapters 4 and 6. Finally, Section 8.4 presents the results corresponding to the task-adapted reconstructions methods of Chapter 7. In addition, results corresponding to wavefront set inpainting solely are also presented.

Our main goal for this chapter is to back up our main statements regarding the high performance and precision of the methods presented in this thesis. In addition, we also want to show that one can also apply these methods to real-world data, even in the absence of ground-truth. This is done by simulating close-to-real phantoms based on splines. We hope that this convinces the reader that our methods are worth trying in real-world applications, and maybe in the future, an actual human patient will benefit from them.

My own contribution: This chapter presents the numerical results and implementation ideas of Chapters 5, 6 and 7. The implementations and writing in this chapter was fully done by myself. The code is also publicly available for replication purposes.

8.1 Digital wavefront set extraction

Let us first explore the implementations and performance of our proposed wavefront set extraction method presented in Section 5.4, also known as DeNSE. This method was trained on different data sets, and outperformed a variety of model and data-driven methods significantly, including recent methods that represented current state-of-the-art.

8.1.1 DeNSE training procedure

We trained the network as described in Section 5.4.2 using stochastic gradient descent to minimize the loss function over a variety of training sets. For the loss function we have used the standard classification loss function, known as *cross-entropy*. Let us assume we are classifying pixels in an image of size $N \times N$, with number of classes given by

$M \in \mathbb{N}$. Let $y_c \in \mathbb{R}^{N\times N}$ be a binary indicator representing the ground truth of the class c, therefore $y_c[i,j] = 1$ if the pixel $[i,j]$ belongs to class c. Similarly let p_c be the probability of each pixel to belong to class c obtained by the architecture, meaning the pixel $[i,j]$ has probability $p_c[i,j]$ to be of class c. Thus, the cross-entropy loss is given by

$$\sum_{i=1,j=1}^{N} \sum_{c=1}^{M} y_c[i,j] \log(p_c[i,j])$$

We used five different data sets to train our classifier and test our algorithm:

1) The first data set consists of patches of the shearlet transform of images made of random ellipses and parallelograms of different contrasts, sizes, and orientations.

2) The second data set contains ellipses and parallelograms convolved with a kernel to generate functions with a higher-order wavefront set.

3) The third data set is the Berkeley Segmentation data set (BSDS500, `https://www2.eecs.berkeley.edu/Research/Projects/CS/vision/bsds/`) provided by the Computer Vision Group of UC Berkeley. It comprises 503 natural images of different types.

4) The fourth data set is the Semantic Boundaries data set (SBD, `http://home.bharathh.info/pubs/codes/SBD/download.html`) with 11355 natural images, again provided by the Computer Vision Group of UC Berkeley.

We depict examples of functions from each of the data sets in Figures 8.1, 8.2, 8.3 and 8.4. To make these data sets suitable for our purposes, we need to equip each image of the data sets with an associated set of labels indicating the associated wavefront set or the set of edges. For the first two and the last data set, standard theoretical results on the wavefront sets of characteristic functions allow us to compute the associated wavefront sets analytically (see Theorem 2.2.11 and Proposition 2.2.14). The segmentation and semantic boundaries data sets, on the other hand, are natural images where such an approach is not possible. These data sets are used to assess the quality of segmentation and contour detection applications, see [54] and [10]. Therefore, every image in these data sets was annotated and has a set of ground truth edges. However, we should point out that this annotated ground truth does not contain all edges of the images, but only those between semantically different parts of the images. We depict the annotated edges in Figures 8.1, 8.2, 8.3, 8.4, and 8.4. In the following subsections, we describe the computation of the associated wavefront sets in detail.

8.1.1.1 Ellipses and parallelograms

The wavefront sets of characteristic functions of ellipses and parallelograms can be identified by Proposition 2.2.14. We also use the fact that if $x \in \mathbb{R}^2$ is a vertex of a parallelogram P then $\{x\} \times \mathbb{S}^1 \subset \mathrm{WF}(\chi_P)$. For sums of these functions, we have, by basic properties of the Fourier transform, that

$$\mathrm{WF}(\chi_{P_1} + \chi_{P_2}) \subset \mathrm{WF}(\chi_{P_1}) \cup \mathrm{WF}(\chi_{P_2}).$$

Note that in this relation we do not have equality in general. Indeed, if $\mathrm{WF}(\chi_{P_1}) \cap \mathrm{WF}(\chi_{P_2}) \neq \varnothing$ then cancellations can occur. We shall neglect this technicality as the probability of cancellations is sufficiently small and assume that the wavefront set of characteristic functions as described above is the union of the respective wavefront sets.

We build this data set by randomly choosing a number of parallelograms and ellipses with random positions and computing the associated ground truth of the wavefront set as described above. The number of parallelograms and ellipses is picked with a random number generator with uniform distribution in the interval $[0, 20)$. The random ellipses are characterized by the center coordinates, the angle of the major axis with respect to the x_1-axis and the size of the major and minor axis. These parameters are randomized also with a random number generator with uniform distribution. The parallelograms are characterized by the length the base and height, the inner angle and the angle of the base with respect to the x_1-axis, these parameters are also randomized using uniformly distributed random numbers. We generated a set of 10,000 patches of shearlet coefficients with size $21 \times 21 \times 49$ of the random ellipses and parallelograms images. We use 7,000 patches for training, 1,000 for validation and 2,000 for testing.

8.1.1.2 Higher-order wavefront data set

The ellipses/parallelograms data set contains images with jump singularities only. To test our method on functions with higher-order singularities, such as ramp singularities, we computed the convolution of the elements of the ellipses/parallelograms data set with a filter h. The filter h is defined by its Fourier transform given by

$$\hat{h}(\xi) = \frac{1}{1+|\xi|}, \text{ for } \xi \in \mathbb{R}^2.$$

It is not hard to see that $P : f \mapsto h * f$ is an elliptic pseudo-differential operator and hence $\mathrm{WF}(h * g) = \mathrm{WF}(g)$ for all $g \in L^2(\mathbb{R}^2)$, see [43, Chapter 8 G] for details. Thus, the convolutions of the elements of the ellipses/parallelograms data set with h have the same wavefront set as the associated ellipses or parallelograms, but of a higher order.

This dataset was generated in the same way the dataset of random ellipses and parallelograms, where in addition we blur each image via the convolution with h before computing the shearlet coefficients. For this case, we have also generated 39,000 patches, with 30,000 for training, 3,000 for validation and 6,000 for testing.

8.1.1.3 Segmentation and semantic boundaries data sets

In the BSDS500 (`https://www2.eecs.berkeley.edu/Research/Projects/CS/vision/grouping/resources.html`) and the SBD (`http://home.bharathh.info/pubs/codes/`

`SBD/download.html`) data sets, the ground truth of the edges is given in form of binary images with 0's at positions where the image is smooth and 1's at locations associated to edges. This annotated edge set is depicted in Figure 8.3.

To compute the orientation of the edges, we used a five-point stencil derivative on the edges to approximate the normal vectors. To detect corners and assign the appropriate orientations we used the Harris corner detector [55]. From these images, we produce patches for the training of the network classifier. However, since the annotated image does not contain all edges we only use patches that are close to these edges for training, validation, and testing.

On the one hand, the BSDS500 dataset consists in 5000 natural images with annotated boundaries, we generated from this images 39,000 patches of shearlet coefficients of size $21 \times 21 \times 49$. We used 30,000 patches for training, 6,000 for testing and 3,000 for validation. On the other hand, the SBD dataset consists of 5623 images with annotated semantic edges. In this data set we also generated 39,000 patches of shearlet coefficients of size $21 \times 21 \times 49$, with the same distribution for training, testing and validation as in the BSDS500 case.

8.1.2 DeNSE performance on datasets

We implemented the training as described in the previous section using the GPU version of Tensorflow. To evaluate the classification quality, we use two quality measures, a mini-batch test average accuracy taken over all mini-batches and the so-called MF-score. The MF-score is computed as the mean of the F-score defined as

$$F := \frac{2PR}{R+P}, \tag{8.1.1}$$

where P is the *precision*, i.e., the number of true positives divided by the sum of true and false positives, and R is the *recall*, i.e., the number of true positives divided by the sum of true positives and false negatives, [102]. The MF-score is often used for evaluating classification performance when the distribution of classes is uneven. This is, for example, the case in edge detection, since there usually are significantly fewer edge points than smooth points in an image. Moreover, these performance measures (test average accuracy and MF-score) enable us to compare with the state-of-the-art [122] on the respective data sets. In addition, the required code to reproduce the results is publicly available in `http://www.shearlab.org/applications`.

The implementations of the other methods that were used to compare the performance of DeNSE were taken from the publicly available github repositories provided by the authors of the methods (all available in the topic: `http://github.com/topics/edge-detection`), with exception of the Canny, Sobel algorithms that were taken from the python library OpenCV (`http://opencv.org/`). The data-driven methods were trained using the hyper-parameters proposed by the authors for the given data set (Berkeley segmentation set and SBD) without further tuning.

8.1.2.1 Results for ellipses/parallelograms

We trained each of the 181 subnetworks as of Section 5.4.2 using 10,000 images as training data, 1,000 images as validation data, and 2,000 images as test data. For each direction θ_i we trained the associated subnetworks using a mini-batch procedure with 86 examples per batch and 3,000 training steps for each. We obtained an average test accuracy of 96.2% (taking the average overall 181 classifiers) and an MF-score of 97.1%. We also notice that the test accuracy of the individual classifiers was higher when classifying angles aligned to the discrete orientations of the underlying shearlet system.

We compared our method on the data set of the shearlet coefficients patches to other classifiers commonly used in machine learning, namely, logistic regression, decision trees, K-nearest neighbors, linear SVM, and random forest, using their implementations in the python library scikit-learn (`http://scikit-learn.org/`). We report the performances of these classifiers in Table 8.1.

Method	**Test accuracy**	**MF-score**
Logistic regression	45.7	48.9
Decision trees	75.2	75.8
Linear SVM	46.5	50.3
K-nearest neighbors	72.7	73.2
Random forest	86.0	86.7
DeNSE	**96.2**	**97.1**

Table 8.1: Ellipses/parallelograms data set performance metrics in percentage.

By construction, the last of the 181 subnetworks corresponds to an edge-detector, where the achieved average test accuracy was 97.5%, and the MF-score was 97.9%, the performance benchmarks with other classical edge classifiers can be found in Table 8.2. Figure 8.1 shows the results on an example of the ellipses/parallelograms data set.

We depict the classification for one instance of the test set of the parallelograms/ellipses data set in Figure 8.2 and compare the results with the classification by the heuristic approach by Yi-Labate-Easley-Krim [120]. We observe that our method performs significantly better in low contrast regions. Moreover, our algorithm appears to be more precise when differentiating between corners and edges. Here, we classify a point as a corner point if the classifiers predict at least two different orientations that differ by more than 10 degrees. In Figures 8.1, 8.2, 8.3 and 8.4, we indicate corners by white dots.

Method	**MF-score**
Canny [24]	49.1
Sobel [107]	40.0
BEL [33]	63.3
Yi-Labate-Easley-Krim [120]	70.3
CoShREM [93]	90.6
DeNSE	**97.5**

Table 8.2: Edge detection performances of edge detection algorithms on the Ellipses/parallelograms data set. The MF-Score is in percentage.

Method	**MF-score**
gPb-owt-ucm [10]	73.7
gPb [10]	71.5
Mean Shift [27]	64.0
Normalized Cuts [105]	64.2
Felzenszwalb, Huttenlocher [41]	61.0
Canny	60.3
CoShREM [93]	75.7
DeepEdge [11]	75.3
DeNSE	**95.4**

Table 8.3: BSDS500 (Berkeley) data set performance metrics in percentage.

8.1.2.2 Higher-order wavefront set data set

We performed wavefront set detection for the higher-order wavefront set data set, using the same procedure as in the ellipses/parallelograms classification. In this case, we used 30,000 patches as training data, 3,000 patches as validation data, and 6,000 patches as test data. We trained on 86-sized mini-batches, with 200,000 training steps. We obtained an average test accuracy of 93.4% and an MF-score of 94.6%. We are not aware of any algorithms specifically build for higher-order wavefront set detection, which is why we do not provide a comprehensive list of results of alternative algorithms in this case. For completeness, we added Figure 8.3 showing an example of the obtained results. We also add two predictions by the algorithm of Yi-Labate-Easley-Krim [120] and the method CoShREM [93].

It is important to mention that the algorithm of Yi-Labate-Easley-Krim is constructed to detect jump singularities and not higher-order singularities. Hence this algorithm is expected to fail on this data set. Indeed, the performance of the algorithm achieves only an MF-score of 30.5%. CoShREM, on the other hand, is built to detect edges and ridges. The performance was significantly better than that of Yi-Labate-Easley-Krim and resulted in an MF-score of 65.4%.

8.1.2.3 Berkeley segmentation set

In the Berkeley segmentation data set, the complexity of the images is considerably higher compared to the images from the ellipses/parallelogram data set. Therefore, we use a significantly larger training set to train the associated classifier. For the classification of each angle, we used 30,000 patches as training data (around 600 patches per image), 3,000 patches as validation data, and 6,000 patches as test data. As in the case of the ellipses/parallelograms, we trained using a mini-batch procedure, with 86 examples per batch, but in this case, using 30,000 training steps for each. We obtained an average test accuracy of 93.1% and MF-score of 95.4%, which is lower than the one obtained in the ellipses/parallelogram. This is likely due to the higher complexity of the patches. One advantage of this and the SBD data set is the existence of several benchmarks including state-of-the-art deep learning-based algorithms.

We compared our method using the available benchmarks on this data set provided by the UC Berkeley Computer Vision Group, we refer to [10] for a more detailed explanation of these methods. In [10], just the MF-score of the competing algorithms was reported. We give the results in Table 8.3. We present one example of the results obtained on the BSDS500 data set in Figure 8.3.

8.1.2.4 Semantic boundary dataset (SBD)

The SBD data set contains significantly more images than the BSDS500 which, as we observe below, improves the overall classification performance slightly. In this case, we used 100,000 patches as training data, 10,000 patches as validation data, and 20,000 patches as test data. We trained on 86-sized mini-batches, with 100,000 training steps. We obtained average test accuracy of 95.3% and MF-score of 96.8%.

This data set has recently been widely used for image segmentation tasks, in particular, it was used on the two deep learning-based image segmentation frameworks proposed by Z. Yu *et al.*, namely the SEAL (Simultaneous Edge Alignment and Learning) [122] and the CASENet (Category-Aware Semantic Edge Detection Network) [121]. We also compared them with the deep learning image boundary detector and classifier proposed (OBDC) by J. Y. Koh *et al.* [69]. The results can be found in Table 8.4. In addition, Figure 8.4 depicts the results in an example of the SBD dataset.

Figure 8.4 shows the results obtained by DeNSE on an example image of the SBD data set, as in the case of the BSDS500 data set, the obtained result admits more edges than the ground truth due to the batch-based approach. Nonetheless, the method outperforms even the specialized algorithms for segmentation over the given data sets.

Method	**MF-score**
OBDC	62.5
CASENet	71.8
CASENet-S	75.8
CASENet-C	80.4
CoShREM	69.7
SEAL	81.1
DeNSE	**96.8**

Table 8.4: Performance on the SBD data set. All values are in percentage.

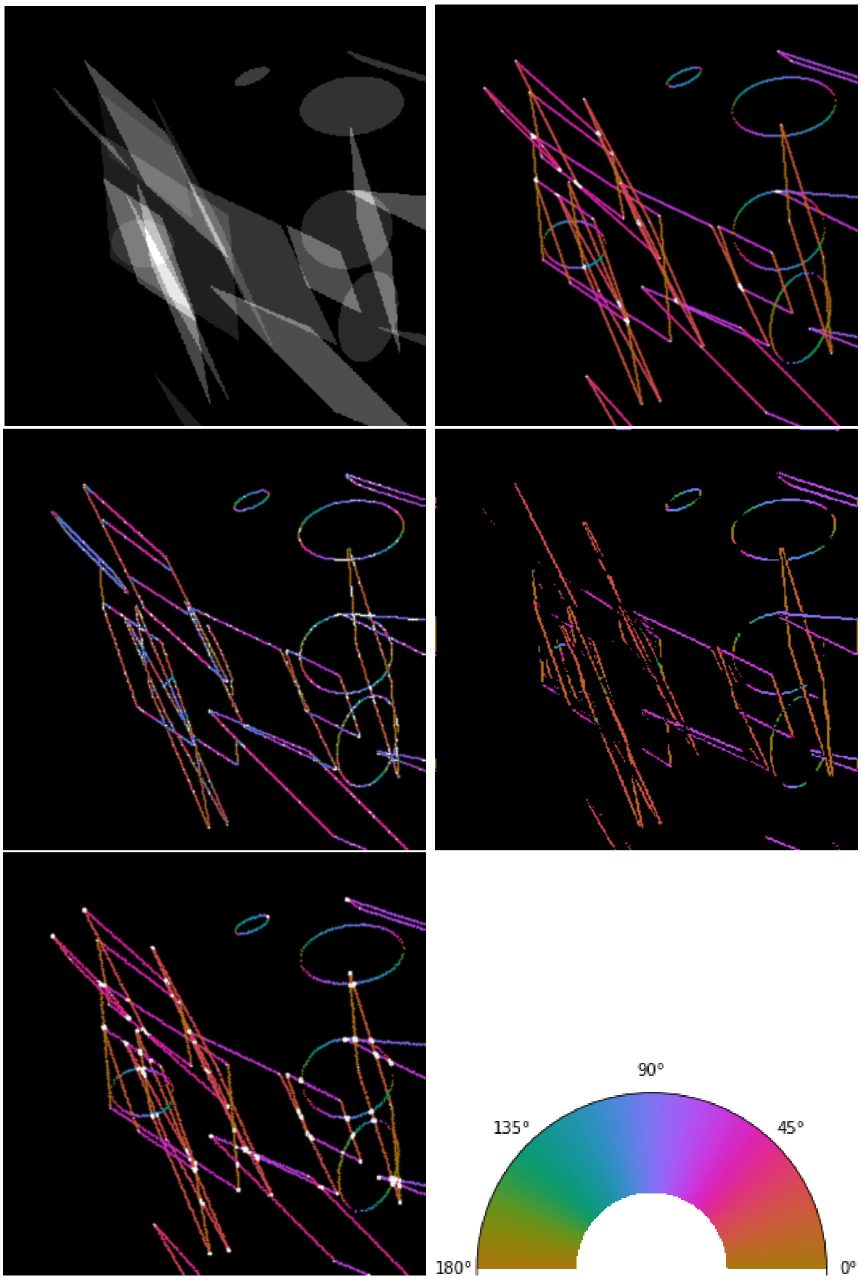

Figure 8.1: Computed edges and orientations of an example of the ellipses/parallelograms data set. Top-left: Input image. Top-right: Orientations, human annotation. Middle-left: Orientations predicted by Yi-Labate-Easley-Krim algorithm. Middle-right: Orientations predicted by CoShREM. Bottom-left: Orientations predicted by DeNSE algorithm. Bottom-right: Color code for normal-directions.

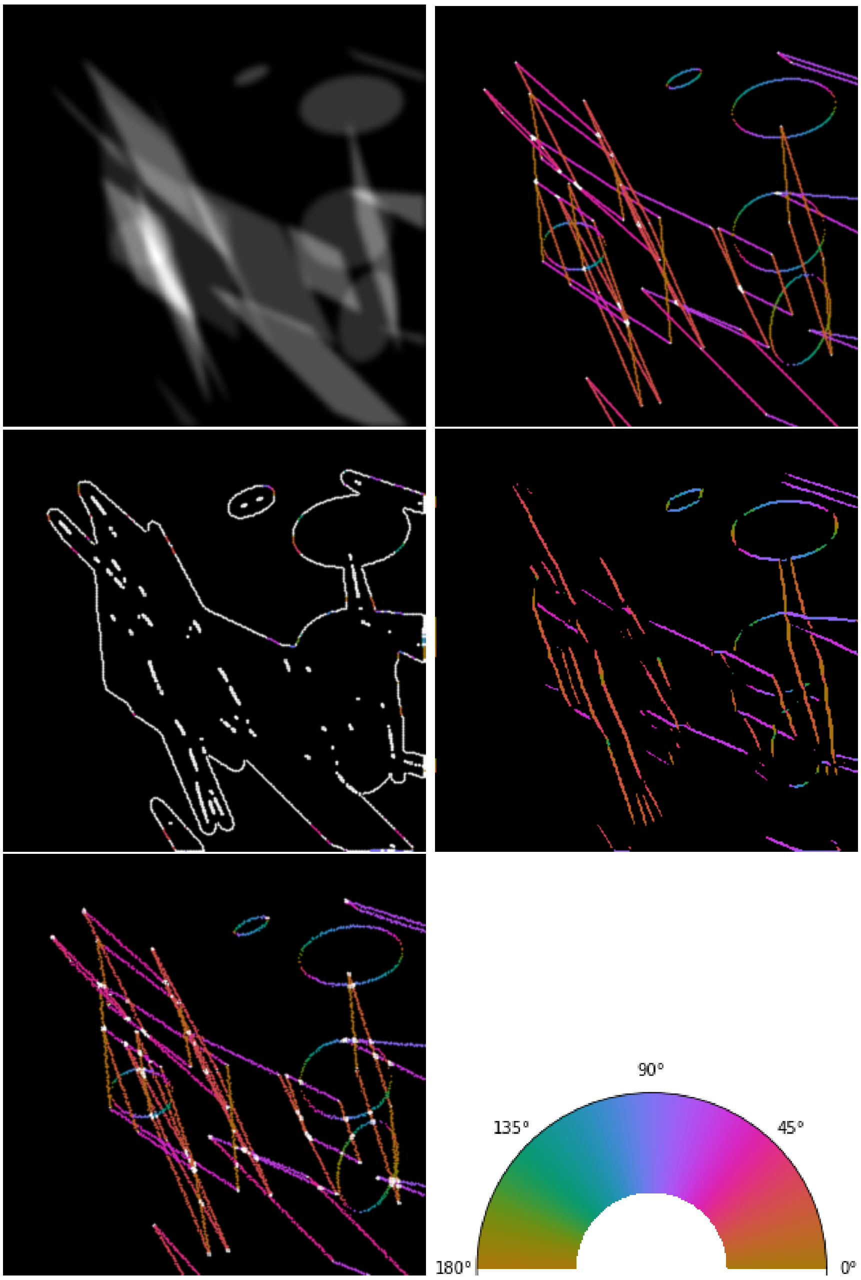

Figure 8.2: Computed edges and orientations of an example of the higher-order ellipses/-parallelograms data set. Top-left: Input image. Top-right: Orientations, human annotation. Middle-left: Orientations predicted by Yi-Labate-Easley-Krim algorithm. Middle-right: Orientations predicted by CoShREM. Bottom-left: Orientations predicted by DeNSE algorithm. Bottom-right: Color code for normal-directions.

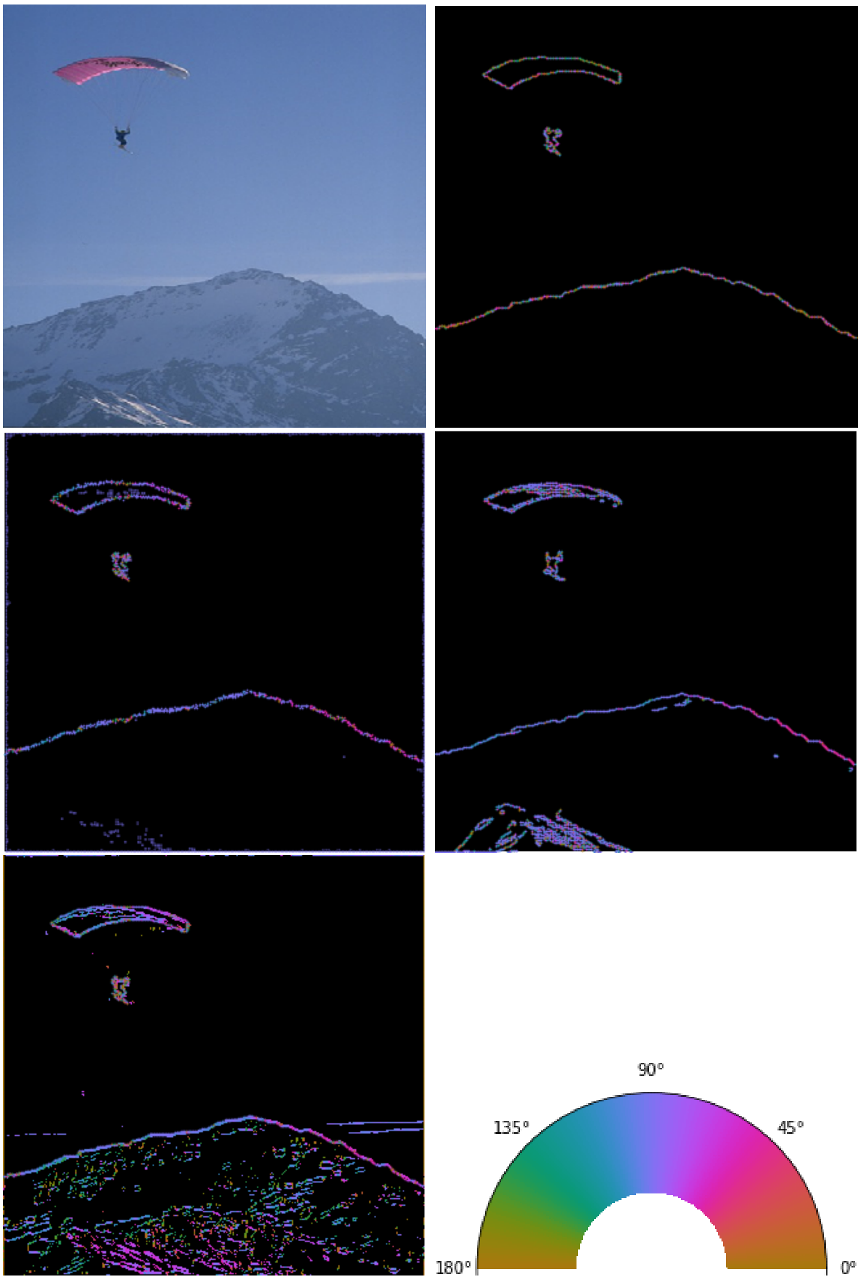

Figure 8.3: Computed edges and orientations of an example of the BSDS500 (Berkeley) data set. Top-left: Input image. Top-right: Orientations, human annotation. Middle-left: Orientations predicted by gPb-owt-ucm. Middle-right: Orientations predicted by CoShREM. Bottom-left: Orientations predicted by DeNSE algorithm. Bottom-right: Color code for normal-directions.

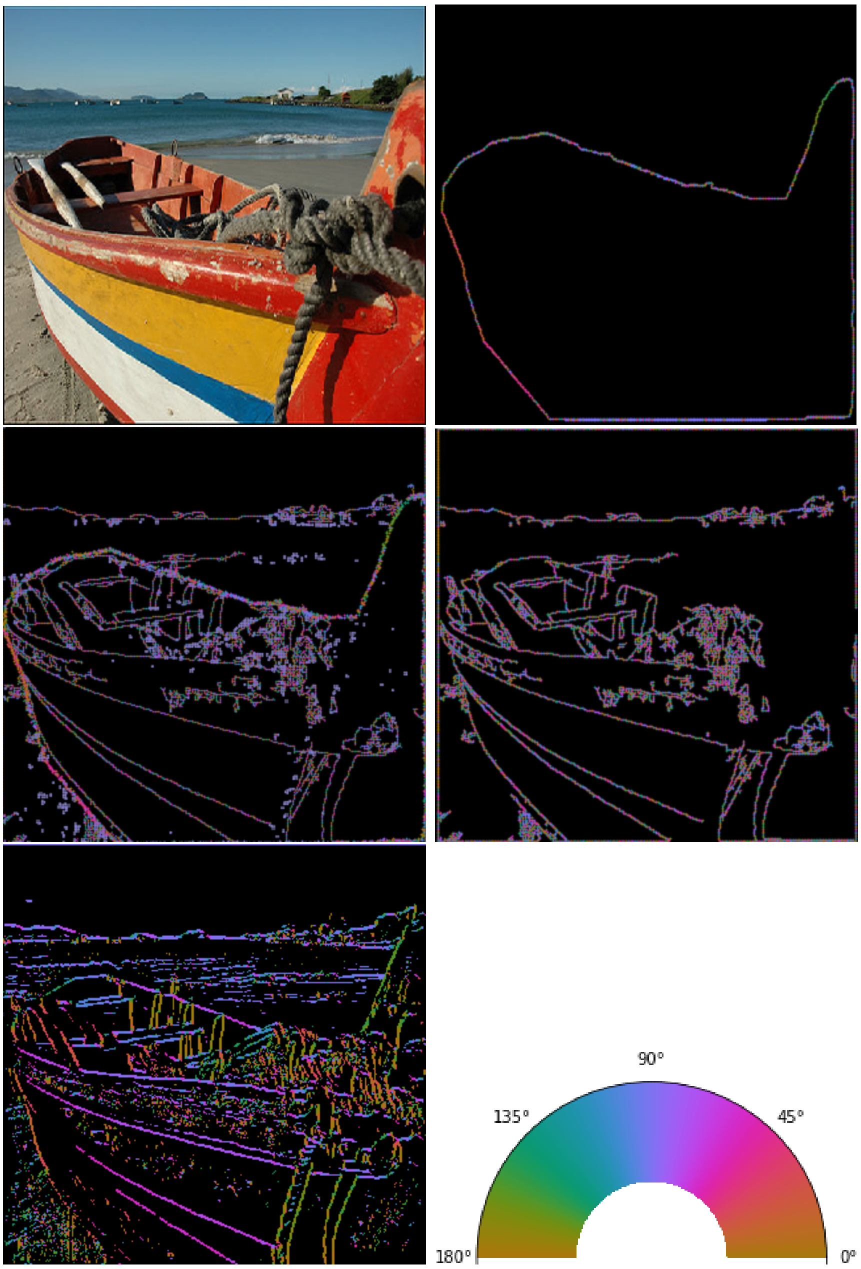

Figure 8.4: Computed edges and orientations of an example of the SBD data set.Top-left: Input image. Top-right: Orientations, human annotation. Middle-left: Orientations predicted by SEAL. Middle-right: Orientations predicted by CoShREM. Bottom-left: Orientations predicted by DeNSE algorithm. Bottom-right: Color code for normal-directions.

8.2 General semantic edge detection

In this section we present the results for numerical experiments of the general case of semantic edge detection. These results are based on the methods presented in Section 5.3. For the case of general semantic edge detection, we trained the shear-CASENet (Figure 5.4) and shear-DDS (Figure 5.5) architectures on the Semantic Boundaries Dataset (SBD). As we know, this database consists of 11,355 images, from which we used 9,035 images for training, 1,050 for evaluation, and 1,050 for testing. Each image has a human-annotated array of edge-pixels with the intensity value as the category number of the object, where this edge belongs to. The SBD dataset consists of a total of 20 categories, including vehicles, animals, and plants. One can download the dataset in `http://home.bharathh.info/pubs/codes/SBD/download.html`, in addition to our code can be found in `http://shearlab.org/applications`, making the experiments fully reproducible.

Both shear-CASENet and shear-DDS were trained on the full shearlet coefficients. Similar to the case of the wavefront set extraction, we use the digital shearlet transform [75] implemented on julia (`http://shearlab.math.lmu.de/software`), with a total of four scales. This produces a shearlet coefficients volume of 49 slices, which was then fed to the proposed architectures. We used the publicly available implementation of the CASENet architecture (`https://github.com/lijiaman/CASENet`). This implementation makes use of the deep learning framework pytorch, making it compatible with our shearlet implementation. Based on this code, we implemented the deep diverse supervision architecture by introducing the information converters and the proposed multi-task loss (see Section 5.3). Also based on this code, we implemented the Shear-CASENet and Shear-DDS architectures by extending the first convolutional layer with the corresponding shearlet channels (see Figures 5.4 and 5.5) and removing the fourth stage of the original architectures.

In addition to CASENet and DDS, we also compared the performance of shear-CASENet and shear-DDS against the deep supervised version of CASENet [121] and SEAL [122]. The performance benchmarks presented in Table 8.5 are done in terms of the mean F-score, in a similar fashion as with the wavefront set extraction benchmarks, by computing the mean of the F-score over all the categories, see Equation (8.1.1). It is visible that the mean-F value is slightly better on the shear-CASENet and Shear-DDS than on the other architectures. The improvement is not as significant as in the case of the wavefront set extraction, since DeNSE was specifically designed for wavefront set extraction and the existing models have general semantic edge detection applications. It is though worth stressing that shear-CASENet and Shear-DDS have significantly fewer parameters than their non-shearlet counterparts.

In Figure 8.5, we depict the results obtained using an example of the SBD dataset. It shows the semantic edges obtained by both CASENet and DDS and their respective shearlet extension. In all the cases the airplane in the picture was correctly classified, but the refinement of the obtained edges is improved in the shearlet version. This strongly suggests that the use of shearlets is well-suited for high performance in semantic edge detection.

Method	**MF-score**
DSN[121]	65.2
SEAL[122]	75.3
Classical CASENet[121]	71.4
Classical DDS[79]	78.6
Shear-CASENet	**75.7**
Shear-DDS	**80.1**

Table 8.5: Semantic edge detection performance on the SBD dataset. All values are in percentage.

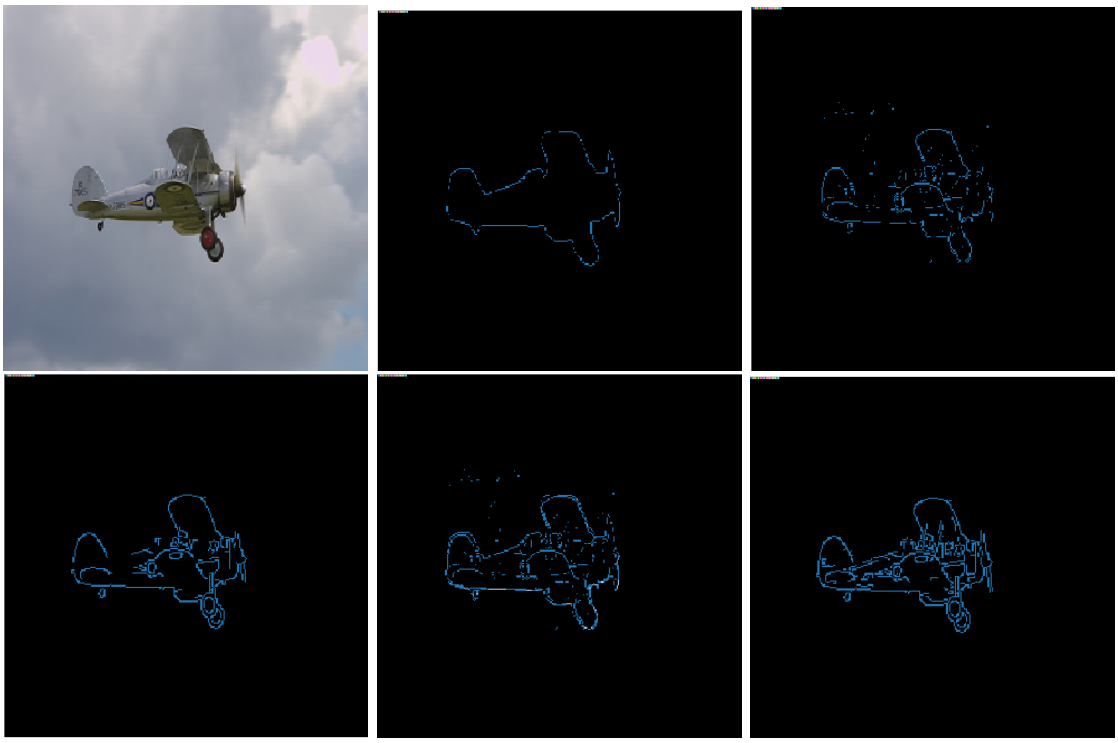

Figure 8.5: Computed semantic edges of an example of the semantic boundaries dataset (SBD). The color blue represents the category of airplane. Top-left: Input image. Top-middle: Semantic edges, human annotation. Top-right: Semantic edges predicted by the classical CASENet architectures. Bottom-left: Semantic edges predicted by the classical DDS architecture. Bottom-middle: Semantic edges predicted by the Shear-CASENet architecture. Bottom-right: Semantic edges predicted by the Shear-DDS architecture.

8.3 Digital microlocal analysis of the learned primal dual

We have implemented the deterministic rules described in Section 6.3, in order to propagate the singularities of the input along the learned primal-dual layers. This also allows us to show the gain in running time with respect of the plain computation of the wavefront set of the output with DeNSE. For the experiments we have trained a modified version of the learned primal-dual algorithm, where we are using as activation function the ReLU non-linearity and no bias in the convolutional layer. Figure 8.6 shows the architecture.

The particular learned-primal dual architecture that we used has ten iterations and dual and primal inputs of five channels each. We trained the network on a set of phantoms composed of random cartoon-like images. Each cartoon-like image is composed of smooth shapes with piece-wise smooth boundaries. In these images, each smooth piece is a spline with the order at most three. Splines are smooth curves or certain order that intersect specific points in the plane, also known as knots and we can define them as follows.

Definition 8.3.1. *Let $n \in \mathbb{N}$ be the order of the spline, and m be the number of knots, namely $t_0 \leq t_1 \leq \ldots \leq t_{m-1}$, a $B-spline$ of order n is the parametric curve*

$$S : [t_0, t_{m=1}] \to \mathbb{R}^2$$

composed by a linear combination of basic $B-$splines $b_{i,n}$ of order n

$$S(t) = \sum_{i=1}^{m-n-2} P_i b_{i,n}(t), \quad t \in [t_{n-1}, t_{m-n}],$$

where P_i are the control points*. There are $m-(n+1)$ control points that form the convex haul. The $m-(n+1)$ basic B-splines of order n can be defined by the formula*

$$b_{j,0}(t) := \begin{cases} 1 & \textit{if } t_j \leq t < t_{j+1}, \\ 0 & \textit{otherwise}, \end{cases}$$

$$b_{j,n}(t) := \frac{t - t_j}{t_{j+n} - t_j} b_{j,n-1}(t) + \frac{t_{j+n=1} - t}{t_{j+n+1} - t_{j+1}} b_{j+1,n-1}(t).$$

Since we know the analytical expression of each boundary, this gives us access to its analytical wavefront set.

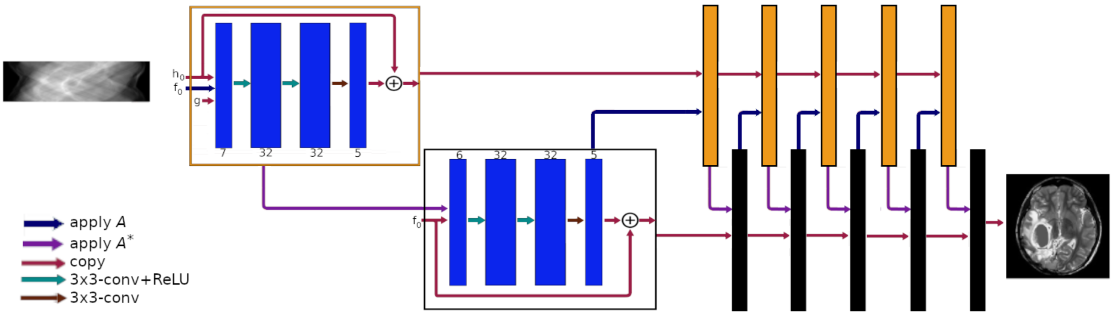

Figure 8.6: Learned primal-dual architecture with sinogram and reconstruction.

Using the digital canonical relation on the wavefront set we can also find the analytical wavefront set of its fully sampled sinogram. Although in the practical case, the data consists of low sampled sinograms, where the visible angles can also be obtained via the digital canonical relation. We depict an example of this image and its wavefront set in Figure 8.8. In addition, one can obtain a subset of the wavefront set of the image by using the microcanonical relation in the wavefront set of the low sampled sinogram. In this context, we refer as a low sampled sinogram to either the low dose setting, with 40 measured angles equally spaced, or the limited-angle setting with a wedge of 80 degrees. This can be done with no need for any reconstruction and represents a strong prior for the reconstruction, this is depicted in Figure 8.7.

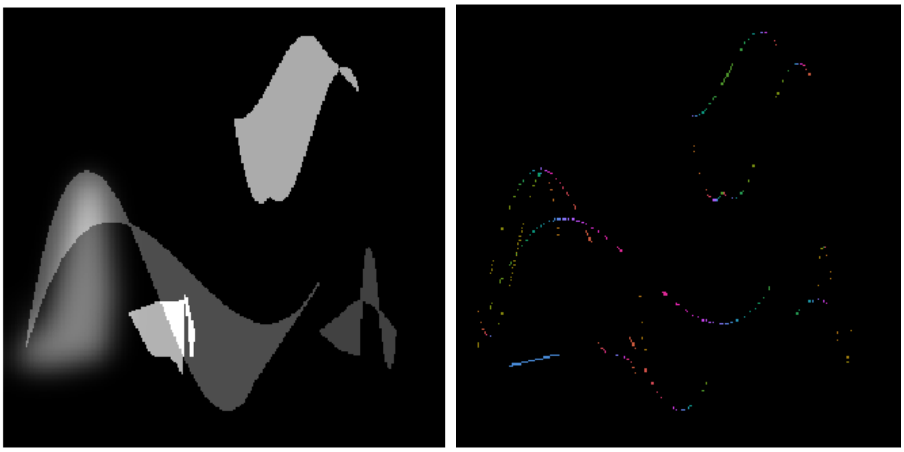

Figure 8.7: Low-dose wavefront set of the image (left) obtained with the digital canonical relation acting on the wavefront set of the low-dose sinogram (left).

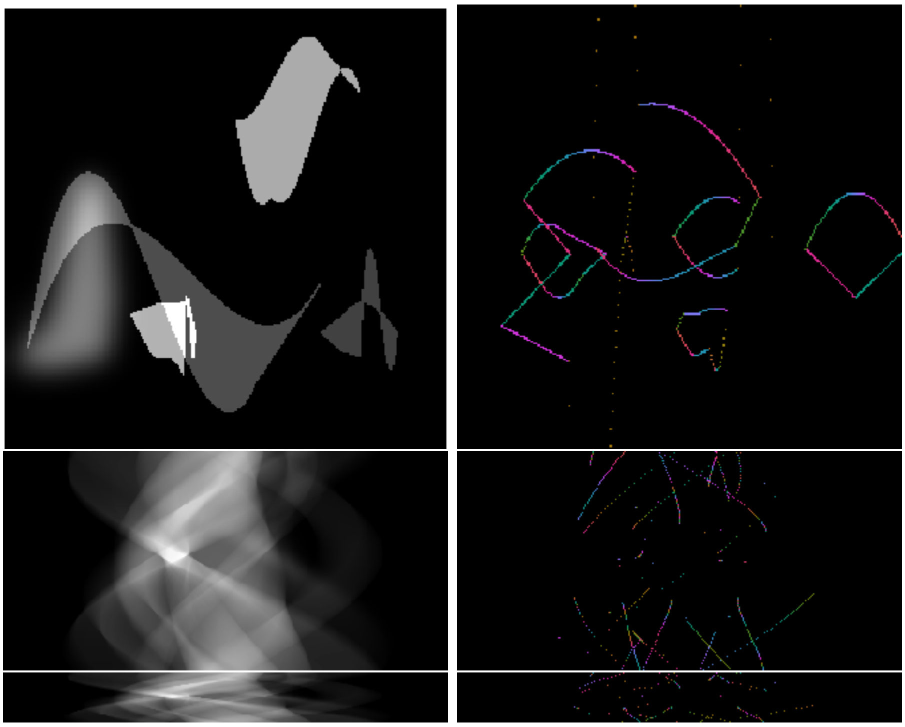

Figure 8.8: Example of cartoon-like image in training set (top-left), its analytical wavefront set (top-right), and the fully sampled sinogram (middle-left) with the corresponding wavefront set (middle-right) computed via the digital microcanonical relation. We also depict the low-dose sinogram (bottom-left), with a dose of 40 measured degrees and its low sampled wavefront set (botom-right).

We compute the ellipticity level of the convolutional operators with the next formula

$$\mathbb{E}(\mathcal{K}_{\boldsymbol{\theta}}) = \min_{i,j\in\{1,\ldots,N\}} |p_{\boldsymbol{\theta_k}}[i,j]|$$

where $p_{\boldsymbol{\theta}_k}[i,j]$ is the amplitude of the operator given by Equation (2.3.9). In other words, we evaluate the absolute value of the amplitude in the Fourier grid and we take the minimum. Therefore, if $\mathcal{E}(\mathcal{K}_{\boldsymbol{\theta}_k}) > 0$, the convolutional operator associated to the layer is elliptic and preserves the singularities (see Figure 8.9). In the next sections, we will present the results on the wavefront set propagation obtained on the distinct primal and dual iterations of the learned-primal dual architecture.

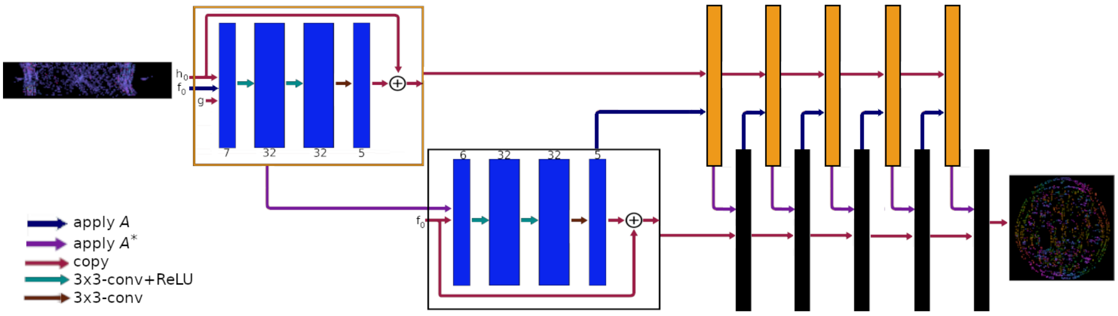

Figure 8.9: Learned primal-dual architecture with the corresponding wavefront sets.

8.3.1 Primal and dual iterations

In Figure 8.10, 8.11 and 8.12, we present the output of the first, sixth, and tenth iteration of the trained primal-dual reconstruction and their wavefront sets obtained with the canonical relation of the layers discussed in Section 4.3. These figures are relevant to present since they are evidence that the theory of digital wavefront set propagation in Section 6.3 holds in real-world applications. One can notice by comparing with the output images of each layer, that the propagated wavefront set is accurate. In addition, these results will be later used on Section 8.4 in the context of task-adapted tomographic reconstruction and wavefront set inpainting, in order to recover the invisible part of the wavefront set from the propagated visible part, just as presented theoretically in Section 7.5.2.

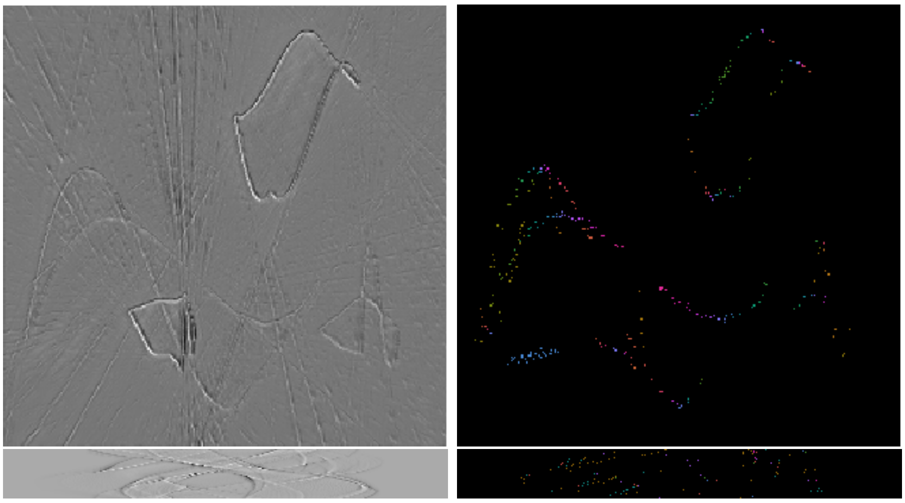

Figure 8.10: Output of the 1st iteration of the primal (upper-left) and dual (upper-right) subnetworks and its wavefront set obtained from the canonical relations presented above. Their wavefront sets are depicted on the right side of each image.

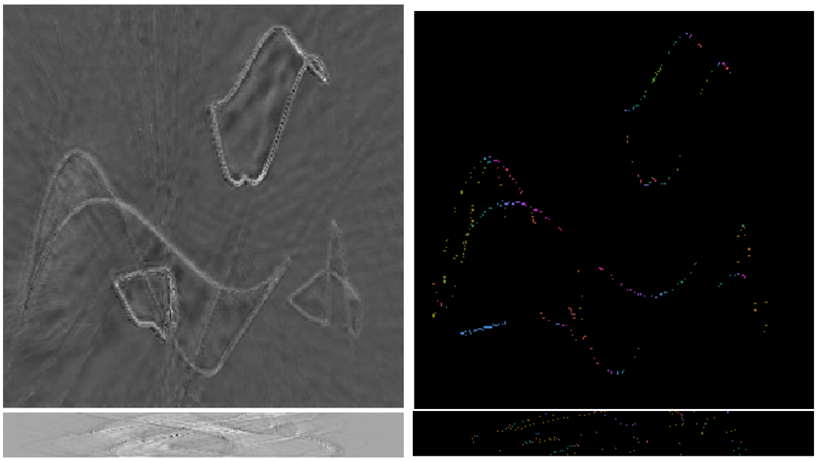

Figure 8.11: Output of the 6th iteration of the primal (upper-left) and dual (upper-right) subnetworks and its wavefront set obtained from the canonical relations presented above. Their wavefront sets are depicted on the right side of each image.

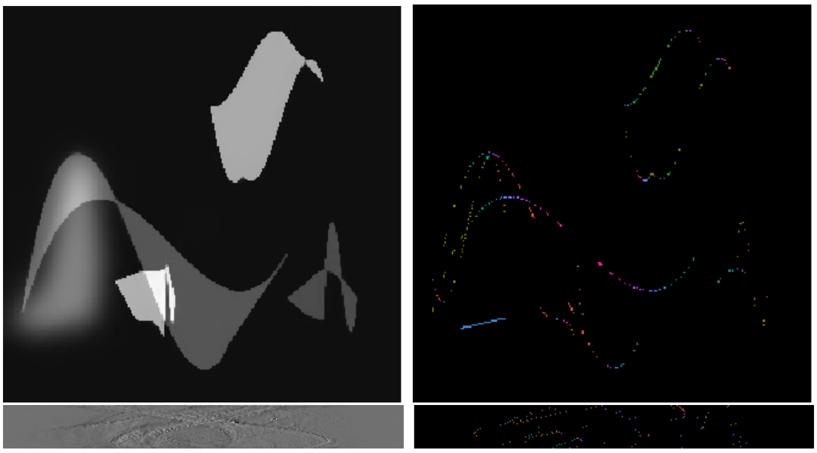

Figure 8.12: Output of the 10th iteration of the primal (upper-left) and dual (upper-right) subnetworks and its wavefront set obtained from the canonical relations presented above. Their wavefront sets are depicted on the right side of each image.

8.3.2 Inner loop of dual and primal iterations

In this section we present the results obtained on the inner loop in the tenth layer of the trained learned primal-dual reconstruction, in both the primal and dual step, this corresponds to the theory presented in Section 6.3. Figure 8.13 shows the output of the first, second, and third convolution on the primal and dual inner loop. Figure 8.14 shows the wavefront set of the output in the first convolutional layer of the primal an dual iteration. In Figures 8.15 and 8.16 we present the output of the Heaviside function applied to the first and second convolutional layer, and its wavefront set obtained with the microcanonical relation of the Heaviside function, correspondingly. Finally, Figures 8.15 and 8.16 show the same results, but for the ReLU function. Notice in addition that the images corresponding to the sinogram and dual steps and their wavefront set are small. This is due to the small number of measured angles, 40 degrees to be exact. The size makes them hard to interpret, but on the other hand, the lowest dose is always the best for the potential human patient. The figures show an accurate propagation of the wavefront set through the inner-loop layers.

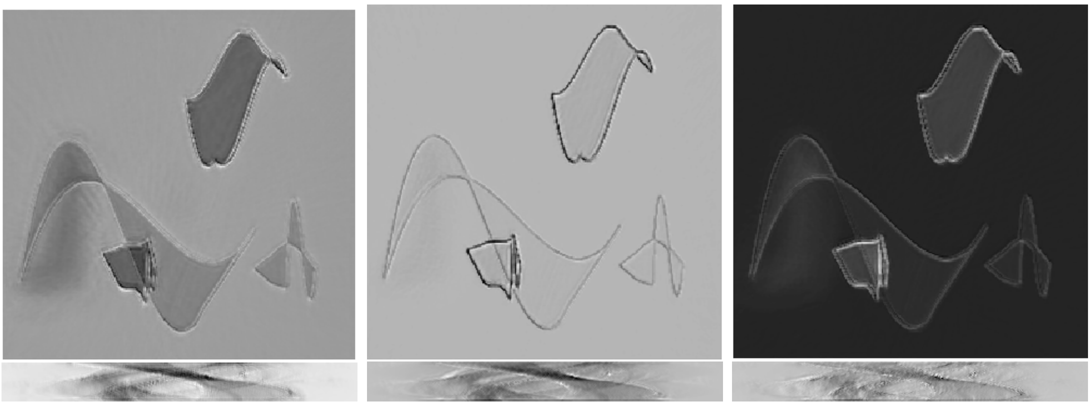

Figure 8.13: Output of the first, second and third convolutional layers (from left to right) of the primal and dual inner loops (from up to bottom) of the 10th iteration layer of the LPD.

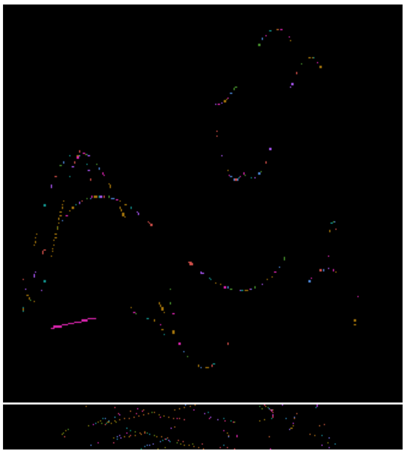

Figure 8.14: The wavefront set of the output of the output of first convolutional layer on the 10th iteration.

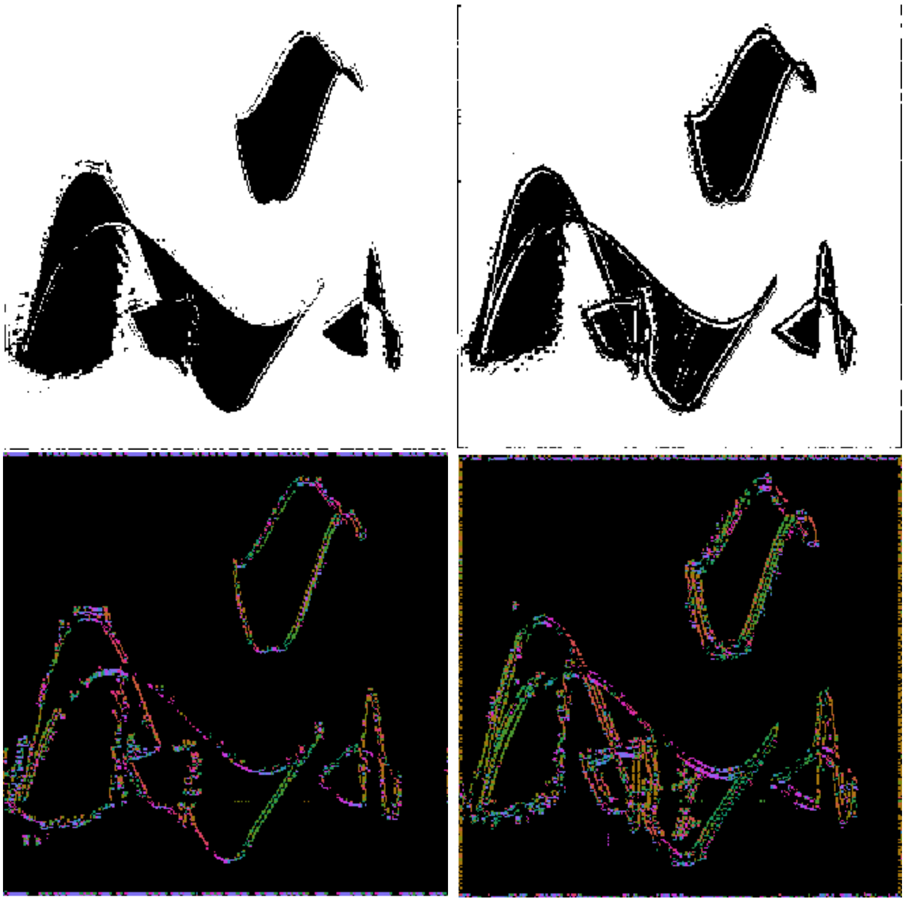

Figure 8.15: Output and wavefront set of the Heaviside function applied to the first and second convolution layer of the primal 10th iteration (up). The wavefront sets (bottom) were computed using the microcanonical relation of the Heaviside function.

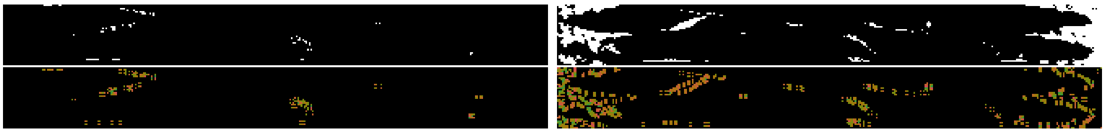

Figure 8.16: Output and wavefront set of the Heaviside function applied to the first and second convolution layer of the dual 10th iteration (up). The wavefront sets (bottom) were computed using the microcanonical relation of the Heaviside function.

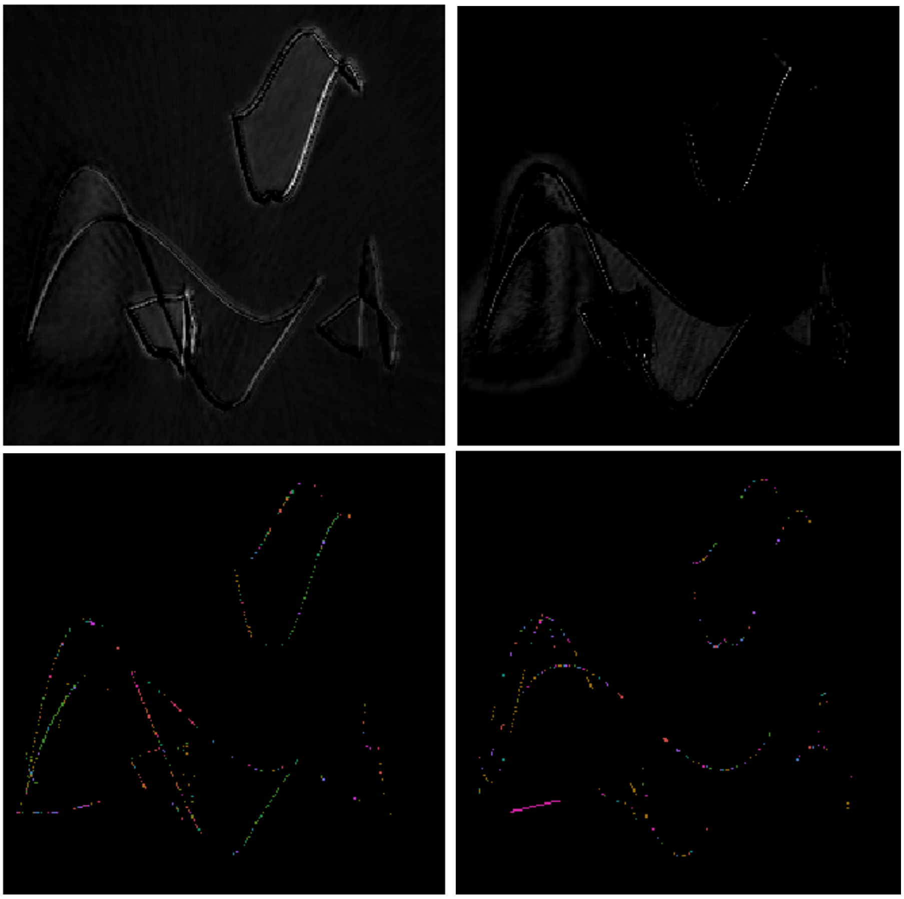

Figure 8.17: Output and wavefront set of the ReLU function applied to the first and second convolution layer of the primal 10th iteration (up). The wavefront sets (bottom) were computed using the microcanonical relation of the ReLU function.

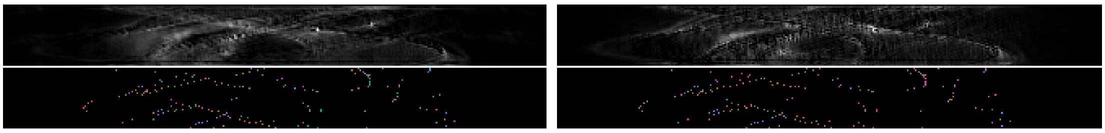

Figure 8.18: Output and wavefront set of the ReLU function applied to the first and second convolution layer of the dual 10th iteration. The wavefront set was computed using the microcanonical relation of the ReLU function.

8.4 Task-adapted reconstruction using digital wavefront sets

In Chapter 7 we explored the framework of task-adapted reconstruction. Our contribution to this framework is based on the use of the wavefront set as a-priori information. In this context, we introduced the tasks of wavefront set extraction (Section 7.5.1) and wavefront set inpainting (Section 7.5.2), the latter being the best of them. In this section, we present the results on both approaches, as well as the results concerning the novel task of wavefront set inpainting.

8.4.1 Wavefront set inpainting

As discussed in Section 7.5.2, we used a U-Net architecture (see Figure 1.1) to inpaint the low dose wavefront set, training on different datasets, one formed by random ellipses and other formed by more realistic phantoms. The realistic phantoms are formed by piece-wise smooth functions with piece-wise smooth boundaries defined with splines with order at most four (as shown in the last section). In the following, we present the results, in every experiment we used a low dose of 40 measured angles (equally spaced). Figure 8.19 shows the results obtained when trained on random ellipses. Figure 8.22 shows the results with a training set formed by realistic ellipses. Finally, Figure 8.23 shows how both models predict an unseen example, the Shepp-Logan phantom, it is clear that the model trained on realistic phantoms performs the best.

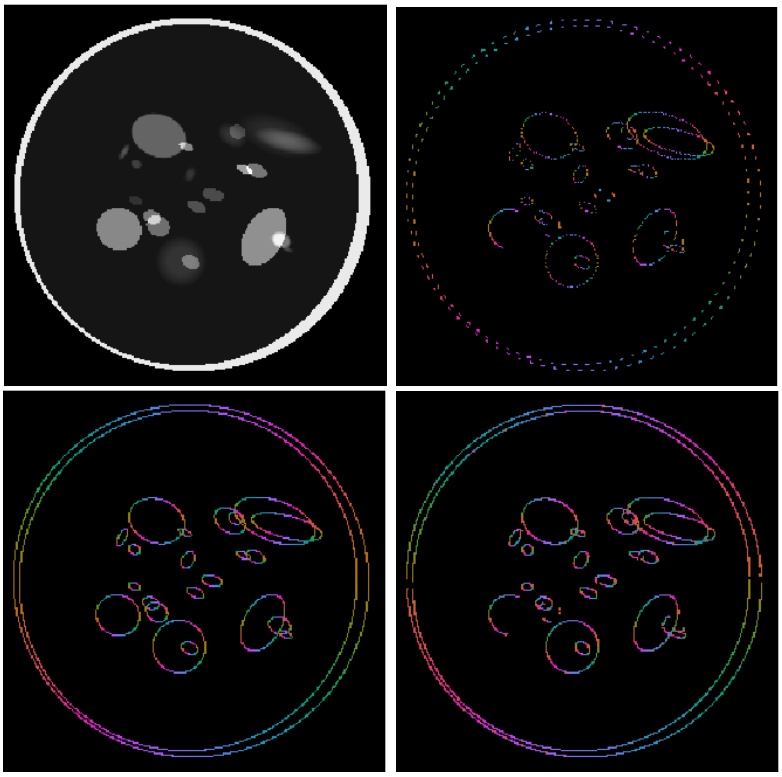

Figure 8.19: An training example of the random ellipses phantom dataset. Top-left: Phantom. Top-left: Ground truth low-dose wavefront set. Bottom-left: Ground truth full-dose wavefront set. Bottom-right: Predicted wavefront set with U-Net.

8.4.2 Task-adapted tomographic reconstruction and wavefront set extraction

As presented in Section 7.5.1, we jointly trained the learned primal-dual reconstruction and the wavefront set extractor architectures, using different constants in the sense of task-adapted reconstruction. We trained the task-adapted architecture training set with random ellipses and with realistic phantoms, individually. Figure 8.20 depicts the results when the constant $C = 0.1$, which emphasizes the task of wavefront set extraction, rather than the reconstruction. In the results, one can see that the reconstruction is poor, but the extracted wavefront set is better. Figure 8.21 depicts the results when $C = 0.9$, which emphasizes reconstruction, resulting in a better reconstruction. In all these cases we generated a total of 10,000 images with random phantoms, where 7,000 were used for training, 1,000 for validating and 2,000 for testing.

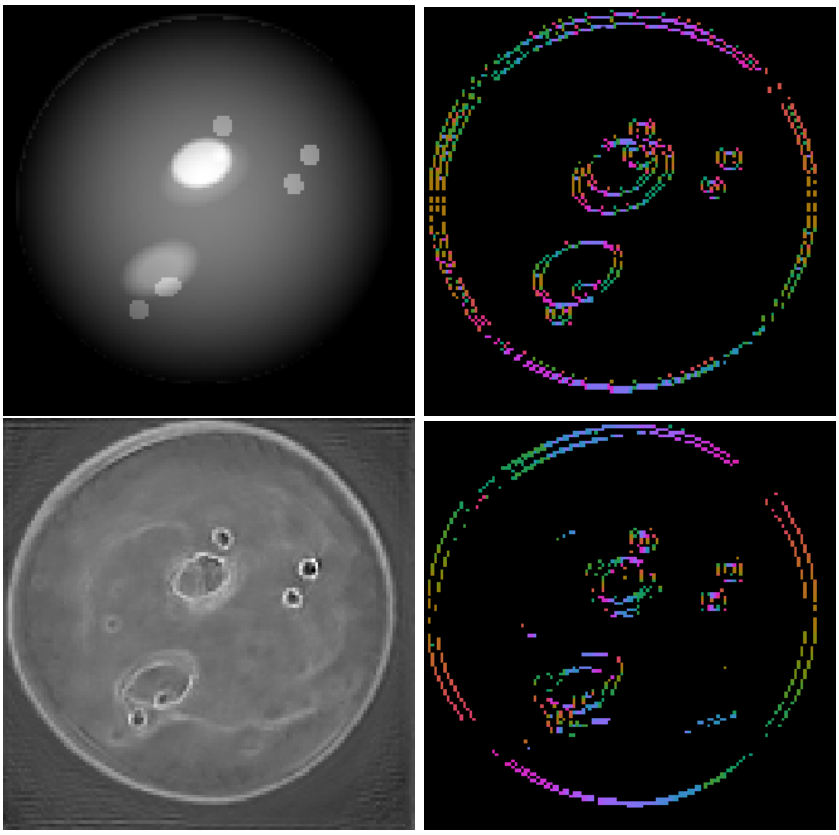

Figure 8.20: Results on the joint CT reconstruction and WFset extraction for constant $C = 0.1$. Top-left: Phantom. Top-left: Ground truth wavefront set. Bottom-left: Reconstruction results. Bottom-right: Wavefront set results.

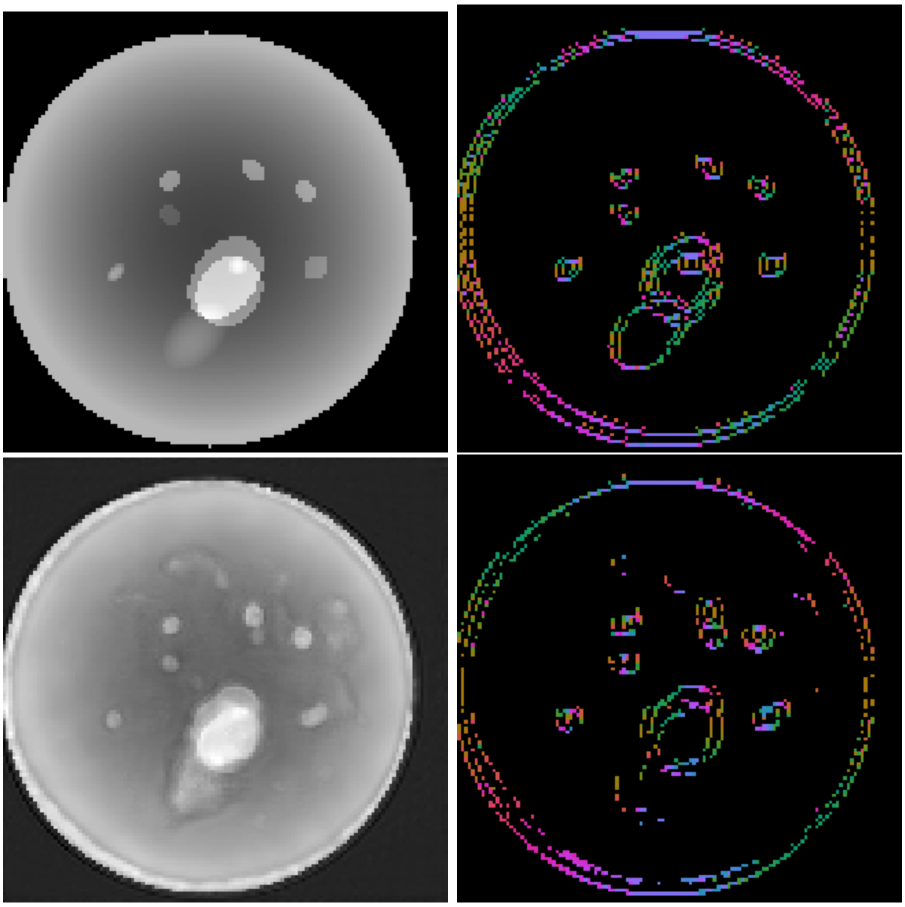

Figure 8.21: Results on the joint CT reconstruction and WFset extraction for constant $C = 0.9$. Top-left: Phantom. Top-left: Ground truth wavefront set. Bottom-left: Reconstruction results. Bottom-right: Wavefront set results.

8.4.3 Task-adapted reconstruction and wavefront set inpainting

We jointly trained the learned primal-dual reconstruction and wavefront set inpainting architecture in the sense of task-adapted reconstruction. As discussed in Section 7.5.2, the wavefront inpainting step involves the wavefront set propagation given by the microlocal analysis of the learned primal-dual presented in this work. Figures 8.24, 8.25, present the results for the case of low dose (40 measured angles) and limited angle (wedge of 40 degrees) in a validation example, respectively. In both cases, the joint approach outperforms the learned primal-dual, being clearer in the limited angle case. In this case, we have trained the model on the ellipses dataset. In addition, we have also trained the models in the realistic dataset based on splines and evaluate them in a real brain CT scan, depicting the results on Figures 8.28 and 8.29.

Figures 8.26 and 8.27 depict general classical benchmarks for the limited angle and low dose tomography, including filtered back projection, Tikhonov, and total variation,

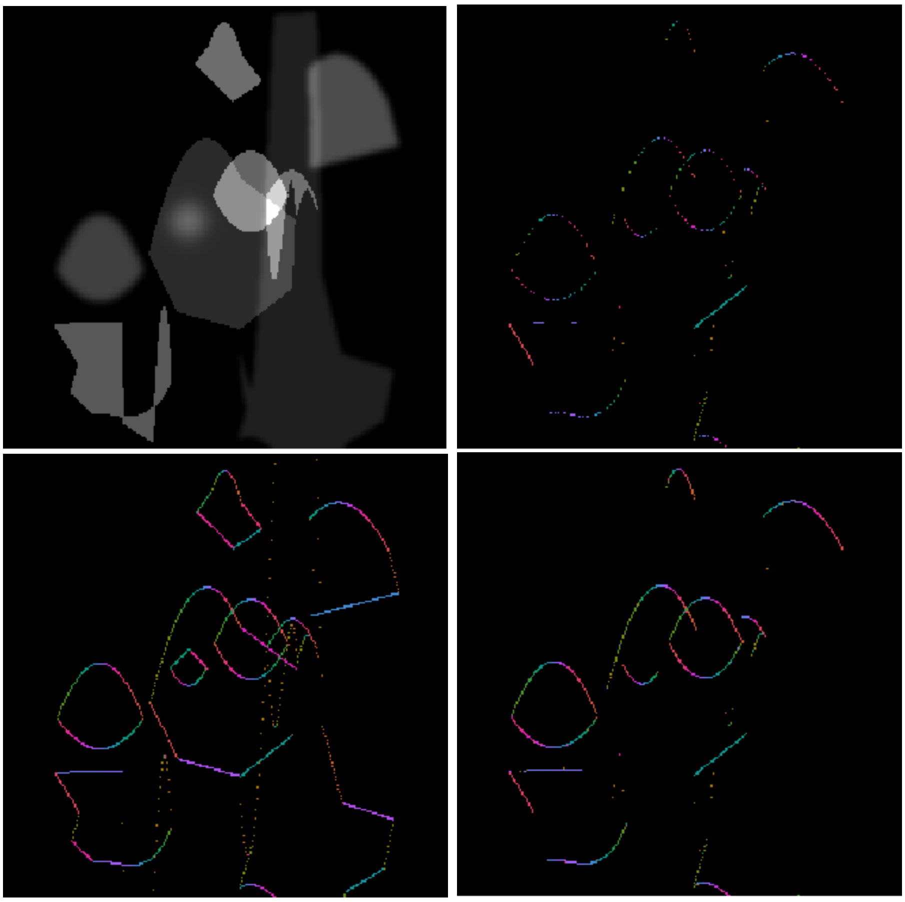

Figure 8.22: An training example of the realistic phantom dataset. Top-left: Phantom. Top-left: Ground truth low-dose wavefront set. Bottom-left: Ground truth full-dose wavefront set. Bottom-right: Predicted wavefront set with U-Net.

both trained on the ellipses dataset. We also present the results of the limited angle case using the Phantom Net architecture (Bubba et al. [17]). In addition, we have also performed reconstructions using as regularization functional both the L_2 and L_1 norm of the shearlet coefficients. We have coined the two methods *Shearlet-L2 sparse* and *Shearlet-L1 sparse*, respectively. In all cases, the joint approach outperforms the other methods. Tables 8.6 and 8.7 respectively present the performance measure in terms of self-similarity (SSIM) and peak signal-to-noise ratio (PSNR). In addition, we have also evaluated the methods on the brain CT scan, where the deep learning approaches were trained on the realistic dataset based on splines. These results are shown in Figures 8.30 and 8.31. Similarly, Tables 8.8 and 8.9 present the performance measure for the realistic dataset evaluation.

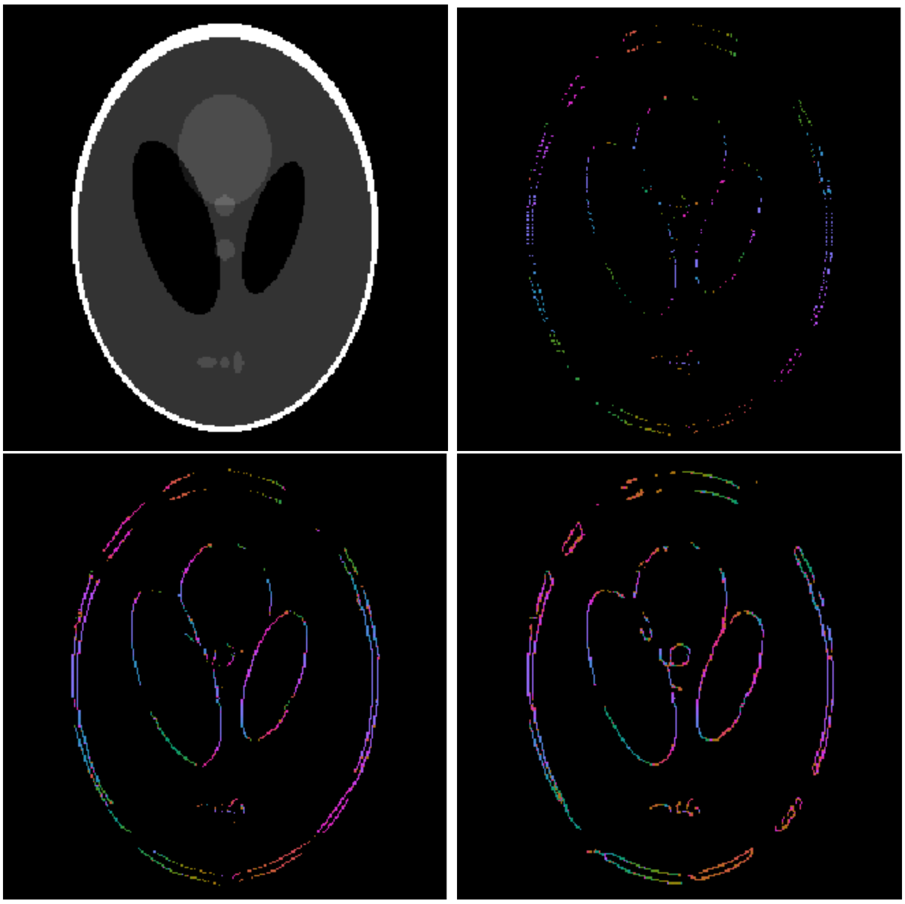

Figure 8.23: Prediction on an unseen example (Shepp-Logan phantom). Top-left: Phantom. Top-left: Ground truth low-dose wavefront set. Bottom-left: Prediction from network trained with realistic phantoms. Bottom-right: Prediction from network trained with random ellipses phantoms.

Method	**SSIM**	**PSNR**
FBP	0.58	20.15
Tikhonov	0.75	26.66
TV	0.90	27.90
Shearlet-L2 sparse	0.76	25.88
Shearlet-L1 sparse	0.83	26.31
LPD	0.91	28.45
Joint approach	**0.98**	**31.12**

Table 8.6: Ellipses dataset performance for general benchmarks for low dose CT.

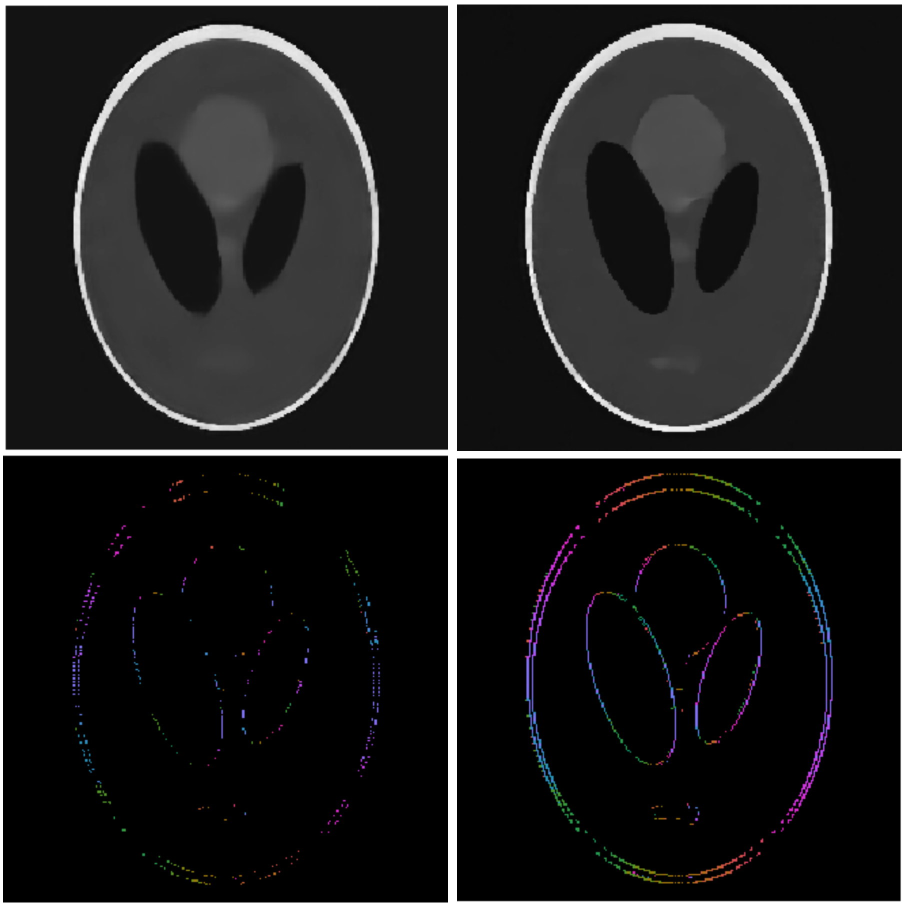

Figure 8.24: Ellipse dataset results on the joint CT reconstruction and WFset inpainting for low dose case, lowd = 40. Top-left: Reconstruction LPD (PSNR 28.51). Top-right: Reconstruction Joint approach (PSNR 32.15). Bottom-left: Low-dose wavefront set. Bottom-right: Inpainted wavefront set.

Method	**SSIM**	**PSNR**
FBP	0.50	17.30
Tikhonov	0.79	21.72
TV	0.83	24.15
Shearlet-L2 sparse	0.80	20.12
Shearlet-L1 sparse	0.82	23.50
PhantomNet [17]	0.92	26.55
LPD	0.90	27.65
Joint approach	**0.97**	**30.33**

Table 8.7: Ellipses dataset performance for general benchmarks for limited angle CT.

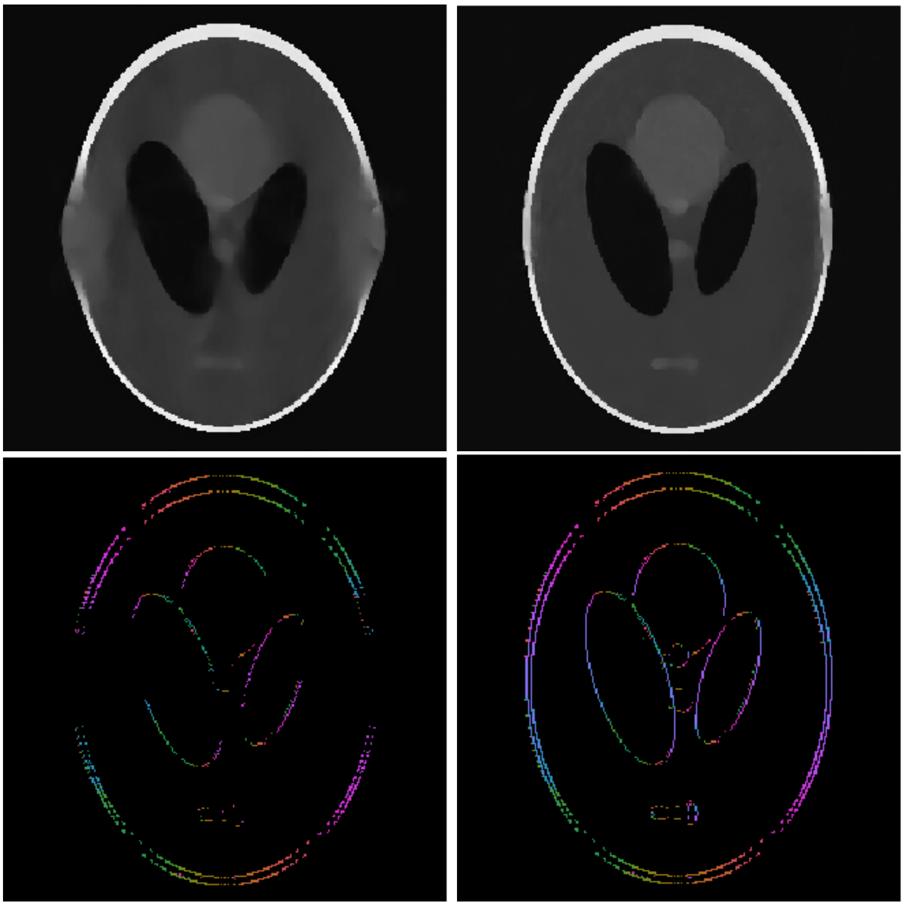

Figure 8.25: Ellipse dataset results on the joint CT reconstruction and WFset inpainting for limited angle case, wedge = 40. Top-left: Reconstruction LPD (PSNR 27.45). Top-right: Reconstruction Joint approach (PSNR 30.24). Bottom-left: Wedge wavefront set. Bottom-right: Inpainted wavefrontset.

Method	**SSIM**	**PSNR**
FBP	0.51	19.90
Tikhonov	0.73	24.77
TV	0.88	26.59
Shearlet-L2 sparse	0.73	24.69
Shearlet-L1 sparse	0.78	25.42
LPD	0.89	27.55
Joint approach	**0.92**	**31.46**

Table 8.8: Realistic dataset performance for general benchmarks for low dose CT.

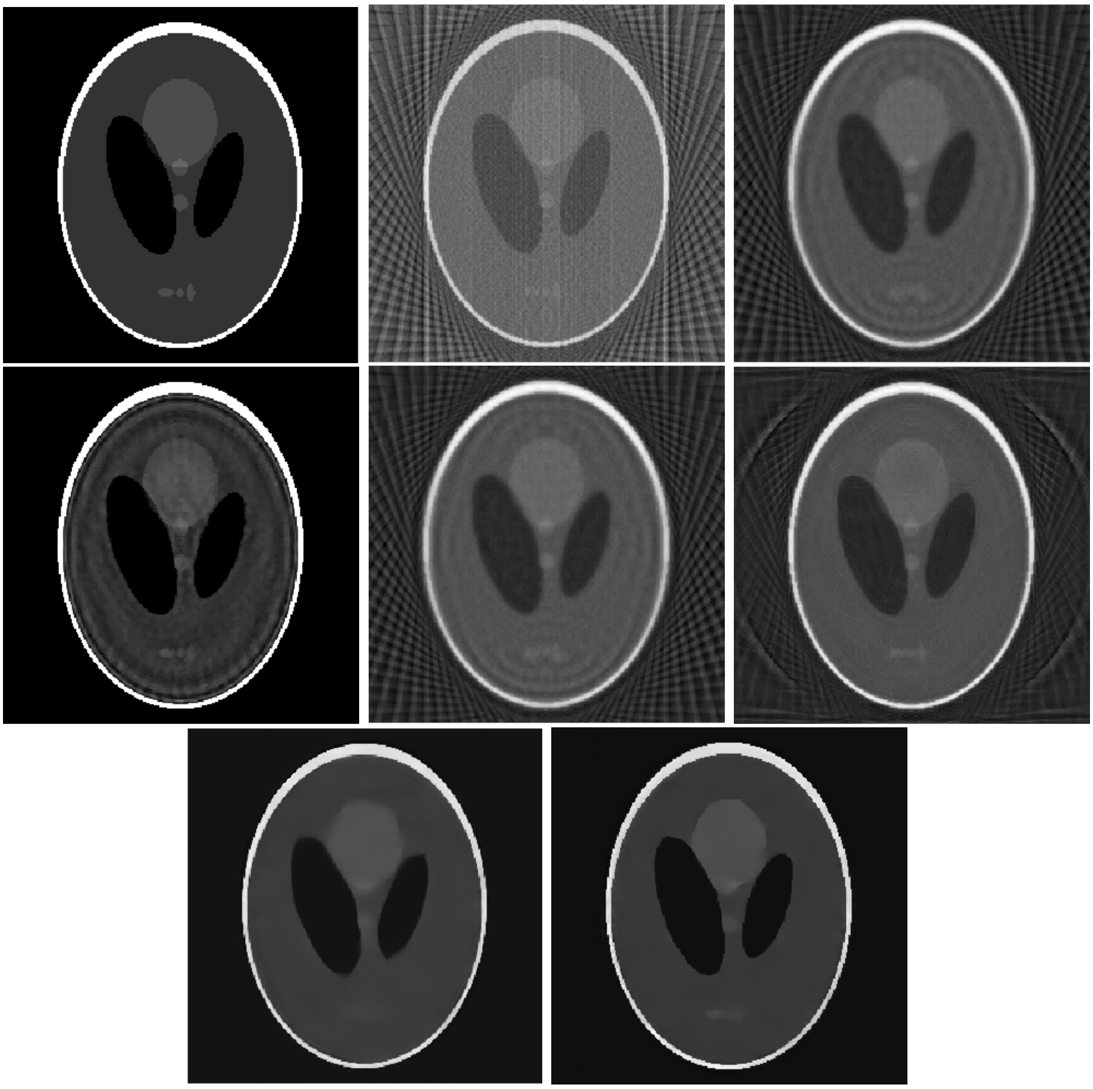

Figure 8.26: Ellipse dataset results for general benchmarks for low dose CT, lowd = 40. 1st row: Ground truth (left), FBP (right). 2nd row: Tikhonov (left), TV (right). 3rd row: Shearlet-L2 sparse (left), Shearlet-L1 sparse (right). 4th row: LPD (left), Joint approach (right).

Method	**SSIM**	**PSNR**
FBP	0.44	14.53
Tikhonov	0.73	22.62
TV	0.83	23.09
Shearlet-L2 sparse	0.70	22.20
Shearlet-L1 sparse	0.76	22.29
PhantomNet [17]	0.87	25.50
LPD	0.86	25.55
Joint approach	**0.95**	**29.80**

Table 8.9: Realistic dataset performance for general benchmarks for limited angle CT.

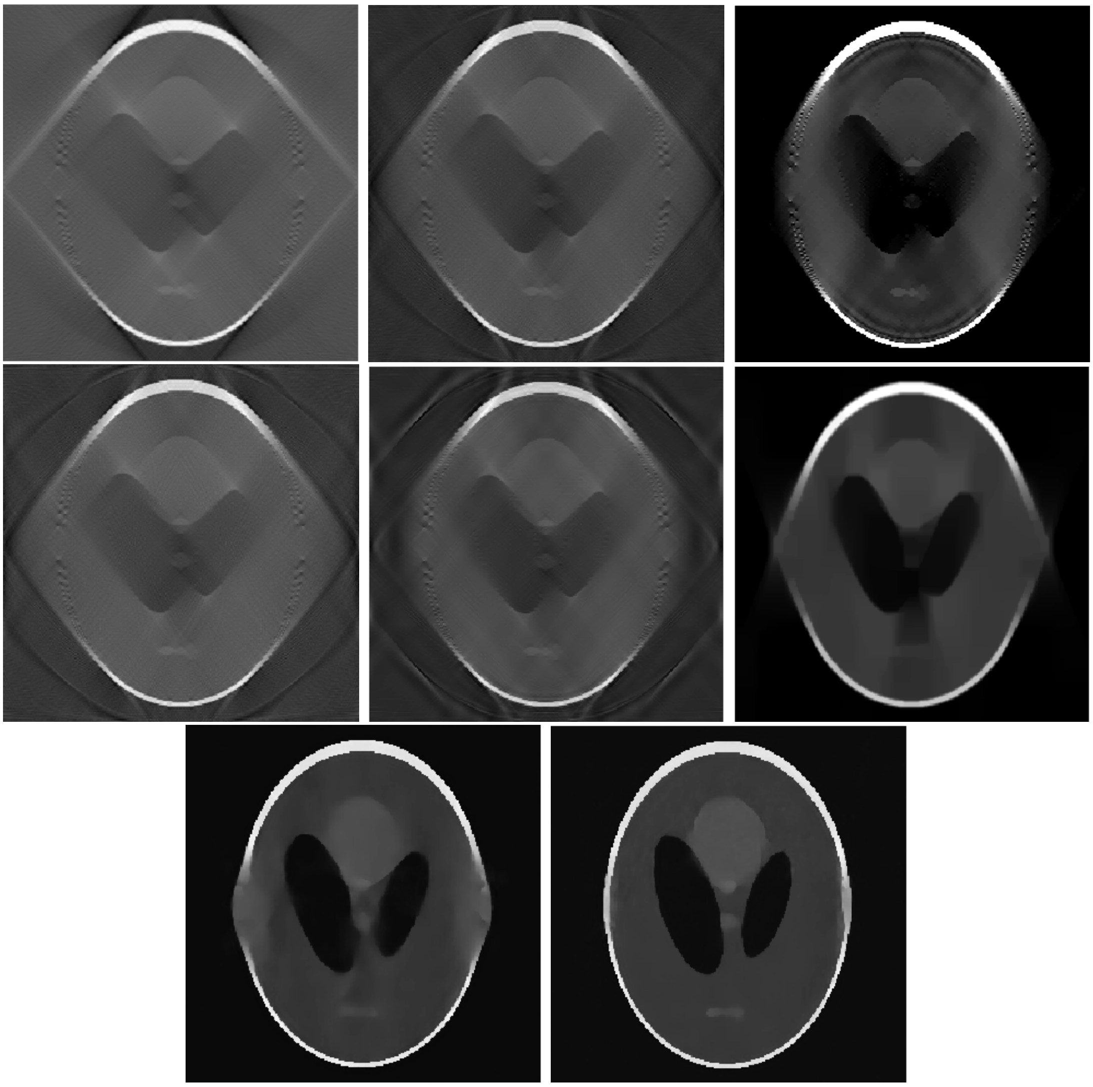

Figure 8.27: Ellipse dataset results for general benchmarks for limited angle CT, wedge = 40. 1st row: FBP (left), Tikhonov (right). 2nd row: TV (left), Shearlet-L2 sparse (right). 3rd row: Shearlet-L1 sparse (left), PhantomNet [17] (left). 4th row: LPD (right), Joint approach (right).

As one can observe, the results presented in Figures 8.26 and 8.27 represent a significant improvement on the reconstruction in comparison with the other methods. In particular, the addition of the wavefront set inpainting improves significantly the performance of the learned primal-dual architecture. We can also see that the improvement of performance extends to the real-data (Figures 8.30 and 8.31). This tells us that the potential of our method to be applied in real case scenarios is very promising. We are going to discuss this and other final remarks and conclusions in the next and final chapter.

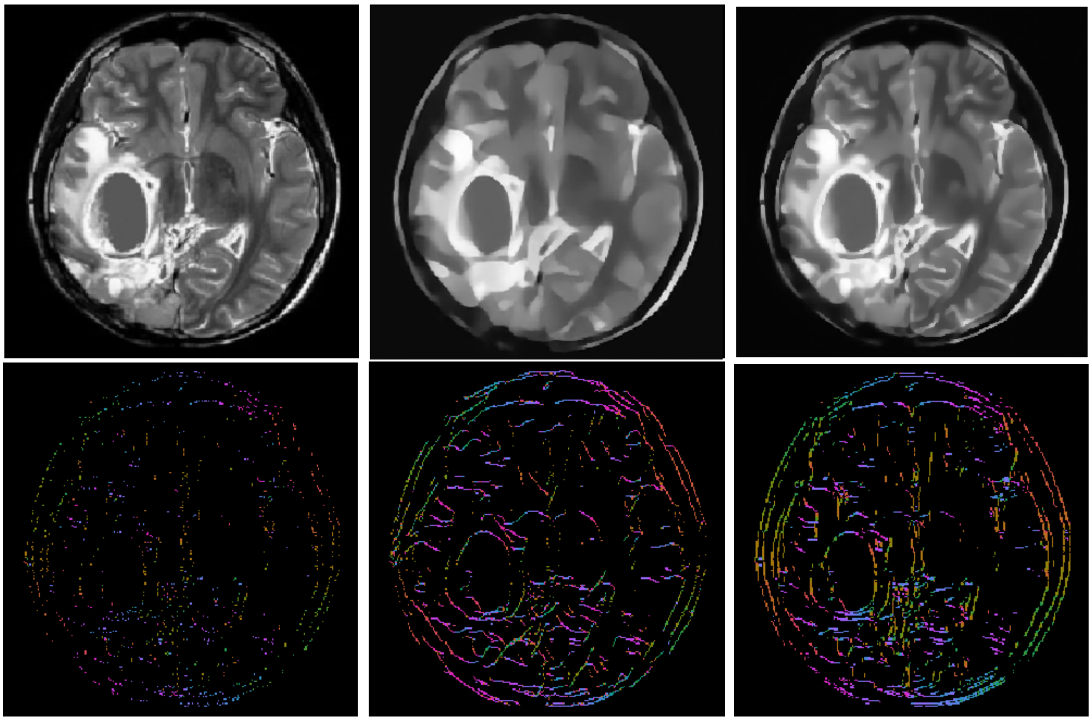

Figure 8.28: Realistic dataset results on the joint CT reconstruction and WFset inpainting for low dose case, lowd = 40. Top left: Ground truth. Top middle: Reconstruction using the learned primal dual algorithm without using the wavefront set information as prior (PSNR 26.91). Top right: Reconstruction using the joint approach introduced in this work (PSNR 30.22). Bottom left: Visible wavefront set of the ground truth extracted by DeNSE. Bottom middle: Wavefront set inpainted using U-Net. Bottom right: Reconstructed wavefront set using our joint approach.

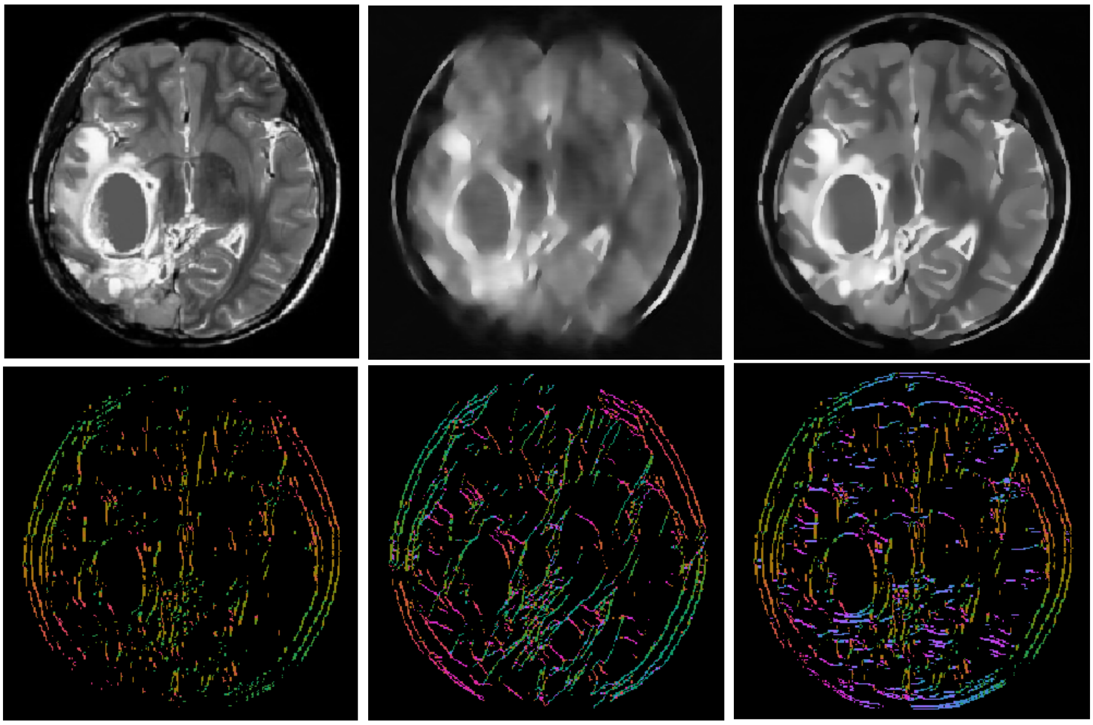

Figure 8.29: Realistic dataset results on the joint CT reconstruction and WFset inpainting for limited angle case, wedge = 40. Top left: Ground truth. Top middle: Reconstruction using the learned primal dual algorithm without using the wavefront set information as prior (PSNR 24.90). Top right: Reconstruction using the joint approach introduced in this work (PSNR 30.20). Bottom left: Visible wavefront set of the ground truth extracted by DeNSE. Bottom middle: Wavefront set inpainted using U-Net. Bottom right: Reconstructed wavefront set using our joint approach.

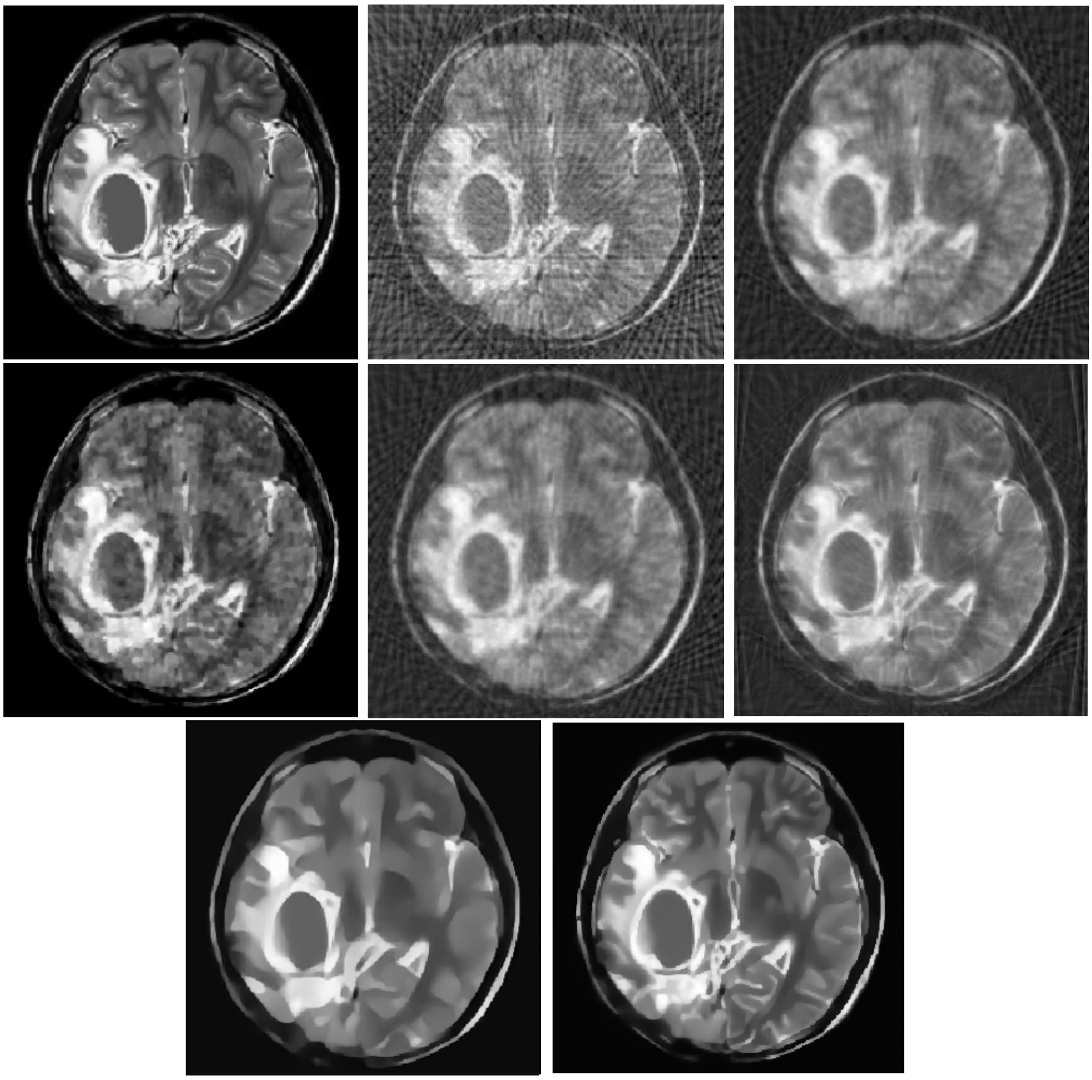

Figure 8.30: Realistic data set results for general benchmarks for low dose CT, lowd = 40. 1st row: Ground truth (left), FBP (right). 2nd row: Tikhonov (left), TV (right). 3rd row: Shearlet-L2 sparse (left), Shearlet-L1 sparse (right). 4th row: LPD (left), Joint approach (right).

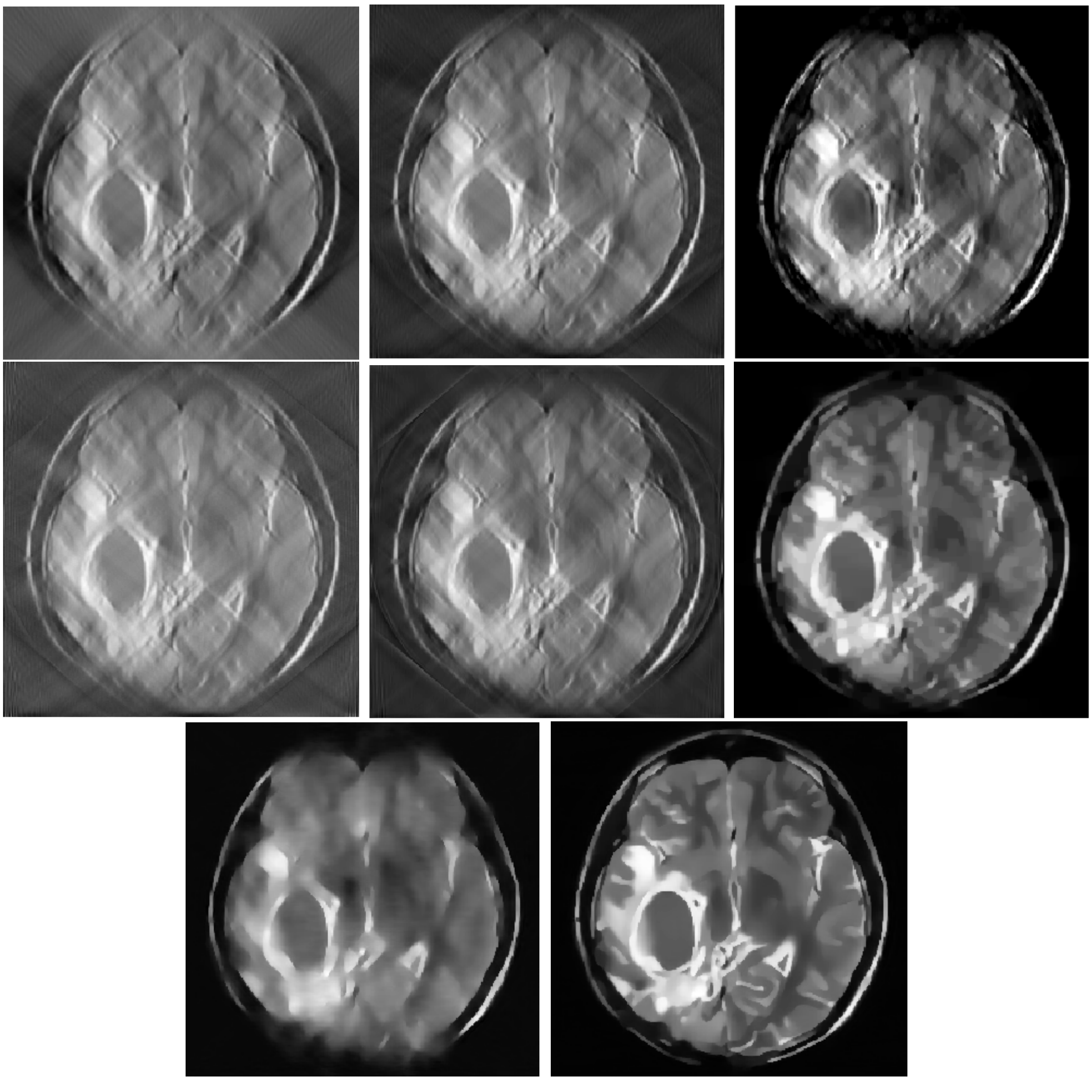

Figure 8.31: Realistic dataset results for general benchmarks for limited angle CT, wedge = 40. 1st row: FBP (left), Tikhonov (right). 2nd row: TV (left), Shearlet-L2 sparse (right). 3rd row: Shearlet-L1 sparse (left), PhantomNet [17] (left). 4th row: LPD (right), Joint approach (right).

9 Conclusion and Outlook

In this final chapter, we present the summary of the thesis as well as a conclusion. In addition, we also want to discuss future promising advances and applications that can take advantage of the framework presented in this thesis. We have introduced in Chapter 1 the basic notions of microlocal analysis, deep learning, and inverse problems. In particular, Section 1.1 discusses the impact of microlocal analysis in inverse problems in imaging, where the microcanonical relation plays an important role – it characterizes the wavefront set of reconstructions. In addition, Section 1.3 presents a diverse set of deep neural network architectures designed to solve inverse problems, from which the learned primal-dual algorithm [3] is the most relevant in our application. These two ingredients in addition to the task-adapted reconstruction form the backbone of our work.

The main contribution of microlocal analysis to the realm of inverse problems is the wavefront set, which characterizes oriented singularities of distributions. Throughout the thesis, we have learned that the wavefront set plays a central role in the morphological formation of images. In addition, for Fourier integral operators, which are abundant in inverse problems, one can characterize the microcanonical relation in an explicit form, as seen in Section 2.4.

In order to use the wavefront set in real-world scenarios, we first need to have a way to extract it explicitly. The continuous shearlet system presented in Section 3.4 is useful for this purpose. Continuous shearlets allow us to characterize the wavefront set in terms of the decay rate of shearlet coefficients. This characterization still depends on the asymptotic analysis of such decay rate, which cannot be achieved in the digital realm. This happens since in digital domains one just have access to finitely many (rather a small number of) coefficients. In this case, the use of deep learning to find patterns on digital shearlet coefficients, and therefore approximate the wavefront set, is almost necessary. In Section 5.4, we introduced the DeNSE algorithm that performs this task with high accuracy.

Although the concept of wavefront set is interesting on its own, it is useful in practice when used to boost the performance on methods for inverse problem regularization. Since recently most inverse problems are solve, at least partially, by the use of neural networks, it is important to study how neural networks act on wavefront sets. In the case of convolutional ResNets, this analysis arises naturally, since these architectures can be regarded as operators that act on spaces of the same dimension. Chapter 4 is dedicated to the study of the microlocal analysis of convolutional ResNets, aiming at characterizing their microcanonical relation. For this purpose, we needed to extend the notion of microcanonical relation to the nonlinear case in order to use activation functions. We have also rewritten ResNets as operators acting on continuous spaces. The main reason for this is the fact that a wavefront set is uniquely defined in the continuous setting. Later on, we have also discretized this analysis in Chapter 6 using the shearlet transform to first obtain a sparse representation of Fourier integral operators and then digitize them while ensuring fast convergence of such digitization. This allowed us to explicitly define the microcanonical relation in the digital case, and to use it for real-world applications. This also allows us to characterize the propagation of singularities under the action of convolutional ResNets. In particular, in Section 6.3 we have presented the microcanonical relation of the different layers present in the learned primal-dual algorithm.

The microcanonical relation of the learned primal-dual algorithm is handy when we are trying to use the wavefront set as prior for tomographic reconstruction. The most natural framework where this arises is the framework of task-adapted reconstruction. In task-adapted reconstruction, one aims to jointly perform an image reconstruction from an inverse problem and adapt it to a decision-making procedure. This is important in the context of biomedical imaging, since the ultimate goal in applications in medicine is making a decision that affects the patient's health. On the other hand, the framework of task-adapted reconstruction can be used as a method to jointly train a reconstruction and a task, where the task is not directly applied in the decision making for medical purposes. The task can also be a way to improve the reconstruction, in other words, a regularizer. In that context, the final product of the microlocal analysis of neural networks, presented in this thesis, is its application to tomographic reconstruction. This application uses the learned primal-dual architecture as a reconstruction operator, and a U-Net architecture to inpaint the wavefront set as a task architecture.

In Section 7.5.2, we presented the general setting of this approach. In this setting, we make use of the microcanonical relation of the learned primal-dual presented in Section 6.3 to propagate the sparsely sampled wavefront set of the tomographic data through the network. The product of this propagation is a sparse wavefront set on the image domain, with no reconstruction needed. Later this sparse wavefront set is inpainted via the U-Net to a densely sampled wavefront set. The joint training of the U-Net and the learned primal-dual allows us to force the reconstruction to have a wavefront set close to its ground-truth. This improves the original reconstruction provided by the classical learned primal-dual architecture, which is the previous state-of-the-art. Our method outperforms widely used model and data-driven reconstruction methods. The numerical experiments that back this statement were presented in Chapter 8.

Finally, we believe that our framework, as general as it is, has great potential in distinct applications of biomedical imaging. In the near future, we aim to apply our method to different real-world problems including MRI and EEG. We also think that general task-adapted reconstruction will very soon become common in medical applications. In addition, the lack of commercial software based on wavefront set extractors demonstrates that the area of applied microlocal analysis is still unexplored.

Being able to characterize the propagation of singularities performed by deep neural networks has a clear impact on the theoretical understanding of the field. In this context, one possible research direction is the study of singularities approximation under deep conv-ResNets, in order to characterize the approximation capabilities of the network itself. In addition, we also have encountered some challenges along our journey, which are planned to be addressed soon. The first of these challenges is the difficulty to analyse the microlocal behavior of non-residual neural networks. This is mainly due to the change in the dimension between the input and output spaces of such networks. This does not allow us to interpret neural networks as operators in the continuous setting. In addition, convolutions are pseudodifferential operators, which makes the microlocal analysis simpler. More general neural networks, such as fully connected networks, transformers and recurrent neural networks lack this property. Therefore, the analysis of these architectures needs to be done with a different approach.

In biomedical imaging, the lack of annotated real-data available is a significant problem. In our case, this lack is mainly due to the fact that it is humanly impossible to annotate the wavefront set of natural images. We have tackled this issue by designing a data simulation procedure, that generates realistic phantoms formed by splines of degree at most four. In this phantom one is able to define the wavefront set analytically, but it is far from being a simulation of real-world images. This impacts our results in the sense that our reconstruction has a *cartoon-ish look*, which is of course not ideal. We plan to work on the simulation of more realistic datasets in the future, using tools provided by modern computer graphics techniques. Sophisticated rendering techniques will allow us to obtain close to realistic imagery with a normal map explicitly defined. This normal map will be handy for defining the corresponding wavefront set, which will allow supervised learning.

Bibliography

[1] J. Adler, S. Lunz, O. Verdier, C.-B. Schönlieb, and O. Öktem. Task-adapted reconstruction for inverse problems. *arXiv preprint arXiv:1809.00948*, 2018.

[2] J. Adler and O. Öktem. Deep bayesian inversion. *ArXiv*, abs/1811.05910, 2018.

[3] J. Adler and O. Öktem. Learned primal-dual reconstruction. *IEEE Trans. Med. Img.*, 37(6):1348–1357, 2018.

[4] F. Andersson, M. V. de Hoop, and H. Wendt. Multiscale discrete approximation of fourier integral operators. *Multiscale model simulation*, 1:111–145, 2012.

[5] H. Andrade-Loarca and G. Kutyniok. tfshearlab: The tensorflow digital shearlet transform for deep learning. *arXiv preprint arXiv:2006.04591*, 2020.

[6] H. Andrade-Loarca, G. Kutyniok, and O. Öktem. Shearlets as feature extractor for semantic edge detection: The model-based and data-driven realm. *Proc. R. Soc. A.*, 2243(476), 2020.

[7] H. Andrade-Loarca, G. Kutyniok, O. Öktem, and P. Petersen. Extraction of digital wavefront sets using applied harmonic analysis and deep neural networks. *SIAM J. Imaging Sci.*, 12(4):1936–1966, 2019.

[8] H. Andrade-Loarca, G. Kutyniok, O. Öktem, and P. Petersen. Extraction of digital wavefront sets using applied harmonic analysis and deep neural networks. *SIAM J. Imaging Sci.*, 12(4):1936–1966, 2019.

[9] H. Andrade-Loarca, G. Kutyniok, O. Öktem, and P. Petersen. Deep microlocal reconstruction for limited-angle tomography. *arXiv preprint arXiv:2108.05732*, 2021.

[10] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(5):898–916, 2011.

[11] G. Bertasius, J. Shi, and L. Torresani. Deepedge: A multi-scale bifurcated deep network for top-down contour detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4380–4389, 2015.

[12] M. Bertero, H. Lantéri, and Z. L. Iterative image reconstruction: a point of view. In *Pubblicazioni del Centro De Giorgi Proceedings of the Interdisciplinary Workshop on Mathematical Methods in Biomedical Imaging and Intensity-Modulated Radiation Therapy (IMRT), Pisa, Italy, October 2007*, pages 37–63, 2008.

[13] M. Bertero, H. Lantéri, and L. Zanni. Iterative image reconstruction: a point of view. In *Proceedings of the interdisciplinary workshop on mathematical methdos in biomedical imgaing and intensity-modulated radiation*, 2008.

[14] T. O. Binford. Inferring surfaces from images. *Artif. Intell.*, 17(1):205–244, 1981.

[15] M. Brady. Computational approaches to image understanding. *ACM Comput. Surv.*, 14(1):3–71, 1982.

[16] C. Brouder, N. V. Dang, and F. Hélein. A smooth introduction to the wavefront set. *Journal of Physics A: Mathematical and Theoretical*, 47(44):443001, oct 2014.

[17] T. A. Bubba, G. Kutyniok, M. Lassas, M. März, W. Samek, S. Siltanen, and V. Srinivasan. Learning the invisible: A hybrid deep learning-shearlet framework for limited angle computed tomography. *Inverse Probl.*, 35(6), 2019.

[18] P. Butzer and R. Nessel. *Fourier Transform Methods and Second-Order Partial Differential Equations*. Fourier Analysis and Approximation. Birkhäuser Verlag, 1971.

[19] E. Candes, L. Demanet, D. L. Donoho, and L. Ying. Fast discrete curvelet transform. *SIAM Multiscale Model. Simul.*, 5:861–899, 2006.

[20] E. Candes, L. Demanet, and L. Ying. Fast computation of Fourier integral operators. *SIAM J. Sci. Comput.*, 29(6):2464–2493, 2007.

[21] E. Candès and D. Donoho. *Curvelets- a surprisingly effective nonadaptive representation for objects with edges*, volume 1 of *Curves and Surface Fitting*. Springer, 2000.

[22] E. J. Candès and D. L. Donoho. New tight frames of curvelets and optimal representation with c^2 singularities. *Comm. Pure. Appl. Math.*, 57:219–266, 2004.

[23] E. J. Candès and D. L. Donoho. Continuous curvelet transform: I. resolution of the wavefront set. *Appl. Comput. Harmon. Anal.*, 19(2):162–197, 2005.

[24] J. Canny. A computational approach to edge detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 8(6):679–698, 1986.

[25] A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40(1):120–145, 2010.

[26] H. Chen, Y. Zhang, M. K. Kalra, F. Lin, Y. Chen, P. Liao, J. Zhou, and G. Wang. Low-dose ct with residual encoder-decoder convolutional neural network (RED-CNN). *IEEE Transactions on Image Processing*, 2017.

[27] D. Comaniciu and P. Meer. Mean shift: a robust approach toward feature space analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2002.

[28] G. Cybenko. Approximation by superpositions of a sigmoid function. *Mathematics of Control, Signals, and Systems*, 2(2):251–257, 1989.

[29] M. Dashti and A. Stuart. *The Bayesian approach to inverse problems.* Handbook of Unceratinty Quantification. Springer-Verlag, 2016.

[30] I. Daubechies. Orthonormal bases of compactly supported wavelets. *Communications on Pure and Applied Mathematics*, 7(41):909–969, 1988.

[31] L. Demanet and L. Ying. Fast wave computation via fourier integral operators. *Mathematics of Computation*, 81(279):1455–1486, 2012.

[32] M. N. Do and M. Vetterli. The contourlet transform: An efficient directional multiresolution image representation. *IEEE Trans. Image Proc.*, 14:2091–2106, 2005.

[33] P. Dollár, Z. Tu, and S. Belongie. Supervised learning of edges and object boundaries. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2006.

[34] D. L. Donoho. Sparse components of images and optimal atomic decomposition. *Constr. Approx.*, 17:353–382, 2001.

[35] R. Duffin and A. Schaeffer. A class of nonharmonic fourier series. *Trans. Amer. Math. Soc*, 72:341–366, 1952.

[36] J. J. Duistermaat. Applications of fourier integral operators. *Séminaire Équations aux dérivées partielles (Polytechnique)*, pages 1–24, 1971-1972.

[37] A. Dutt and V. Rokhlin. Fast fourier transforms for nonequispaced data. *SIAM J. Sci. Comp.*, 14:1368–1393, 1993.

[38] M. E. Davison. The ill-conditioned nature of the limited angle tomography problem. *SIAM J. Appl. Math.*, 43, 04 1983.

[39] L. C. Evans. *Partial Differential Equations.* Providence: American Mathematical Society, 1998.

[40] S. Evans and P. B. Stark. Inverse problems as statistics. *Inverse Problems*, 18:R1–R55, 2002.

[41] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient graph-based image segmentation. *Int. J. Comput. Vis.*, 2004.

[42] D. J. Field, A. Hayes, and R. F. Hess. Contour integration by the human visual system: Evidence for a local ”association field”. *Vision Res.*, 33(2):173–193, 2011.

[43] G. B. Folland. *Introduction to partial differential equations.* Princeton University Press, Princeton, NJ, 1995.

[44] I. M. Gel'fand, M. I. Graev, and N. Y. Vilenkin. *Generalized functions. Volume 5, Integral geometry and representation theory.* Academic Press, New York, 1966.

[45] M. Genzel and G. Kutyniok. Aymptotic analysis of inpainting via universal shearlet systems. *SIAM J. Imag. Sc.*, 4(7):2301–2339, 2014.

[46] S. Ghosal and A. W. Van der Vaart. *Fundamental of nonparametric bayesian inference.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2017.

[47] K. Gröchenig. *Foundations of Time-Frequency Analysis.* Birkhäuser, 2001.

[48] P. Grohs. Continuous shearlet frames and resolution of the wavefront set. *Monatsh. Math.*, 164(4):393–426, 2011.

[49] P. Grohs and Z. Kereta. Continuous parabolic molecules. *Research Report*, 2015.

[50] K. Guo, G. Kutyniok, and D. Labate. Sparse multidimensional representations using anisotropic dilation and shear operators. In *Wavelets and Splines*, pages 189–201, Nashville, TN, 2005. Nashboro Press,.

[51] K. Guo and D. Labate. Sparse shearlet representation of fourier integral operators. *Electronic Research Announcements in Mathematical Science*, 14:7–19, 2007.

[52] K. Guo and D. Labate. Representation of fourier integral operators using shearlets. *Journal of Fourier Analysis and Applications*, 14:327–371, 2008.

[53] J. Hadamard. Sur les problèms aux dérivées partielles et leur signification physique. *Princeton University Bulletin*, pages 49–52, 1902.

[54] B. Hariharan, P. Arbelaez, L. Bourdev, S. Maji, and J. Malik. Semantic contours from inverse detectors. In *Proceedings of the 2011 International Conference on Computer Vision*, ICCV '11, pages 991–998, Washington, DC, USA, 2011. IEEE Computer Society.

[55] C. Harris and M. Stephens. A combined corner and edge detector. In *Fourth Alvey Vision Conference*, pages 147–151, 1988.

[56] K. He, X. Zhang, and J. Sun. Deep residual learning for image recognition. In *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[57] F. Heide, M. Steinberger, Y.-T. Tsai, M. Rouf, D. Pajak, D. Reddy, O. Gallo, J. Liu, W. Heidrich, K. Egiazarian, J. Kautz, and K. Pulli. Flexisp: A flexible camera image processing framework. *ACM Trans. Graph.*, 33(6):231:1–231:13, 2014.

[58] M. Heinrich, M. Stille, and T. Buzug. Residual u-net convolutional neural network architecture for low-dose ct denoising. *Current Directions in Biomedical Engineering*, 4:297–300, 09 2018.

[59] S. Helgason. *The Radon transform.* Birkhäuser, Boston, MA, 2nd edition, 1999.

[60] J. Hilbe and A. Robinson. *Methods of Statistical Model Estimation.* CRC Press, 2013.

[61] L. Hörmander. Fourier integral operators I. *Acta Math.*, 127:79–183, 1971.

[62] L. Hormander. *The Analysis of Linear Partial Differential Operators. I, Distribution Theory and Fourier Analysis.* Grundlehren Der Mathematischen Wissenschaften. Springer, 1990.

[63] L. Hörmander. *The analysis of linear partial differential operators I: Distribution theory and Fourier analysis.* Classics in Mathematics. Springer, Berlin, 2nd edition, 2003.

[64] X. Hu, Y. Liu, and B. Ren. Learning hybrid convolutional features for edge detection. *Nuerocomputing*, 2018.

[65] K. H. Jin, M. T. McCann, and M. Unser. Deep convolutional neural network for inverse problems in imaging. *IEEE Transactions on Image Processing*, 26(9):4509–4522, 2016.

[66] J. P. Kaipio and E. Somersalo. *Statistical and Computational Inverse Problems.* Applied Mathematical Sciences. Springer-Verlag, 2005.

[67] E. Kang, J. Min, and J. C. Ye. Wavenet: a deep convolutional neural network using directional wavelets for low-dose x-ray ct reconstruction. *Medical physics*, 44(10):e360–e375, 2017.

[68] A. Kaur, A. Raj, N. Jayanthi, and S. Indu. Inpainting of irregular holes in a manuscript using unet and partial convolution. In *2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA)*, pages 778–784, 2020.

[69] J. Y. Koh, W. Samek, K.-R. Müller, and A. Binder. Object boundary detection and classification with image-level labels. In *German Conference on Pattern Recognition*, pages 153–164. Springer, 2017.

[70] V. P. Krishnan and E. T. Quinto. *Microlocal analysis in tomography.* Handbook of Mathematical Methods in Imaging. Springer, 2015.

[71] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

[72] G. Kutyniok and D. Labate. Resolution of the wavefront set using continuous shearlets. *Trans. Amer. Math. Soc.*, 361(5):2719–2754, 2009.

[73] G. Kutyniok and D. Labate. *Shearlets. Multiscale analysis for multivariate data*, volume 1 of *Applied and Numerical Harmonic Analysis.* Birkhäuser, 2012.

[74] G. Kutyniok, W.-Q. Lim, and R. Reisenhofer. ShearLab 3D: Faithful digital shearlet transforms based on compactly supported shearlets. *ACM Trans. Math. Softw.*, 42(1):5:1–5:42, 2016.

[75] G. Kutyniok, W.-Q. Lim, and X. Zhuang. Digital shearlet transforms. In *Shearlets*, pages 239–282. Springer, 2012.

[76] S. L. Sur l'impossibilité de la multiplication des distributions. *C. R. Acad. Sci. Paris*, 238:847–848, 1954.

[77] S. Lang. *Real and functional analysis*, volume 142. Springer Science & Business Media, 2012.

[78] F. Liese and K.-J. Miescke. *Statistical Decision Theory: Estimation, Testing, and Selection.* Springer Series in Statistics. Springer-Verlag, 2008.

[79] Y. Liu, M.-M. Cheng, D.-P. Fan, L. Zhang, J.-W. Bian, and D. Tao. Semantic edge detection with diverse deep supervision. *IEEE transactions on pattern analysis and machine learning*, 2019.

[80] D. Ludwig. The Radon transform on Euclidean space. *Communications on Pure and Applied Mathematics*, 19:49–81, 1966.

[81] S. Mallat and W. Liang Hwang. Singularity detection and processing with wavelets. *IEEE Trans. on Information Theory*, 38(2):617–643, 1992.

[82] A. Markoe. *Analytic Tomography.* Encyclopedia of mathematics and its applications. Cambridge University Press, 2006.

[83] R. J. I. Marks. *Introduction to Shannon sampling and interpolation theory.* Springer Science & Business Media, 2012.

[84] D. Marr and E. Hildreth. Theory of edge detection. *Proc. R. Soc. Lond. B*, 207(1167):187–217, 1980.

[85] W. McCulloch and W. Pitts. A logical calculus of ideas immanent in nervous activity. *Bull. Math. Biophys.*, 5:115–133, 1942.

[86] P. Michor. *Manifolds of Differentiable Mappings.* Shiva Publishing Ltd., 01 1980.

[87] J. Mueller and S. Siltanen, editors. *Linear and Nonlinear Inverse Problems with Practical Applications*, volume 10 of *Computational Science and Engineering.* Society for Industrial and Applied Mathematics, United States, 2012.

[88] J. M. Prewitt. Object enhancement and extraction. *Pict. Process. Psychopictorics*, 10(1):15–19, 1970.

[89] E. T. Quinto. Singularities of the X-ray transform and limited data tomography in $\mathbb{R}^2$ and $\mathbb{R}^3$. *SIAM J. Math. Anal.*, 24(5):1215–1225, 1993.

[90] E. T. Quinto. An introduction to x-ray tomography and radon transforms. In *Proceedings of Symposia in Applied Mathematics*, volume 63, 2006.

[91] E. T. Quinto and O. Öktem. Local tomography in electron microscopy. *SIAM J. Appl. Math.*, 68(5):1282–1303, 2008.

[92] M. Reed and B. Simon. *Methods of mathematical physics: Fourier analysis, self-adjointness.* Academic Press, 1975.

[93] R. Reisenhofer, J. Kiefer, and E. J. King. Shearlet-based detection of flame fronts. *Exp. Fluids*, 57, 11 2015.

[94] R. Reisenhofer and E. J. King. Edge, ridge, and blob detection with symmetric molecules. *SIAM J. Imaging Sci.*, 12(4):1585–1626, 2019.

[95] L. G. Roberts. *Machine perception of three-dimensional solids.* PhD thesis, Massachusetts Institute of Technology, 1963.

[96] Y. Romano, M. Elad, and M. Peyman. The little engine that could: Regularization by denoising (red). *SIAM J. Imaging Sci.*, 10(4):1804–1844, 2017.

[97] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing.

[98] L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D.*, 1–4(60):259–268, 1992.

[99] W. Rudin. *Real and complex analysis.* McGraw-Hill, 1987.

[100] D. E. Rumelhart, G. E. Hinton, and R. Williams. *Learning internal representations by error propagation.* Parallel Distributed Processing: Explorations in the Microstructure of Cognition. MIT Press, 1986.

[101] L. Ruthotto and E. Haber. Deep neural networks motivated by partial differential equations. *J. Math Imaging Vis*, 2018.

[102] Y. Sasaki. The truth of the F-measure. *Teach Tutor mater*, 1(5):1–5, 2007.

[103] M. Sato. Regularity of hyperfunctions solutions of partial differential equations. In *Actes du Congrès international des mathématiciens*, volume 2, pages 785–794, Paris, 1971. Gauthier-Villars.

[104] T. Schuster. The method of approximate inverse: theory and applications. *Lecture Notes in Mathematics*, 1906, 2007.

[105] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Trans. on Pattern Anal. Mach. Intell.*, 1997.

[106] H. F. Smith. A parametrix construction for wave equations with $c^{1,1}$ coefficients. *Ann. Inst. Fourier*, 48:797–835, 1998.

[107] I. Sobel. An isotropic 3x3 image gradient operator. *Presentation at Stanford A.I. Project 1968*, 2014.

[108] E. M. Stein. *Harmonic Analysis: Real-variable Methods, Orthogonality, and Oscillatory integrals.* Princeton University Press, 1993.

[109] E. C. Stueckelberg and T. A. Green. Elimination des constantes arbitraires dans la theéorie des quanta. *Helv. Phys. Acta*, 24:153–174, 1951.

[110] M. Tan and Q. V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *ArXiv*, abs/1905.11946, 2019.

[111] M. E. Taylor. *Pseudodifferential operators.* Princeton University Press, Princeton, NJ, 1981.

[112] A. N. Tikhonov, A. Goncharsky, V. V. Stepanov, and A. Yagola. Numerical methods for the solution of ill-posed problems. In *Mathematics and its Applications*, 1995.

[113] V. Torre and T. A. Poggio. On edge detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 8(2):147–163, 1986.

[114] F. Trèves. *Introduction to pseudodifferential and Fourier integral operators*, volume 1 of *The University Series in Mathematics.* Plenum Pres, 1980.

[115] G. Uhlmann and A. Vasy. The inverse problem for the local geodesic X-ray transform. *Invent. Math.*, 205, 10 2012.

[116] M. V. de Hoop, H. Smith, G. Uhlmann, and R. D. V. der Hilst. Seismic imaging with the generalized radon transform: A curvlet transform perspective. *Inverse Problems*, 25(2):25005–25021, 2009.

[117] J. Velikina, S. Leng, and G.-H. Chen. Limited view angle tomographic image reconstruction via total variation minimization. *Proceedings of SPIE*, 6510, 2007.

[118] G. Wang. A perspective on deep imaging. *IEEE Access*, 7:8914–8924, 2016.

[119] Q. Xu, H. Yu, X. Mou, L. Zhang, J. Hsieh, and G. Wange. Low-dose x-ray ct reconstruction via dictionary learning. *IEEE Transaction on medical imaging*, 31(9):1682–1697, 2012.

[120] S. Yi, D. Labate, G. R. Easley, and H. Krim. A shearlet approach to edge analysis and detection. *IEEE Trans. Image Process.*, 18(5):929–941, 2009.

[121] Z. Yu, C. Feng, M.-Y. Liu, and S. Ramalingam. Casenet: Deep category-aware semantic edge detection. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA*, pages 21–26, 2017.

[122] Z. Yu, W. Liu, Y. Zou, C. Feng, S. Ramalingam, B. V. Kumar, and J. Kautz. Simultaneous edge alignment and learning. *arXiv preprint arXiv:1808.01992*, 2018.