ABSTRACT

Title of dissertation: Harmonic Analysis and Machine Learning

Michael Pekala
Doctor of Philosophy, 2018

Dissertation directed by: Professor Wojciech Czaja and
Professor Doron Levy
Department of Mathematics

This dissertation considers data representations that lie at the interesection of harmonic analysis and neural networks. The unifying theme of this work is the goal for robust and reliable machine learning. Our specific contributions include a new variant of scattering transforms based on a Haar-type directional wavelet, a new study of deep neural network instability in the context of remote sensing problems, and new empirical studies of biomedical applications of neural networks.

Harmonic Analysis and Machine Learning

by

Michael Pekala

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2018

Advisory Committee:
Professor Wojciech Czaja, Chair/Advisor
Professor Doron Levy, Co-Chair/Co-Advisor
Professor Radu Balan
Professor Maria Cameron
Professor Daniel Butts

Acknowledgments

So many people have helped me over the years that to attempt to name them all would inevitably lead to omission. Thanks so much to all my past co-workers and collaborators! I'd like to thank all my former colleagues at APL, especially Phillipe Burlina, Dennis Lucarelli, Scott Peacock, and I-Jeng Wang. I can't imagine better people to work with and learn from!

I would also like to thank my committee: Dr. Balan whose AMSC project sequence was one of my best experiences at UMD, Dr. Cameron for her wonderful RIT which rekindled my enthusiasm for graduate study, and Dr. Butts for thoughtful suggestions on the dissertation. Most especially I thank my advisors, Dr. Czaja and Dr. Levy, for their encouragement, guidance, and generosity over these past years. I can't thank them enough for all they have done for me.

Finally, I would like to thank my family: my parents and Dan for always encouraging me, and Laurie and Josh for their patience, love, and support. Words are inadequate; I can't wait to play trucks with you guys again!

Table of Contents

List of Tables

List of Figures

List of Abbreviations and Notation

$\mathbb{C}$	The set of complex numbers
$\mathbb{R}$	The set of real numbers
$\mathbb{Z}$	The set of integers
$\mathcal{L}$	A lattice
$L^2(\mathbb{R}^n)$	The space of square integrable functions from $\mathbb{R}^n \to \mathbb{C}$
$L^1(\mathbb{R}^n)$	The space of (Lebesgue) integrable functions from $\mathbb{R}^n \to \mathbb{C}$
$GL_n(\mathbb{R})$	The general linear group of invertible $n \times n$ matrices
$SL_n(\mathbb{R})$	The special linear group of invertible $n \times n$ matrices with $\lvert \det \cdot \rvert = 1$.
χ_E	The indicator function on the measurable set E

AE	Adversarial Examples
AMD	Age-related Macular Degeneration
APL	Applied Physics Laboratory
CDW	Composite Dilation Wavelet
CHCDW	Crystallographic Haar-type Composite Dilation Wavelet
CNN	Convolutional Neural Network
CPU	Central Processing Unit
CWT	Continuous Wavelet Transform
DME	Diabetic Macular Edema
DNN	Deep Neural Network
DWT	Discrete Wavelet Transform
DCWT	Discrete Composite Wavelet Transform
GP	Gaussian Process
GPU	Graphical Processing Unit
KS	Kolmogorov-Smirnov
ML	Machine Learning
MNIST	Modified National Institute of Standards and Technology database
MRA	Multiresolution Analysis
MSE	Mean Square Error
MSPE	Mean Square Prediction Error
OCT	Optical Coherence Tomography
ONB	Orthonormal Basis
ReLU	Rectified Linear Unit
SAR	Synthetic Aperture Radar
UMD	University of Maryland

Chapter 1: Preliminaries: Network-based Feature Representations

This dissertation considers data representations that lie at the intersection of harmonic analysis and neural networks. The unifying theme of this work is the desire for robust and reliable machine learning. Our specific contributions include a new variant of scattering transforms based on a Haar-type directional wavelet, a new study of deep neural network instability in the context of remote sensing problems, and new empirical studies of biomedical applications of neural networks.

Deep neural networks (DNNs), which generate data representations by learning from data, have recently commanded a prominent role in the machine learning (ML) community. Despite the their unquestionably good performance and the vast amount of research attention, DNNs are still not completely understood from a mathematical perspective. Even relatively simple feedforward convolutional neural networks lack a complete mathematical characterization. At the other end of the design spectrum, there is a long and rich history of manually designed filters whose mathematical properties are comparatively well understood. One path toward bringing added mathematical rigor to network-based representations lies in the scattering transform framework devised by Stéphane Mallat. This framework brings together mathematical properties of designed filters within a network-based archi-

tecture reminiscent of feed forward neural networks. The underlying motivation is to explicitly systemize desirable properties of feature representations, a task which DNNs seem to be able to do implicitly but in a manner that is not well understood nor guaranteed.

Part of our motivation for studying these representations is a desire for robust and reliable ML. Questions about generalization performance and robustness are not new in the ML community; however the prominence of deep learning together with recent observations that these networks may lack robustness to certain designed perturbation in the signal input space have elevated their attention of late. So-called *adversarial examples* have raised a number of questions about what is and is not being captured by neural networks. While the landscape of neurally-inspired architectures will undoubtedly continue to evolve (e.g. the capsule networks recently advocated by Hinton [131]) there are currently open and exciting research opportunities in understanding the behavior of these deep neural networks.

This thesis makes a contribution to this broader topic by the introduction of two new algorithms and associated empirical studies. Our first algorithmic exploration embeds a composite Haar-type directional wavelet within a scattering transform framework. We find that a scattering transform when equipped with this Haar-type wavelet demonstrates good performance on benchmark image classification problems involving signals with fine, edge-like structure. However, they underperform scatterings based on wavelets traditionally employed for this kind of signal (e.g. Morlet wavelets). Follow-on experiments involving different signals (not dominated by fine edge structure) closes this performance gap somewhat. Due to

their geometric underpinnings, the Haar-type wavelets are also capable of generating images with certain aesthetic properties. This work is the topic of Chapter 2.

Our second study proposes an algorithm to construct adversarial examples (AE) which capture physical considerations that arise in remote sensing settings. This venue for AE has not yet been well explored, and ours is one of the first ever studies to consider this problem explicitly. We identify key issues and explore how the support of designed perturbations is related to the attack success rate. Our study lays a foundation for future work which incorporates practical domain constraints into the AE optimization problem. Our final application involves biomedical imaging; in particular, retinal image analysis. These applications are the topics of Chapter 3 and Chapter 4.

Chapter 2: Modern Applications of Crystallographic Wavelets

This chapter explores modern applications of a particular class of directional Haar-type wavelets whose mathematical properties are well understood but whose possible practical applications have not been widely studied. At the same time we also observe that there exist in the literature a number of *scattering transform* frameworks which are designed to provide a possible bridge between wavelet frames and modern deep neural nets. Therefore, our primary goals in this chapter are to (a) obtain insight into the performance characteristics of these Haar-type wavelets relative to more popular directional techniques (such as shearlets) and (b) to explore possible roles for these Haar-type wavelets within scattering transform frameworks, with an eye towards future possible synergies with neural network techniques.

Before proceeding, we pause to mention that the field of wavelets is quite broad and encompasses a great deal of work that has been developed since its origins in the 1980s. A comprehensive summary of this field is simultaneously beyond the scope of this document and the credentials of the author; however, the interested reader is encouraged to consult [83] for an extensive collection of papers covering the early development of wavelets assembled by leading experts in the field. There are similarly many too many fine books on wavelets to list them all; we mention

[14, 15, 84, 113, 138].

In this chapter we will be especially interested in *Composite Dilation Wavelets* (CDW) [77, 79, 80] which construct representations for $L^2(\mathbb{R}^n)$ by means of a pair of dilation operators coupled with a suitable notion of translation (either discrete or continuous). The use of a pair of dilation operators distinguishes CDW from more traditional wavelets which relied upon a single dilation operator. Generalization to multiple dilation operators provides greater flexibility to capture anisotropic characteristics of signals. The CDW framework includes as special cases a number of popular wavelet systems, including ridgelets [32], curvelets [33], and shearlets [46, 47, 78, 93].

Many of these CDW systems confer desirable mathematical properties. For example, Candés and Donoho proved that, for a "cartoon-like" function f (i.e. piecewise continuous functions with discontinuities along C^2 curves), the best $M-$term curvelet approximation f_M of f converges quadratically [34]

$$||f - f_M||^2 \leq CM^{-2}(\log_2 M)^3.$$

This is in contrast to traditional (tensor product) wavelet expansions which only provide an $\mathcal{O}(M^{-1})$ approximation error rate [113]. Furthermore, it has been shown that this quadratic convergence rate is essentially optimal, assuming one is using a non-adaptive wavelet system. Analogous results have been developed for other CDW systems, such as shearlets [76].

However, images dominated by edge-like structures do not necessarily satisfy

the "cartoon-like" property for which the aforementioned CDW systems yield optimal results. Furthermore, many of these systems (e.g. shearlets and curvelets) are compactly supported in the frequency domain. For images dominated by discontinuities in the time domain it is natural to ask whether wavelets explicitly designed to localize features in the time domain may provide some representational benefit.

This desire to capture oriented structure in signals motivates our study of a somewhat less well-known variant of CDW, the *Crystallographic Haar-type Composite Dilation Wavelets* (CHCDW), originally developed by [96] and subsequently extended by various authors, including [19, 115]. CDW are deemed to be "Haar-type" when they satisfy the following two properties:

1. There exists an associated multi-resolution analysis (MRA);

2. The scaling function (and therefore the wavelets) associated with the MRA are constructed by linear combination of characteristic functions.

The first property is fairly standard and stems from a goal to analyze signals at multiple resolutions; a concise definition is provided in Definition 2.1.2 and the theory of MRA is by now well-established (many details can be found in references such as [83, 84, 113]). However, the second requirement, that the scaling function φ be a linear combination of characteristic functions, is a less conventional property. Intuitively this constraint on the support of the scaling function suggests that Haar-type systems may be especially well-suited for localizing discrete or piecewise-constant features in the time domain. Of course, the celebrated Heisenberg uncertainty principle [16] suggests that this comes at the cost of good frequency localization and

therefore Haar-type systems should be less effective at representing signals with strong oscillatory components.

However, detailed empirical studies into the practical performance of these systems is generally lacking. Noteworthy exceptions include: [98] where the authors compare Haar-type CDW to curvelets and contourlets in the context of image denoising applications, [97] where the authors compare Haar-type CDW to classical Haar wavelets for image denoising, and [50] where the authors demonstrate the advantage of Haar-type CDW over shearlets for inpainting binary images when used within a total variation minimization framework.

In this chapter we describe in detail the CHCDW framework, with a focus on the discrete setting and associated computational considerations. This begins with a detailed description of an algorithm developed by Dr. Benjamin Manning to implement the CHCDW. We then describe a new implementation of this algorithm that provides a substantial improvement in runtime performance, the benefits of which are best appreciated when the underlying signal is non-trivial in size. Using this algorithm, we conduct a study of directional sensitivity by comparing CHCDW to other wavelets systems in a number of signal processing applications where edge-like structure dominates. These experiments are new but their motivation is inspired, in part, by the novel analysis of directional structures for shearlets that was originally presented in [145]. We then present numerical experiments related to classification problems. In addition to exploring practical applications of this wavelet, our end objective is to place this wavelet in the context of a larger discussion regarding stable and reliable feature representations for signal processing problems. To this end we

will also explore integration with scattering transform frameworks and ultimately towards synergies with neural network techniques.

2.1 Composite Dilation Wavelets: Overview

Before defining composite dilation wavelets and their variants, we must introduce some basic building blocks. Note that no new results are presented in this section and much of the introductory material is adapted from [19].

Since our interest is ultimately in processing discrete signals, the underlying structure of a lattice plays a key role. Given m linearly independent vectors $v_1, \ldots, v_m \in \mathbb{R}^n$ the associated *lattice* $\mathcal{L}$ is the collection of vectors obtainable by means of the linear combinations

$$\mathcal{L}(v_1, \ldots, v_m) = \{x_1 v_1 + x_2 v_2 + \ldots + x_m v_m \,|\, x_1, x_2, \ldots, x_m \in \mathbb{Z}\}\,. \tag{2.1}$$

Equivalently, if V is the $n \times m$ matrix whose columns are $\{v_i\}_{i=1}^m$, the lattice can be expressed in matrix form

$$\mathcal{L}(V) = \{Vx : x \in \mathbb{Z}^m\}.$$

The matrix V is said to be the *generating matrix* of $\mathcal{L}$ and the lattice $\mathcal{L}$ is said to be *full rank* if $m = n$. Note that lattices are not associated with a unique generating matrix. We also recall a useful result regarding equivalences between lattices. A matrix U is called *unimodular* if it is a square integer matrix with determinant equal to 1 or -1.

Theorem 2.1.1 ([57] Theorem 3.2)**.** *Matrices A and B generate the same lattice $\mathcal{L}$, that is $\mathcal{L}(A) = \mathcal{L}(B)$, if and only if $A = BU$ where U is a unimodular matrix.*

We can now define the two fundamental operations necessary to construct wavelets: translation and dilation. Let $f \in L^2(\mathbb{R}^n)$ and define *translation* $T_k : L^2(\mathbb{R}^n) \to L^2(\mathbb{R}^n)$ of f by $k \in \mathcal{L}$ as $T_k f(x) = f(x - k)$. Note that, by defining translation as a fully discrete operator, we are distinguishing this framework from the semi-discrete wavelet setting[1] which is also quite common in the literature. Let $f \in L^2(\mathbb{R}^n)$ and c be an invertible matrix; then *dilation* $D_c : L^2(\mathbb{R}^n) \to L^2(\mathbb{R}^n)$ of f by c is defined to be the operation $D_c f(x) = |\det c\,|^{-1/2} f(c^{-1}x)$.

With these basic definitions, we can now define our wavelet system of interest. Let a be an *expanding matrix* (i.e. a matrix whose eigenvalues all have modulus greater than 1) and let $A = \{a^j : j \in \mathbb{Z}\}$ denote an associated collection of expanding matrices. Let B be a subgroup of $GL_n(\mathbb{R})$, $\mathcal{L}$ be a full rank lattice, and $\Psi = \{\psi_1, \ldots, \psi_L\} \subset L^2(\mathbb{R}^n)$ a collection of functions. A composite dilation wavelet (CDW) is defined as follows:

Definition 2.1.1 ([19] Definition 1)**.** *The collection of functions $\Psi = (\psi_1, \psi_2, \ldots, \psi_L) \subset L^2(\mathbb{R}^n)$ is a* composite dilation (multi-)wavelet *if there exists a collection of expanding matrices A, a group of invertible matrices B, and a full rank lattice $\mathcal{L}$ such that the collection of functions*

$$A_{a,B,\mathcal{L}}(\Psi) = \{D_a D_b T_k \psi_i : a \in A, b \in B, k \in \mathcal{L}, i = 1, \ldots, L\}, \tag{2.2}$$

[1] By which we mean the setting where dilation is discrete and translation is continuous; sometimes called "semi-continuous" or just "continuous" if the discrete nature of dilation is irrelevant.

forms a an orthonormal basis of $L^2(\mathbb{R}^n)$.

A few comments about Definition 2.1.1 are in order. While ideal for our purposes, this definition is neither the first proposed in the literature nor the most general definition possible. CDW were originally introduced by Guido Weiss and his collaborators [77, 79, 80] and generalizations of Definition 2.1.1 typically involve extensions to Parseval frames (e.g. [18]). It is also worth mentioning that the CDW framework includes a number of familiar wavelet settings as special cases. For example, a standard (one-dimensional) wavelet is obtained by setting $n = 1, a = 2$ and $\mathcal{L} = \mathbb{Z}$. Alternatively, in higher dimensions if B is the one-element group consisting of the identity matrix, then one recovers the standard multiwavelet definition. As mentioned previously, shearlets are a special case of CDW; one example uses the (infinite cardinality) group of integer shear matrices for the dilation group B:

$$B = \left\{ \begin{pmatrix} 1 & j \\ 0 & 1 \end{pmatrix} : j \in \mathbb{Z} \right\}.$$

An important concept in wavelet theory is the analysis of signals at multiple scales by means of a multiresolution analysis (MRA). This notion of an MRA applies naturally to CDW as well [80]:

Definition 2.1.2. *The collection* $\{V_j\}_{j\in\mathbb{Z}}$ *of closed subspaces of* $L^2(\mathbb{R}^n)$ *is an* $(a, B, \mathcal{L})$-multiresolution analysis *if all of the following conditions are satisfied:*

M.1 $V_j \subset V_{j+1}$, *where* $D_a V_{j+1} = V_j$;

M.2 *Closure*$(\bigcup_{j\in\mathbb{Z}} V_j) = L^2(\mathbb{R}^n)$;

M.3 $\bigcap_{j\in\mathbb{Z}} V_j = \{0\}$;

M.4 *There exists a function* $\varphi \in V_0$ *such that* $\{D_b T_k \varphi : b \in B, k \in \mathcal{L}\}$ *is an orthonormal basis for* V_0.

The function φ from Item **M.4** is called the *(composite) scaling function* of the given MRA and plays a key role in constructing the corresponding wavelets. For classical wavelets this might be accomplished via the Smith-Barnwell equation (see e.g. [83] for details); however, we will defer discussing explicit constructions until after we further specialize CDW in Section 2.1.1.

In addition to MRA, there are other properties of wavelet systems that are of interest, such as compact support, regularity (i.e. smoothness), and accuracy. In this context, accuracy of a function f is the highest degree p such that all multivariate polynomials with degree less than p are exactly reproducible from translates of f on some lattice [31]. When the wavelet scaling function has accuracy, this provides additional information on classes of functions that can be represented exactly. We mention accuracy and regularity for completeness, but will not focus on them further as they are outside the scope of this thesis. However, we do mention that, for crystallographic wavelets (which will be introduced in Section 2.1.1), the work of [115] provides the theoretical machinery necessary to develop crystallographic wavelets with given accuracy and also presents numerical results for one dimensional signals. We also mention a result by P. Houska, who proved that, if one seeks compactly supported MRA with regularity, then B must be a finite group [87].

Compactly supported wavelets are desirable from a computational standpoint, especially when the support is simple (e.g. as opposed to the fractal "twin-dragon"

structure described in [115] which arises from wavelets employing a single quincunx dilation). One such class of wavelets are the minimally supported frequency (MSF) composite dilation wavelet which are constructed from characteristic functions of compact sets in the frequency plane [77, 80]. These wavelets provide excellent localization in the frequency domain at the cost of performance in the time domain. Analogously, a wavelet is termed *Haar-type* if it is constructed from characteristic functions in the time domain. A further specialization developed by [96] is the following:

Definition 2.1.3. *A wavelet system is called a* Haar-type composite dilation wavelet *if its multiwavelets ψ_i are constructed from linear combinations of characteristic functions in the time domain and there is an associated MRA.*

In the next section we will explore a specific Haar-type CDW whose properties are further modified by a lattice assumption.

2.1.1 Crystallographic Haar-type CDW

Even after specializing to a Haar-type CDW, there still remain a number of degrees of design freedom available in (2.2), including the specific structure of the group B, the choice of dilation matrix a, and the specific composition of the scaling function used to realize the MRA. Motivated by the discrete lattice-based setting upon which our digital signals are supported, an additional property of substantial interest is the *crystallographic condition*

Definition 2.1.4 ([19] Definition 3)**.** *A group of invertible matrices G and a full*

rank lattice $\mathcal{L}$ *satisfy the* crystallographic condition *if* $\mathcal{L}$ *is invariant under the action of* G*, i.e.* $G(\mathcal{L}) = \mathcal{L}$.

Intuitively, it makes sense that lattice preservation would enter into the picture; otherwise, one may have to resort to interpolation or other approximations in the course of computing the wavelet transform (beyond the approximations involved in initially digitizing the signal). For example, the lattice-preserving properties of shearing operators have been cited as an advantage of shearlets over curvelets; the latter rely upon rotations to capture directional information, and these rotations do not preserve the integer lattice [101].

Crystallographic Haar-like wavelets arise from the dual desire for Haar-type CDW together with a group B that is lattice-preserving. In particular, the desire for an orthogonal system together with B and $\mathcal{L}$ that satisfy the crystallographic condition enforces a special structure upon these groups which is specified by Theorem 2.1.2.

Theorem 2.1.2 ([19] Theorem 1)**.** *Suppose* $\{D_b T_k \phi(x) : b \in B, k \in \mathcal{L}\}$ *is an orthogonal system. If* $B(\mathcal{L}) = \mathcal{L}$*, then* B *is a finite subgroup of* $SL_n(\mathbb{R})$ *and* $\Gamma := B \ltimes \mathcal{L}$ *is a crystallographic group.*

The group operation associated with $B \ltimes \mathcal{L}$ in Theorem 2.1.2 is that of a standard semi-direct product

$$(b_1, k_1) \cdot (b_2, k_2) = (b_1 b_2, b_2^{-1} k_1 + k_2). \tag{2.3}$$

That this outer product is a crystallographic group implies that $|B|$ is finite since

Definition 2.1.5 ([19] Definition 4)**.** *Let G be a group of orthogonal transformations and $\mathcal{L}$ be a full rank lattice for $\mathbb{R}^n$. Then G is a* crystallographic group *if $\mathcal{L}$ is a subgroup of G and the quotient group (also called the point group) $G/\mathcal{L}$ is finite.*

With these additional restrictions on $B, \mathcal{L}$ one can take a slightly different perspective on the elements of (2.2). Since the elements of B are unitary, the expanding matrix a will be solely responsible for providing control over scaling. Conceptually, the dilation group B can now be thought of as a way to enrich the notion of translation by means of its action together with the underlying lattice $\mathcal{L}$.

In particular, the lattice $\mathcal{L}$ can be thought of in a group-theoretic sense as isomorphic to the abelian group of integers $\mathbb{Z}^n$ whose elements correspond to all possible ways of "shifting" the wavelets ψ_i. In the crystallographic CDW setting, the group $\Gamma = B \ltimes \mathcal{L}$ gives rise to a new, more general way to "shift" wavelet supports. Each $\gamma = D_b T_k \in \Gamma$ is defined pointwise on $\mathbb{R}^n$ as

$$\gamma(x) = b(x - k), \tag{2.4}$$

whose inverse

$$\gamma^{-1} = D_{b^{-1}} T_{-bk}, \tag{2.5}$$

follows from (2.3)

$$(b, k) \cdot (b^{-1}, -bk) = (bb^{-1}, bk - bk) = (I, 0).$$

One then defines the generalized shift operator $L_\gamma : L^2(\mathbb{R}^n) \to L^2(\mathbb{R}^n)$ as $L_\gamma f(x) = f(\gamma^{-1}(x))$. Then the CDW definition (2.2) can be written in an equivalent form that highlights this notion of a generalized translation

$$A_{a,\Gamma}(\Psi) = \{D_a L_\gamma \psi_i : a \in A, \gamma \in \Gamma, i = 1, \ldots, L\}. \tag{2.6}$$

We now have a specialization of CDW where the commutative group of translations $\mathbb{Z}^n$ has been replaced with a new (non-commutative) group Γ whose elements γ are the wavelet shifting parameters.

We also introduce a number of assumptions on the scaling matrix a which, while not strictly necessary to guarantee the existence of a wavelet system, admit some simpler conditions for proving existence of an MRA. In a classical (non-composite) wavelet system, it is common to make the assumption that $a(\mathcal{L}) \subset \mathcal{L}$, a kind of "closure" property with respect to the lattice. After going through all the trouble to ensure B is lattice-preserving, it would be a pity to fall off the lattice in the process of applying the scaling operator. Therefore, in our CHCDW setting, we also assume $a(\mathcal{L}) \subset \mathcal{L}$. One also makes the additional "normalizing" assumption that $aBa^{-1} = B$. A simple consequence of these two assumptions is that $a\Gamma a^{-1}$ is then a subgroup of Γ.

Lemma 2.1.1. *If a scaling matrix satisfies $a(\mathcal{L}) \subset \mathcal{L}$ and $aBa^{-1} \subset B$, then $a\Gamma a^{-1}$ is a subgroup of Γ.*

Proof. We demonstrate this using the one-step subgroup test. First we observe that

$a\Gamma a^{-1}$ is a non-empty subset of Γ since for any $\gamma \in \Gamma$

$$\begin{aligned} a\gamma a^{-1}(x) &= ab(a^{-1}x - k), \\ &= aba^{-1}x - aba^{-1}ak, \\ &= aba^{-1}(x - ak), \\ &= \tilde{b}t_{ak}(x), \end{aligned}$$

where $\tilde{b} \in B$. Now it remains to show that for all $g_1, g_2 \in a\Gamma a^{-1}$ we also have $g_1 g_2^{-1} \in a\Gamma a^{-1}$. From a direct calculation we observe

$$\begin{aligned} (aba^{-1}, ak) \cdot (ac^{-1}a^{-1}, a\ell) &= (aba^{-1}ac^{-1}a^{-1}, (ac^{-1}a^{-1})^{-1}ak + a\ell) \\ &= (abc^{-1}a^{-1}, ack + a\ell) \\ &= (a\tilde{b}a^{-1}, a\tilde{\ell}), \end{aligned}$$

where $\tilde{b} \in B$ and $\tilde{\ell} \in \mathcal{L}$. □

More substantial results (and ones which directly motivates the above assumptions on the scaling matrix a) are given by Lemma 2.1.2, which provides conditions for an orthonormal system, and Theorem 2.1.3 which provides a sufficient condition for constructing a crystallographic wavelet with an associated MRA. Before stating these theorems, it is first necessary to introduce the notion of a tiling set.

Definition 2.1.6 ([19] Definition 5)**.** *Let G be a group of invertible matrices and let R and W be measurable sets in $\mathbb{R}^n$. Then R is called a* G-tiling set of W *if W*

is a disjoint union of the images of R under the action of G, i.e.

$$W = \bigcup_{g \in G} gR \quad \textit{with} \quad \mu(g_1 R \cap g_2 R) = 0 \ \textit{ for all } \ g_1 \neq g_2 \in G.$$

We designate a special tiling set, $R \subset \mathbb{R}^n$, called a *fundamental region* of the crystallographic group $B \ltimes \mathcal{L}$ if R is a $(B \ltimes \mathcal{L})$-tiling set of $\mathbb{R}^n$. We can now state a useful sufficient condition for showing that the action of generalized translation produces an orthonormal system

Lemma 2.1.2 ([19] Lemma 4)**.** *Suppose $B(\mathcal{L}) = \mathcal{L}$, R is a B-tiling set of S, and $\varphi = |\mu(R)|^{-1/2}\chi_R$. If $S = \cup_{b \in B} bR$ is a $\mathcal{L}$-tiling set of $\mathbb{R}^n$ then $\{D_b T_k \varphi : b \in B, k \in \mathcal{L}\}$ is an orthonormal system.*

This is intimately related to MRA condition Item **M.4** and, in addition to being interesting in its own right, is also a useful piece used to prove a useful sufficient condition for producing a crystallographic MRA

Theorem 2.1.3 ([19] Theorem 2)**.** *Suppose $B(\mathcal{L}) = \mathcal{L}$. Let $\varphi = |\mu(R)|^{-1/2}\chi_R$ and $V_0 = \overline{span}\{D_b T_k \varphi : b \in B, k \in \mathcal{L}\}$. Let $a \in GL_n(\mathbb{R})$ be an expanding matrix with $|\det(a)| = L + 1$. Furthermore, suppose the following three conditions also hold:*

(i) *$S = \bigcup_{b \in B} bR$ is a $\mathcal{L}$-tiling set of $\mathbb{R}^n$,*

(ii) *$a(\mathcal{L}) \subset \mathcal{L}$ and normalizes B, i.e. $aBa^{-1} = B$,*

(iii) *There exist $b_0, \ldots, b_L \in B$ and $k_0, \ldots, k_L \in \mathcal{L}$ such that*

$$aR = \bigcup_{i=0}^{L} (b_i R + b_i k_i). \tag{2.7}$$

Then the sequence of subspaces $\{V_j := D_a^{-j} V_0\}_{j \in Z}$ *is an MRA for* $L^2(\mathbb{R}^n)$ *and* φ *is a composite scaling function for this MRA.*

Thus, Theorem 2.1.3 tells us that to build a Haar-type scaling function for CHCDW it is sufficient to find a fundamental region R of $B \ltimes \mathcal{L}$ and a suitably well-behaved expanding matrix a that jointly satisfy Item **(iii)**. A concrete example which leverages this theorem will be provided in Section 2.2.1.

2.2 A Discrete Wavelet Transform for Crystallographic CDW

In addition to a principled way to represent multiscale signals, an advantage of wavelets with an associated MRA is that they admit efficient implementations. In this section we discuss a "cascade algorithm" proposed by [50] to implement a discrete wavelet transform for CHCDW. Ultimately, our goal is to analyze a function f by computing inner products of f with the collection of functions $A_{a,B,\mathcal{L}}(\Psi)$ from Definition 2.1.1. To implement this calculation efficiently, we will exploit properties of the MRA associated with the CHCDW. As part of our exposition we provide some additional (but straightforward) details that were not included in the original reference. We also describe briefly a new implementation of the transform developed by this author which demonstrates the practical computational benefits of the original algorithm.

A consequence of the MRA is that the scaling function φ from Item **M.4** is

refinable, which means there exists $\{c_\gamma\}_{\gamma\in\Gamma} \subset \mathbb{C}$ so that

$$\varphi(x) = \sum_{\gamma\in\Gamma} c_\gamma \varphi(\gamma a x), \tag{2.8}$$

where the series converges in L^2 norm. The coefficients $\{c_\gamma\}_{\gamma\in\Gamma}$ are typically referred to as the *low pass filter* or the *refinement mask*. Since the wavelets $\{\psi_i\}_{i=1}^L$ reside in the MRA function space V_1, there also exist *high pass filters* $\{d_\gamma^i\}_{\gamma\in\Gamma}$ such that

$$\forall i = 1, \ldots, L \quad \psi^i(x) = \sum_{\gamma\in\Gamma} d_\gamma^i \varphi(\gamma a x). \tag{2.9}$$

These filters will allow us to implement wavelet analysis and synthesis. Of primary interest will be the case where φ is compactly supported and all but finitely many of the $\{c_\gamma\}_{\gamma\in\Gamma}$ and $\{d_\gamma^i\}_{\gamma\in\Gamma}$ are zero. This implies that each ψ^i is also compactly supported and one can create a discrete composite wavelet transform (DCWT) that allows for fast multiscale signal decomposition.

We pause to present a few useful relations. For refinable functions, dilating

both sides of (2.8) by a gives the relation

$$\begin{aligned} D_a\varphi(x) &= D_a \sum_{\gamma\in\Gamma} c_\gamma \varphi(\gamma a x) \\ &= \sum_{\gamma\in\Gamma} c_\gamma D_a L_{\gamma^{-1}}\varphi(ax) \\ &= \sum_{\gamma\in\Gamma} |\det a\,|^{-1/2} c_\gamma L_{\gamma^{-1}}\varphi(a^{-1}ax) \\ &= |\det a\,|^{-1/2} \sum_{\gamma\in\Gamma} c_\gamma L_{\gamma^{-1}}\varphi(x) \\ &= |\det a\,|^{-1/2} \sum_{\gamma\in\Gamma} c_\gamma \varphi(\gamma x). \end{aligned} \tag{2.10}$$

It is also helpful to observe how the dilation and generalized translation operators "commute":

$$\begin{aligned} D_a L_{\eta^{-1}} f(x) &= |\det a|^{-1/2} f(\eta(a^{-1}x)) \\ &= |\det a|^{-1/2} f(ba^{-1}x - bk) \\ &= |\det a|^{-1/2} f(a^{-1}aba^{-1}x - a^{-1}abk) \\ &= D_a f(a(ba^{-1}x - bk)) \\ &= D_a f(a\eta a^{-1}(x)) \\ &= L_{a\eta^{-1}a^{-1}} D_a f(x). \end{aligned} \tag{2.11}$$

Let $f \in L^2(\mathbb{R}^n)$ be a signal we wish to process and let $\varphi \in L^2(\mathbb{R}^n)$ be the scaling function associated with a CHCDW. For $j \geq 0$ and $\eta \in \Gamma$ the we define the

scaling coefficients at scale j

$$s_j(\eta) = \langle f, D_{a^j} L_{\eta^{-1}} \varphi \rangle. \tag{2.12}$$

Similarly, the *wavelet coefficients* at scale j are defined to be

$$t_j^i(\eta) = \langle f, D_{a^j} L_{\eta^{-1}} \psi^i \rangle, \qquad i = 1, \ldots, L. \tag{2.13}$$

The CHCDW representation for f will then consist of the collection of sequences $\{t_1^i, \ldots, t_J^i, s_J\}_i$, that is the union of the wavelet coefficients at scales $1, \ldots, J$ together with the scaling coefficients at scale J.

The inner products in (2.12),(2.13) will be implemented iteratively by enlisting the aid of (2.8),(2.9). In particular, the idea is to construct an operator that produces the sequences s_{j+1} and t_{j+1} directly from the sequence s_j. A direct calculation,

which utilizes the linearity of dilation and (2.10),(2.11), gives that

$$\begin{aligned} s_{j+1}(\eta) &= \langle f, D_{a^{j+1}} L_{\eta^{-1}} \varphi \rangle \\ &= \langle f, D_{a^j} D_a L_{\eta^{-1}} \varphi \rangle \\ &= \langle f, D_{a^j} L_{a\eta^{-1}a^{-1}} D_a \varphi \rangle \\ &= \langle f, D_{a^j} L_{a\eta^{-1}a^{-1}} (|\det a|^{-1/2} \sum_{\gamma\in\Gamma} c_\gamma L_{\gamma^{-1}} \varphi) \rangle \\ &= \langle f, |\det a|^{-1/2} \sum_{\gamma\in\Gamma} c_\gamma D_{a^j} L_{a\eta^{-1}a^{-1}\gamma^{-1}} \varphi \rangle \\ &= |\det a|^{-1/2} \sum_{\gamma\in\Gamma} \bar{c}_\gamma \langle f, D_{a^j} L_{a\eta^{-1}a^{-1}\gamma^{-1}} \varphi \rangle \\ &= |\det a|^{-1/2} \sum_{\gamma\in\Gamma} \bar{c}_\gamma \langle f, D_{a^j} L_{(\gamma a\eta a^{-1})^{-1}} \varphi \rangle \\ &= |\det a|^{-1/2} \sum_{\gamma\in\Gamma} \bar{c}_\gamma \, s_j(\gamma a\eta a^{-1}). \end{aligned} \tag{2.14}$$

Note that, since $a\eta a^{-1} \in \Gamma$ (recall Lemma 2.1.1), the term $\gamma a\eta a^{-1}$ is the product of two elements of Γ and therefore the above expression is well-defined. Re-writing (2.14) in operator form one defines the *approximation operator* H to be

$$H s_j(\eta) = |\det a|^{-1/2} \sum_{\gamma\in\Gamma} \bar{c}_\gamma s_j(\gamma a\eta a^{-1}) = s_{j+1}(\eta). \tag{2.15}$$

An analogous procedure yields the *detail operators* G_i

$$G_i s(\eta) = |\det a|^{-1/2} \sum_{\gamma\in\Gamma} \bar{d}^i_\gamma s(\gamma a\eta a^{-1}) = t_{j+1}(\eta) \quad \forall i = 1, \ldots, L. \tag{2.16}$$

Equations (2.15) and (2.16) are analogous to the approximation and detail operators

for the classical discrete wavelet transform, suitably modified by [50] for the CHCDW setting. Their adjoints are

$$H^* s(\eta) = |\det a|^{-1/2} \sum_{\gamma \in \Gamma} s(\gamma^{-1}) c_{\eta a \gamma a^{-1}}, \quad G_i^* s(\eta) = |\det a|^{-1/2} \sum_{\gamma \in \Gamma} s(\gamma^{-1}) d^i_{\eta a \gamma a^{-1}}. \tag{2.17}$$

Thus, (2.15), (2.16), (2.17) provide a mathematical recipe for computing signal synthesis and analysis for CHCDW. There remain a few outstanding questions, such as how to obtain $s_0(\eta) = \langle f, L_{\eta^{-1}} \varphi \rangle$ (i.e. the zeroth level scaling coefficients needed to initiate the cascade algorithm) and how to obtain the filter coefficients. For the latter, the process is somewhat involved and we refer to [50] for some additional discussion regarding the conditions for perfect signal reconstruction as well as [115] for some discussion on identifying filter coefficients in one dimension. A concrete instance of a CHCDW design will be explored in more detail in Section 2.2.1.

As for s_0, in [50], the authors propose committing the usual "wavelet crime" [138]. Indeed,

$$s_0(\eta) = \langle f, L_{\eta^{-1}} \varphi \rangle \approx f(\eta^{-1}(0))$$

provides a reasonable approximation for these zeroth level coefficients. Note that this means s_0 consists of $|B|$ (i.e. the order of B) copies of the input signal since

$$f(\eta^{-1}(0)) = f(b^{-1}(0 + bk)) = f(k).$$

The computational complexity of the CHCDW cascade algorithm is driven by the cost of the filtering operations (2.15),(2.16), (2.17). To implement a convolution

Runtime	# Calls	Line	Code
17.510	172800	89	`x = circshift(s(:,:,bc_i), -delta_row, 1);`
17.360	172800	90	`x = circshift(x', -delta_col, 1)';`
33.038	172800	92	`out(:,:,c_i) = out(:,:,c_i) + c_gamma * x;`

Table 2.1: The three most expensive lines of code from profiling CHCDW software (when performing 100 trials of an analysis task).

of the form (2.15) for an $N \times N$ image we must compute $|B|N^2$ scalar multiplications for each nonzero filter coefficient in the collections $\{c_\gamma\}_{\gamma\in\Gamma}$ and $\{d^i_\gamma\}_{\gamma\in\Gamma,i=1,\ldots,L}$. If one assumes the number of nonzero filter coefficients is constant and independent of image size (e.g. it is 3 or less in all our experiments) the computational cost for a single layer of the CHCDW transform is $\mathcal{O}(N^2)$. The number of stages in the transform is a function of the scaling matrix; for a dyadic scaling matrix $a = 2I$, where I is the identity, there are $\log_2$ steps in the transform yielding a total algorithm cost of $\mathcal{O}(N^2 \log_2(N))$. For a two-dimensional quincunx scaling matrix with $|\det a| = 2$, the number of steps in the transform doubles (as the image width and height are reduced alternately by a factor of 2) but the order of the computation remains the same.

There are two implementations of the CHCDW algorithm the author is aware of, both implemented in the MATLAB programming language. The first is the original (not explicitly optimized) version written by Dr. Benjamin Manning; the second, a newer implementation created by this author which makes careful use of data structures to more efficiently represent signals in the case where all wavelet coefficients can be explicitly retained in main memory using a dense (i.e. non-sparse) representation. Profiling our new implementation (see Table 2.1) demonstrates that

the bulk of the calculations indeed correspond to filtering operations; in particular, they represent the (periodic) translation on the lattice by k and the multiplication and summation entailed by the approximation and detail operators. Runtime from these three lines represents approximately 85 percent of the total algorithm runtime for an image of size 512×512 (computed over 100 trials).

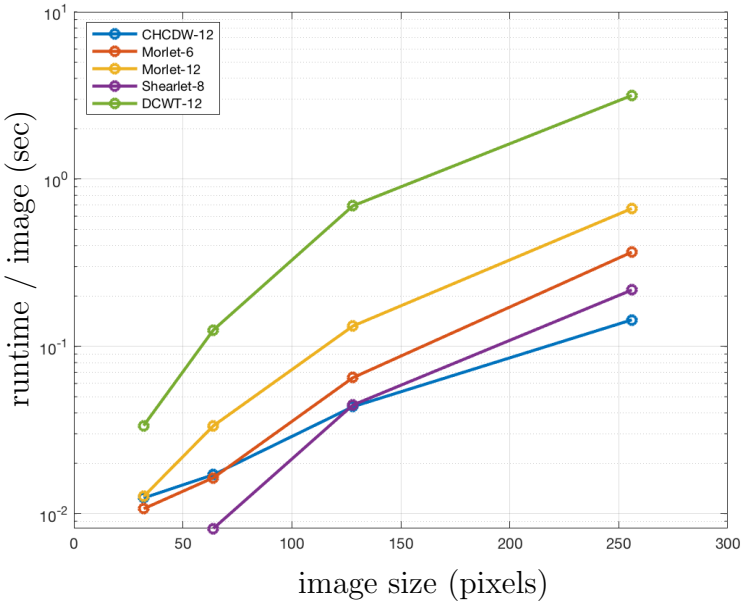


Figure 2.1: Runtimes for various wavelet algorithms (CPU-only implementations) as a function of image size. The y-axis depicts average runtime per image taken over 100 trials (lower is better; note the log scale). Our implementation, CHCDW-12, is among the best performers in this set and performs especially well for larger image size.

Figure 2.1 shows a comparison of runtime performance for various wavelet analysis algorithms. In particular, we compare the unoptimized CHCDW transform (designated DCWT-12) and the new CHCDW implementation (designated CHCDW-12) both using the same 12 element group B and a dyadic scaling matrix $a = 2I$. For reference, we also provide runtimes for a two dimensional wavelet trans-

form included with the scatnet [137] software package and a shearlet transform [59] . In the case of the wavelet transform, we include results for Morlet wavelets with six and twelve angular parameters. For the shearlet transform, we configure the transform to compute 8 directions at each scale (note that the shearlet codes prefer images that are at least 64×64 pixels in size, hence we omit the data point for the 32×32 case). The figure shows the average runtime to compute an analysis transform for a two-dimensional grayscale image as a function of the image dimension (i.e. the N in an $N \times N$ image). All algorithms were implemented in MATLAB and utilize only CPU operations (although we note the authors of the shearlet transform do have a GPU version of their codes available; for fairness, we consider only CPU implementations). All runtimes were computed on a MacBook Air with an Intel i5 CPU with 2 cores (a fairly limited platform for processing). The figure indicates our implementation has a favorable runtime profile. In particular, the difference between the unoptimized and optimized CHCDW algorithm implementations is greater than factor of 10 in the case of images with dimensions 256×256 (mean runtimes of 3.12 vs. 0.15 seconds per image). There are many techniques for improving the efficiency of the convolution operator in general (e.g. see [123]) and it is quite possible that our runtime performance could be further improved; however, given the relatively good performance of our current algorithm it is likely that other factors pose more substantial obstacles to the adoption of CHCDW for practical applications.

2.2.1 An Example Crystallographic Wavelet

In this section we describe a particular CHCDW, developed by Dr. Benjamin Manning and introduced in [50], whose application will be the focus of our subsequent experiments. As mentioned in [50], this wavelet is analogous to (but distinct from) the final example appearing in [96]. For this wavelet, one takes $a = 2I$ and B to be isomorphic to the 12 element dihedral group D_{12} that has the standard presentation

$$\langle x, z \,|\, a^6 = x^2 = e, xax = a^{-1} \rangle. \tag{2.18}$$

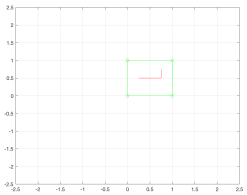
One departure from [96] is that, unlike the typical realization of D_{12}, one defines the elements of B to incorporate shears to facilitate the mapping of hexagonal lattices to canonical lattices. The individual group elements are

$$b_1 = I \qquad b_2 = \text{rotate}_{-90} \cdot \text{shear}_y, \quad b_3 = \text{rotate}_{-90} \cdot \text{shear}_x, \quad b_4 = \text{rotate}_{180},$$

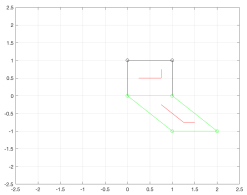
$$b_5 = \text{rotate}_{90} \cdot \text{shear}_y, \quad b_6 = \text{rotate}_{90} \cdot \text{shear}_x, \quad b_7 = \text{flip}_y \cdot \text{shear}_y, \quad b_8 = \text{flip}_{y=x},$$

$$b_9 = \text{flip}_x \cdot \text{shear}_x, \quad b_{10} = \text{flip}_x \cdot \text{shear}_y, \quad b_{11} = \text{flip}_{y=-x}, \quad b_{12} = \text{flip}_y \cdot \text{shear}_x$$

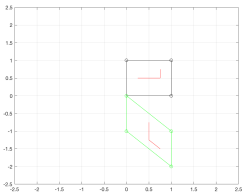
With respect to (2.18) the identity matrix b_1 is of course the group identity element e; elements b_2 and b_7 play the roles of a and x, respectively. A visual representation of the action of the elements B on a unit square is provided in Figure 2.2. In the remainder we will refer to this particular CHCDW as "CHCDW-12", although we observe that it is not the only CHCDW that can be derived from group B with twelve elements (Table 2 in [19] provides a complete taxonomy).



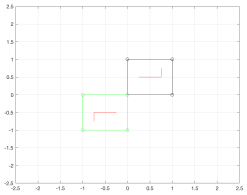
(a) $b_1\chi_S$



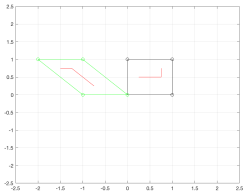
(b) $b_2\chi_S$



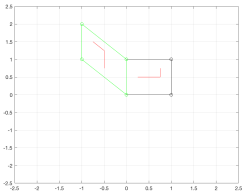
(c) $b_3\chi_S$



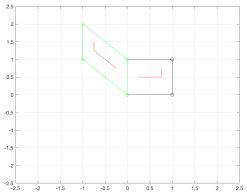
(d) $b_4\chi_S$



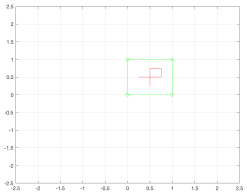
(e) $b_5\chi_S$



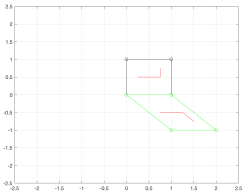
(f) $b_6\chi_S$



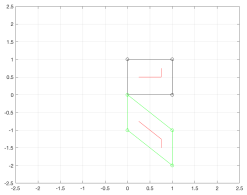
(g) $b_7\chi_S$



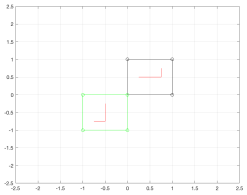
(h) $b_8\chi_S$



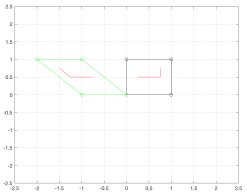
(i) $b_9\chi_S$



(j) $b_{10}\chi_S$



(k) $b_{11}\chi_S$



(l) $b_{12}\chi_S$

Figure 2.2: The unit square χ_S (black and red lines) and the action of elements of B on this region (green and red lines).

While the unit square is useful to gain an intuition regarding the action of elements of B, it is also instructive to consider the specialized subsets of $\mathbb{R}^n$ associated with Theorem 2.1.3. Consider the candidate fundamental region $R = \{(0,0), (\frac{1}{2}, 0), (\frac{1}{3}, \frac{1}{3})\}$. The left panel of Figure 2.3 depicts both R (the triangle labeled $b_1 R$) and also the set $S = \bigcup_{b \in B} bR$. The right panel of the same figure shows nine lattice translates $k \in [-1, 0, 1] \times [-1, 0, 1]$ of S. From inspection it is clear that the complete set of translates of S by $\mathcal{L}$ will result in a tiling of $\mathbb{R}^2$ as required by Theorem 2.1.3 Item **(i)**. For the choice of dyadic scaling matrix $a = 2I$ the normalization property follows immediately since

$$aba^{-1} = (2I)b(2^{-1}I) = b.$$

We also observe that the lattice $\mathbb{Z}^2$ is trivially preserved by dyadic scaling and hence Theorem 2.1.3 Item **(ii)** holds. For the choices $\{b_1, b_5, b_7, b_{12}\}$ and associated translations $\{(0,0)^T, (0,1)^T, (1,-1)^T, (1,0)^T\}$ we demonstrate that (2.7) is indeed satisfied numerically; see Figure 2.4. Thus, we have demonstrated (if somewhat informally) that the conditions of Theorem 2.1.3 all hold and therefore the CHCDW has an associated MRA. That this CHCDW has an associated MRA was of course known by the authors of [50] (it was constructed to have this property); here we have merely made explicit the theoretical justification.

The structure of fundamental region provides insight into the local spatial structure of signals that are likely to be well-captured by this wavelet representation. A demonstration that further illustrates how the geometry of the fundamental

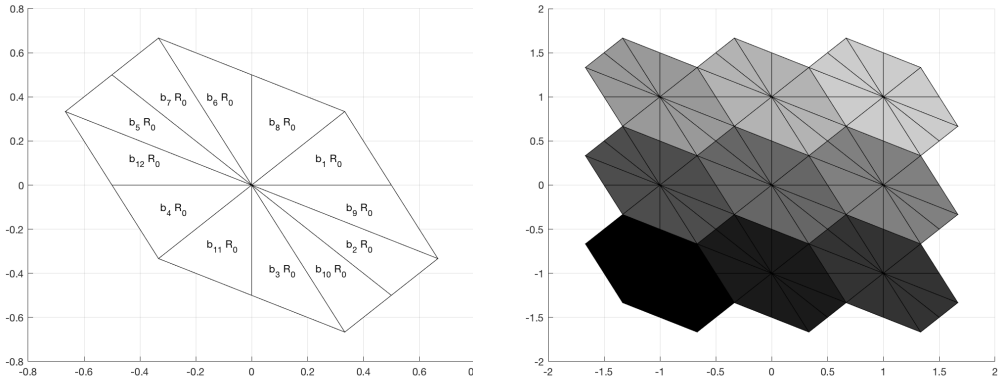


Figure 2.3: Visualizing $S = \bigcup_{b \in B} bR_0$ and its $\mathcal{L}$-tiling property.

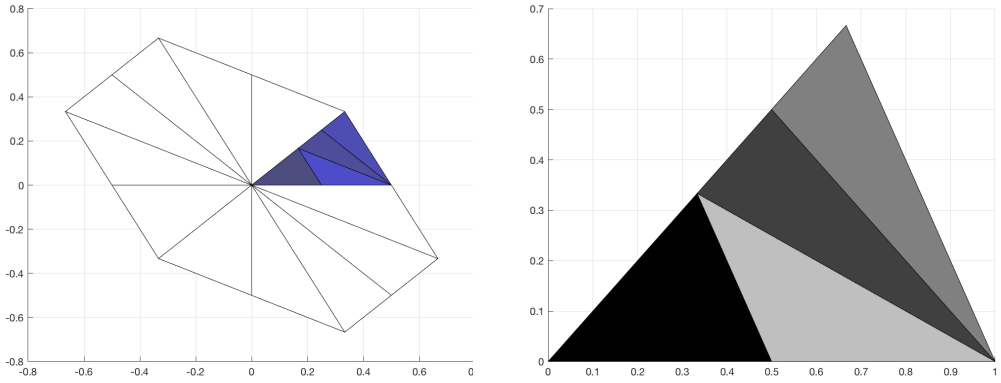


Figure 2.4: Supports for the three multiwavelets ψ_i are denoted by blue triangles (left); each multiwavelet takes a constant value on each blue region. Also shown is a visual demonstration that Theorem 2.1.3 Item **(iii)** holds (right).

Figure 2.5: Image reconstructions using block-sparse wavelet coefficients.

region plays a role is possible by reconstructing a few images using only the coarse scale coefficients. Figure 2.5 shows two few canonical images (left) and their reconstructions from only the wavelet coefficients at or above scale 2^5 and 2^4 respectively. The resulting "Picasso-ification" of the images arises from the structure of the fundamental region. While perhaps not quite as striking as neural network procedures for style transfer (e.g. [68]) the figure does suggest that the Haar-like properties of the transform provide some natural mechanism for introducing a kind of "style" by construction (e.g. vs data-driven techniques as in deep learning).

Coefficients	hyperplane (256×256)	barbara (512×512)	cameraman (256×256)	polygons1 (256×256)	thin-rectangle (256×256)
$t^i_{-1}(\eta)$	0.24	0.58	0.53	0.98	3.15
$t^i_{-2}(\eta)$	0.23	0.53	0.50	0.98	3.30
$t^i_{-3}(\eta)$	0.28	0.63	0.88	1.58	6.97
$t^i_{-4}(\eta)$	0.25	0.73	1.37	3.16	15.46
$t^i_{-5}(\eta)$	0.17	1.14	1.66	6.20	18.03
$t^i_{-6}(\eta)$	0.10	1.80	2.43	11.33	23.21
$t^i_{-7}(\eta)$	0.05	2.69	3.20	15.45	13.75
$t^i_{-8}(\eta)$	0.03	2.22	3.72	23.78	9.38
$t^i_{-9}(\eta)$		3.15			
$s_{-J}(\eta)$	98.66	86.53	85.71	36.53	6.74
Total	100.00	100.00	100.00	100.00	100.00

Table 2.2: Distribution of energy across CHCDW-12 coefficients at each scale for various test images. Numbers in table indicate a percentage relative to the scaling coefficients at scale $j = 0$, i.e. $||s_0(\eta)||_2$.

That the CHCDW transform is orthogonal follows from its MRA property. As an empirical confirmation of this fact, we present a Table 2.2 where, for a variety of images, we compute the energy (i.e. $||\cdot||_2$) for the wavelet coefficients at each scale $\{t^i_j\}_i$ as well as the scaling coefficients s_J. It is interesting to note that, for the natural images, the bulk of the wavelet energy is captured by the scaling coefficients s_J. In contrast, for the binary images of simple geometric shapes the wavelet coefficients capture a relatively larger portion of the orignal signal energy. This suggests that, for algorithms are based on properties of the wavelet coefficients $t^i_j(\eta)$, the CHCDW may be at a relative disadvantage for natural images. We will explore this insight in more detail in Section 2.3.

2.3 Applications

In this section we explore a few practical applications for CHCDW. For all of our experiments we use the CHCDW-12 wavelet introduced in Section 2.2.1.

2.3.1 Directional Analysis

Anisotropic methods, that is, those whose properties differ according to the direction of measurement, have long been of interest in harmonic analysis. For example, in image processing applications the ability to localize structural features (such as edges, segments, corners, etc.) is highly desirable. Classically, continuous wavelet transform (CWT) techniques were employed for these analysis tasks, with discrete, dyadic wavelets typically reserved for synthesis tasks (such as denoising and data compression) [8][2]. In the CWT, angle, scale, and position localization is obtained by filtering signals with the collection of functions

$$A_{A,R}(\Psi) = \{D_a D_r T_y \psi : a \in A, r \in R, y \in \mathbb{R}^2\}, \tag{2.19}$$

where R is a collection of rotation matrices and A is a collection of scaling matrices. This can be seen as a continuous variant of (2.2) where there is more flexibility in selecting the dilations and the function ψ (typical requirements are that ψ be reasonably well-localized in space and frequency and satisfy the minimal admissibility

[2]Even more recently, Mallat elected to use a CWT technique when devising his scattering transform framework instead of a discrete wavelet system, which is telling as he is a pioneer of the latter technique.

requirements for a wavelet [9]). Features $S(a, \theta, b)$ are computed by taking inner products of $A_{A,R}$ with a signal/image (analogous to (2.13)).

In this framework, the properties of ψ and the choices for A, R drive directional and positional sensitivity. Defining precisely what constitutes good directionality can be somewhat application dependent. One precise definition proposed by [8] that a wavelet ψ is *directional* if the effective support of its Fourier transform $\hat{\psi}$ is contained in a convex cone whose apex is at the origin of frequency space. In particular, this implies that the frequency support of $\hat{\psi}$ should be away from zero. One canonical example, the 2-D Marr (or "mexican hat") wavelet

$$\psi_H(u) = (2 - u^T A u) \exp\left(-\frac{1}{2} u^T A u\right),$$

has its frequency support centered at the origin regardless of how one chooses the anisotropy matrix A and therefore fails to satisfy this definition. Indeed, empirical studies have confirmed that this wavelet is better suited for detecting point singularities than oriented edges [9]. In contrast, a prototypical anisotropic tool is the 2-D Morlet wavelet, which is constructed by taking the product of a plane wave and a Gaussian envelope

$$\psi_{Morlet}(u) = \exp(ik_0 \cdot u) \exp\left(\frac{1}{2}\left(\epsilon^{-1} x^2 + y^2\right)\right) + C. \tag{2.20}$$

The parameter $k_0 \in \mathbb{R}^2$ is a wave vector and $\epsilon \geq 1$ is an anisotropy parameter which stretches the Gaussian envelope along the x-axis; together, these two pa-

rameters control the sensitivity to discontinuities along the x-axis. The constant C is a correction term used to ensure the wavelet satisfies the admissibility condition $\hat{\psi}_{Morlet}(0) = 0$ (here, the wavelet is assumed to be suitably regular e.g. $\psi \in L^1(\mathbb{R}^2) \cap L^2(\mathbb{R}^2)$). This wavelet indeed satisfies the above notion of directionality and has also demonstrated excellent angular sensitivity empirically [9]. Note that failing to satisfy the above definition of directionality does not mean a wavelet cannot be used to productively analyze directional content. For example, [8] describe gradient wavelets designed to detect corners in images which do not satisfy this strict definition. Empirical analysis therefore has a role to play in characterizing wavelet behavior.

One method used to quantify the resolving power of a wavelet is to analyze the transforms of particular signals. For example, [8] recommend to analyze the perfomance of ψ with three general types of tests:

1. Evaluating the *impulse response* of ψ by evaluating a Dirac delta function.

2. Using the autocorrelation of ψ to evaluate the *correlation length* in each variable a, θ, b,

$$K(a, \theta, y|1, 0, 0) = \langle \psi_{a,\theta,b} | \psi \rangle = \frac{1}{a} \int \psi(x) \overline{\psi} \left(\frac{1}{a} r_{-\theta}(x - y) \right) dx.$$

3. For testing specific properties, such as the ability to detect a discontinuity or angular selectivity in a particular direction, one may use custom *benchmark signals* to characterize wavelet performance.

Detailed examples of these analyses for the Morlet and Marr wavelets can be found in [9].

A number of more recent directionally sensitive methods fall under the umbrella of CDW systems, such as ridgelets [32], curvelets [33], directional Gabor systems [73, 117], and shearlets [78]. Of these methods, shearlets have proven particularly popular in applications. While provably optimal for the class of cartoon-like functions, there remains interest in understanding their directional analysis properties. For example, [145] provides new insight into the representational efficiency of shearlets in the special case of discontinuities taking the form of two-dimensional hyperplanes and [90] considers angular sensitivity and resolution of various wavelets in the context of digital mammography images. These can be seen as instances of the "benchmark" signal analysis suggested by [8].

For the CHCDW-12 wavelet, we do not have quite the same flexibility as compared to some of the classical CDW methods. For example, the structure of the group B limits to some extent how explicitly one can parameterize a notion of angle beyond what is inherently captured by the elements of B. Additionally, the orthogonality property of CDW suggests that autocorrelation analysis will not be productive. Finally, the Haar-like structure of the scaling function in the time domain indicates that this wavelet will not have a frequency domain profile that will be well contained within a cone as required by the Definition of [8]. Nevertheless, performance on benchmark signals is still of relevance and may even be particularly interesting since the expected results are not entirely clear from the outset (unlike the Morlet wavelet, where k_0 and ϵ from (2.20) have a slightly more intu-

itive interpretation when it comes to detecting discontinuities). In the remainder of this section we ask what kind of directional capabilities might be possible with the CHCDW-12, despite the relatively strong constraints upon its construction.

2.3.1.1 Empirical Studies

In this section we consider a "benchmarking" experiment in the spirit of [8]. However, we take a slightly different approach in that we investigate angular sensitivity relative to globally pooled representations derived from CDW wavelets. This pooling operation averages over spatial location information in the spirit of the translationally invariant representations described in [26]. Our goal is to obtain insight into the behavior of CHCDW-12. To do so, we will compare it with two other directional representations: shearlets and Morlet wavelets. While the benchmarking study we pursue here may not be ideal for these latter two systems (for which we may have additional analytical understanding of their directional sensitivities) the ultimate motivation is to understand the relative behavior of CHCDW-12.

Fix a benchmark signal, and let $x_i \in \mathbb{R}^d, i = 1, \ldots, N$ be the digital representation of rotations of the benchmark by angles $y_i \in \mathbb{R}, i = 1, \ldots N$ (i.e. N is the number of angles considered). For example, Figure 2.6 shows an example where the benchmark signal is a thin rectangle designed to emulate a simple edge-like structure. The goal is to estimate y_i from pooled CDW features derived from x_i. Let ψ_j be a wavelet associated with a CDW system (e.g. CHDW-12). Then we define an associated pooled wavelet representation for each image via $c_{i,j} = ||x_i \star \psi_j||_1$. The

motivation for this particular representation is two-fold: marginalizing over spatial parameters distills the representation down to features most closely associated with directional properties of interest and this representation is directly related to a first-order scattering transform (one that is maximally translation-invariant as it does not compute any local descriptors). As observed by [26] this is a fairly crude signal representation; however, our goal here is to gain understanding about directional sensitivity and not to optimize the solution of a particular problem.

As a model of the relationship between the $c_{i,j}$ and the y_i, we frame a LASSO [142] regression

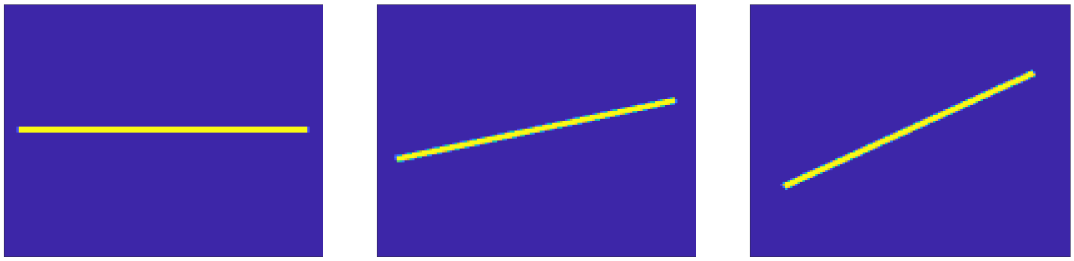
$$(\hat{\alpha}, \hat{\beta}) = \arg\min \left\{ \sum_{i=1}^{N} \left(y_i - \alpha - \sum_j \beta_j c_{i,j} \right)^2 \right\} \quad \text{subject to } \sum_j |\beta_j| \leq \lambda,$$

where $\hat{\alpha}, \hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_d)$ are regression coefficients and λ is a tuning parameter. We then use as a metric for "angular sensitivity" the mean square error (MSE) of the residual,

$$MSE(y, \hat{y}) = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2,$$

where $\hat{y}_i = \alpha + \sum_j \beta_j c_{i,j}$. In all our experiments the choice of λ is determined by a 10-fold crossvalidation procedure; we use the largest value of λ where the corresponding estimate is within one standard error of the minimum MSE computed over a range of possible values.

Benchmark signals for these experiments consist of simple geometric shapes that have been rotated about the center of the image. All images are binary and



(a) 0 degrees (b) 15 degrees (c) 30 degrees

Figure 2.6: Examples of rotating a benchmark signal (here, an edge-like structure). The task is to estimate the rotation angle from a globally pooled wavelet representation.

128×128 pixels in size. For the wavelet and shearlet systems we use the same implementations referenced in Section 2.2, namely the Morlet wavelets included as part of the ScatNet scattering transform framework [137] and the 2D shearlet implementation of [59]. For the Morlet wavelet systems we employ systems with 4, 6, and 12 equally spaced angles and each having $\log_2(128) = 7$ scales. For the shearlet system, we employ systems with 8, 16, and 32 angles each having 3 scales. In the case of the shearlet system, the number of angles must be a power of two and we use the largest number of scales supported by the software for images of 128×128 pixels. For CHCDW-12, we use our own implementation as described in Section 2.2. When reporting results we provide, for each system, the number of dimensions in the associated representation (recall we marginalized over the two spatial dimensions) and the mean square error (MSE) for each feature/signal pair.

Edge-like Signals For our first experiment, we focus on benchmark signals with edge-like structure. We begin with the thin edge-like rectangles depicted in Figure 2.6, which we then rotate from 0 through 135 degrees and solve the aforemen-

Wavelet System	# feature dimensions	MSE
Morlet-4	28	1.55
Morlet-6	42	1.75
Morlet-12	84	1.74
Shearlet-8	25	1.81
Shearlet-16	49	1.71
Shearlet-32	97	1.84
CHCDW-12	84	2.88

Table 2.3: Mean square prediction errors for the edge-like benchmark signal shown in Figure 2.6. The naming convention"*a*-*b*" indicates that CDW system *a* was configured to use *b* "directions". We observe generally low reconstruction error for all three algorithms, but observe that CHCDW-12 appears less well suited comparatively.

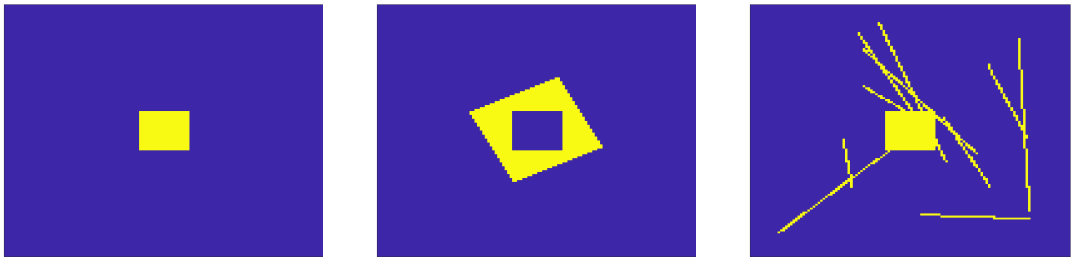
tioned LASSO problem. The results are summarized in Table 2.3. While all CDW systems demonstrated a good ability to capture directional information for this benchmark signal, it appears that CHCDW-12 is least-well adapted of the three to this signal type. This is perhaps unsurprising as the Morlet wavelets in particular were originally designed to capture fine edge-like structures.

As a generalization to less synthetic edge-like signals, we also ran a similar experiment where the benchmark signals are now digits from the MNIST data set. In this case, we have grayscale images which have been resized from the native dimensions of 28×28 to 32×32 pixels (to accomodate wavelet systems that prefer images with dimensions that are a power of 2). The results are summarized in Table 2.4. In general, these results are consistent with those from the thin rectangle experiment described previously. We also remark that this relative performance on MNIST is qualitatively consistent with findings involving a fundamentally different Haar-type wavelet in the context of a scattering network [39].

	MNIST digit									
Feature Type	0	1	2	3	4	5	6	7	8	9
CHCDW-12	0.76	0.66	0.70	0.73	0.70	0.71	0.72	0.73	0.74	0.73
Morlet-4	0.67	0.65	0.68	0.65	0.69	0.66	0.66	0.68	0.64	0.67
Morlet-6	0.70	0.63	0.66	0.64	0.68	0.66	0.66	0.69	0.66	0.69
Morlet-12	0.71	0.65	0.69	0.69	0.69	0.70	0.70	0.68	0.71	0.70
Shearlet-4	10.81	1.43	2.30	4.14	1.47	2.44	2.60	2.91	7.22	2.80
Shearlet-8	3.25	0.87	0.97	1.59	0.97	1.20	1.68	1.33	2.26	1.11

Table 2.4: Mean MSE for angular regression on MNIST digits. Again, while performance is reasonable throughout, results suggest that the Morlet wavelet may have a slight advantage in this setting. Note also the relatively small sizes of these images (rescaled to 32×32 pixels) is not ideally suited for the shearlet codes.

Geometric Signals in Noise For our second set of experiments, we generalize beyond edge-like structure to signals with more substantial spatial support. In particular, we will now consider rotations of a square object in isolation and in the presence of certain kinds of noise; the first noise model introduces a confuser square which circumscribes the target square and the second noise model introduces randomly drawn edges which are overlaid in the scene. Examples of images associated with these three models are presented in Figure 2.7 and Table 2.5 provides the corresponding regression results. Before discussing the results, we first pause to mention that for more traditional noise models (e.g. additive Gaussian noise or "salt-and-pepper" binary noise) the CHCDW-12 system performs quite poorly compared to the more conventional shearlet and wavelet systems. Some of our empirical explorations suggested that the CHCDW-12 tends to perform better for signals characterized by piecewise constant elements with non-trivial spatial support. These informal observations motivated the exploration into these somewhat less conventional noise settings. Our first noise model therefore considers a situation where there is an interesting overlap between piecewise constant signals and our



(a) No noise. (b) Distractor object. (c) Overlaid lines.

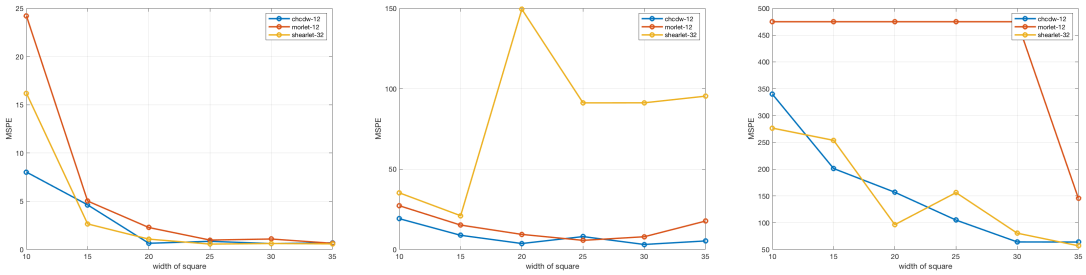
Figure 2.7: Visualization of noise models used in our angular regression experiments. The goal is to estimate the rotation angle of the square in the center of the image (shown here with 0 rotation). All images are binary and 128×128 pixels in size.

second experiment considers settings where fine edge-like structure may act as a distractor rather than the primary signal of interest. For example, one might imagine this as a very crude approximation of cluttered LIDAR scenes where weaker returns from nearby obstacles introduce edge-like noise.

Based on Table 2.5 it appears that when equipped with a sufficient number of directions, all wavelet systems do a reasonable job solving this regression problem. The shearlet system appears to struggle in the setting where there is a distractor object and the Morlet system appears to be misled when fine edge-like structure is a distractor rather than the signal of interest. In all of these settings, the CHCDW-12 system seems to be relatively robust. One might naturally ask whether these results generalize across scales. Figure 2.8 shows results for experiments where the dimension of the target square varies. A similar overall trend is evident; overall, these empirical studies suggest that there may be certain signal "regimes" for which unconventional wavelets can potentially provide some value relative to the more commonly studied directional systems.

Wavelet System	# dims	No noise	Distractor object	Overlaid lines
CHCDW-12	84	**0.72**	**6.06**	**79.36**
Morlet-7-4	28	312.49	350.97	≥ 475
Morlet-7-6	42	9.55	56.75	≥ 475
Morlet-7-12	84	2.21	**11.71**	≥ 475
Shearlet-3-8	25	9.97	292.32	221.78
Shearlet-3-16	49	1.90	144.18	197.23
Shearlet-3-32	97	**0.78**	59.00	**127.49**

Table 2.5: Mean square error for our angular regression problem under the noise models shown in Figure 2.7). Error estimates are for an angular regression problem where the object of interest is rotated from 0° to 75° degrees.



(a) No noise. (b) Distractor object. (c) Overlaid lines.

Figure 2.8: MSE for estimating rotation angles of squares as a function of the target shape size. The three panels correspond to the shapes depicted in Figures 2.7a to 2.7c. The CHCDW-12 wavelet is generally among the best performer across all scales in these three scenarios.

Conclusions In summary:

1. These experiments suggest that, while all three systems provide good performance for edge-like structure, the Morlet system may have a slight advantage in this regard. This is not an absolute statement, however, and other considerations (image size, global pooling, etc).

2. Our noise model experiments hint that there may be some advantage in certain settings, e.g. where fine edge-like structure is not a feature of interest but rather a source of noise.

3. Not shown explicitly here were experiments with more traditional noise models (e.g. additive Gaussian noise) for which CHCDW-12 appears to be entirely ill-suited. This suggests that the practical applications of this wavelet may be somewhat specialized.

2.3.2 Classification via Scattering Transforms

2.3.2.1 Overview

Scattering transforms, originally devised by Stéphane Mallat [26, 114], are a framework for generating signal representations inspired, in part, by the recent widespread success of deep neural networks (DNNs). Despite their unquestionably good performance and the vast amount of research attention [104], DNNs are still not completely understood from a mathematical perspective. Even the relatively simple feedforward convolutional neural networks lack a complete mathematical character-

ization and this task becomes even more challenging as architectural complexities are added (recurrence, dropout, exotic normalizations, etc.). Designing DNNs and tuning their hyperparameters (e.g. number of layers, learning rates, filter sizes) is non-trivial and requires some combination of numerical experimentation and manual craftwork. Scattering transforms start with the cascade of filtering and nonlinear operators that lie at the heart of DNNs and seek to place this structure on firmer mathematical footing. In particular, the goal is to provide a setting where principled statements can be made about the properties of the resulting representation.

Unlike DNNs, scattering frameworks (in their original incarnation) make use of filters that are designed a-priori, rather than learned from data. In essence, they are cascades of directional wavelet filters interleaved with pointwise nonlinearities. The underlying mathematical motivation is to eliminate uninformative variability from the feature representation for the signal processing problem at hand. In image classification problems, for instance, affine transformations such as rotation and scaling typically do not change the underlying signal content but have substantial impact to the resulting Euclidean signal representations. In the classical handwritten digit classification problems such as MNIST, translating a digital image of a digit does not change its class but there is a large ℓ_2 distance between the original and translated digital signals. Such variability is not helpful to a downstream classifier; this motivates a mathematical desire to produce feature representations that are invariant to these actions.

Of course not all variability is irrelevant. There is other (often nonlinear) variability in signal representations whose magnitude can be expected to impact the

signal content proportionally. Back to the handwritten digit classification setting, small elastic distortions do not change the signal content (digit type) but larger deformations could represent a fundamental change in the underlying signal (e.g. deforming a digit 1 into a 7). Thus, one does not desire invariance to such deformations, but instead prefers a representation which is stable to the deformation, meaning it linearizes the elastic deformation so that any variability can be more readily adjudicated by a classifier.

Under certain assumptions related to the filter banks, Mallat is able to show that his scattering tranform feature representations exhibit translation invariance (in a limiting sense), stability to diffeomorphisms, and an energy preservation property that ensures the resulting representation is non-trivial. However, Mallat's scattering transform framework is designed for continuous (or semi-discrete) transforms and the associated guarantees require the underlying wavelets satisfy a very specific admissibility criterion that is not readily satisfied by most modern wavelet systems. Thus, the original scattering is a beautiful mathematical accomplishment but one that entails relaxation and approximations in order to realize a digital implementation.

In an attempt to make scattering transforms more practical, Bölcskei and Wiatowski introduced their own variant of Mallat's scattering that provide separate notions of translation invariance and stability [146, 147, 148]. While it does not provide exactly the same guarantees, their framework imposes fewer requirements on the underlying wavelet system and also introduces additional constructs relevant to DNNs, such as explicit notions of inter-layer pooling. In particular, as long

as the wavelet is suitably bounded (which will always be true in the discrete setting), the nonlinearities are mild (Lipschitz-continuous), and the pooling operations satisfy mild conditions, then one obtains a notion of translation invariance which varies as the network grows deeper and Lipschitz continuity of the feature extractor. These requirements are sufficiently benign that it opens the door to devising hybrid representations consisting of both learned and hand-crafted filters.

A number of other extensions to these scattering transforms have been explored in the litereature as well. This include different wavelet systems, such as Gabor systems [48, 49, 53, 108, 109], roto-translation groups [119], and time-frequency scatterings [7]. There has also been prior work with scattering Haar wavelets [39, 40]; however, instead of building directly upon composite dilation wavelets this work considers hierarchies of additions, subtractions, and absolute values over pairs of coefficients. The specific pairings used in a given network are selected by various optimization strategies, whereas in our setting the MRA determines the structure of the calculations. Scattering transforms have also been generalized to Lipschitz systems [13] and hybrid representations that include both learned and manually crafted wavelets [120].

Our primary interest is in (fully) discrete wavelets and numerical implementations thereof. Even if one sets aside this computational perspective temporarily, the CHCDW wavelet system does not satisfy the wavelet admissibility criterion associated with Mallat's scattering transform and it would require some non-trivial modifications to the construction in order to better localize the frequency. Therefore, our focus in the remainder is upon CHCDW and the discrete scattering framework

of Bölcskei and Wiatowski. We begin by describing more precisely the scattering framework itself, which is fairly consistent between the two systems. Then we provide some additional details regarding the discrete scattering framework and briefly conform that CHCDW is indeed compatible with this framework. Finally we present our main contribution, which is a comparative analysis of CHCDW with two other popular directional composite dilation wavelets.

2.3.2.2 Discrete Scattering Transforms

Both the scattering transforms of Mallat and of Bölcskei and Wiatowski are based on cascades of filtering and nonlinear operations whose cascades are arranged in a tree-like structure. This architecture is depicted in Figure 2.9. While both share the same general framework, the two approaches differ in terms of requirements upon the operations and the resulting guarantees they provide. For reasons mentioned in the previous section, we will limit our focus on the discrete framework of [148] and adopt the corresponding notation in the remainder.

The core building block of the discrete scattering transform is a triple (or *module*) denoted by (Ψ_m, ρ_m, P_m) where Ψ_m is a bank of filters used to implement a convolutional transform, ρ_m is a pointwise nonlinearity, and P_m is a pooling operation; the subscript m denotes that these operations are used at layer m of the scattering tree. The complete collection of these operations (i.e. for the entire scattering tree) is termed a *module sequence* and is denoted $\Omega := ((\Psi_m, \rho_m, P_m)_{0 \leq m \leq M})$. The precise definitions and conditions upon these components is described below.

The filter bank Ψ_m: To be admissible in the sense of Bölcskei and Wiatowski, the filter bank used to implement the convolutional operations must satisfy a frame property. Let $I_N := \{0, \ldots, N-1\}$ be a finite index set and let $H_N := \{f : \mathbb{Z} \to \mathbb{C} :\}$ denote the set of N-periodic discrete-time signals Let Λ be a finite index set. Then a collection of filters A collection of filters $\Psi_\Lambda = \{g_\lambda\}_{\lambda \in \Lambda}$ is called a *convolutional set with Bessel bound* $B \geq 0$ if

$$\sum_{\lambda \in \Lambda} ||f \star g_\lambda||_2^2 \leq B||f||_2^2, \quad \forall f \in H_N. \tag{2.21}$$

As observed in [148], this condition is equivalent to

$$\sum_{\lambda \in \Lambda} |\hat{g_\lambda}[k]|^2 \leq B, \quad \forall k \in I_N. \tag{2.22}$$

and hence every finite set $\{g_\lambda\}_{\lambda \in \Lambda}$ is a convolutional set with Bessel bound $B^\star = \max_{k \in I_N} \sum_{\lambda \in \Lambda}$. The core requirement upon the filter banks Ψ_m is that they are convolutional sets with Bessel bounds; clearly this condition is very mild and opens the door to a number of interesting opportunities. Since the CHCDW-12 system is constructed from a finite linear combination of compactly supported signals (recall Equations (2.15) and (2.16)) this condition holds. As it is an orthonormal system, we also know that the Bessel bound $B = 1$. The only detail is the original condition is stated in terms of N-periodic signals, which we accomodate in our digital implementation by working with the periodization of all signals (following [50]).

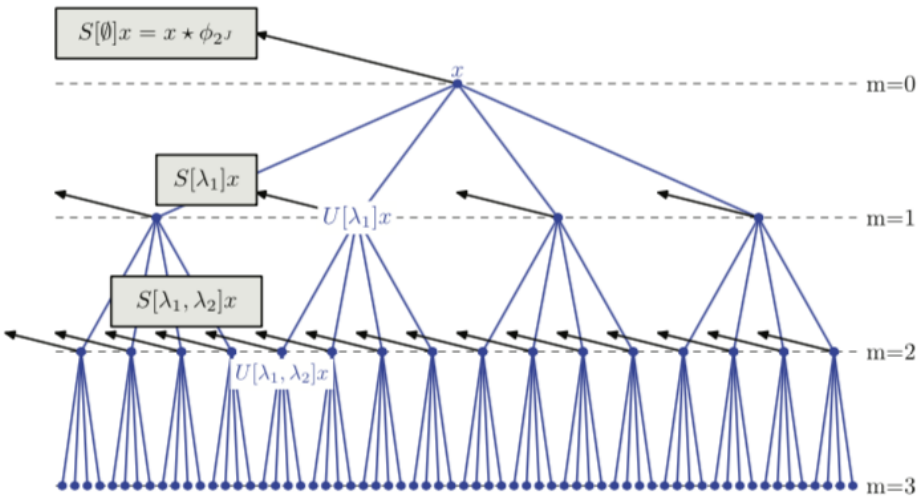


Figure 2.9: Scattering transform framework. Figure 2 in [26].

Non-linearities ρ_m: The non-linearity $\rho_m : \mathbb{C} \to \mathbb{C}$ is a function which transforms points locally. In particular, to be a module sequence requires that act pointwise and satisfy the Lipschitz property $|\rho(x) - \rho(y)| \leq L|x - y|$ for all $x, y \in \mathbb{C}$ and for some scalar $L > 0$. A number of candidate functions satisfy this requirement; of particular interest to us is the modulus non-linearity $\rho(x) = |x|$ which is clearly pointwise and has Lipschitz constant 1. This is the nonlinearity we will use in all of our scattering experiments.

Pooling Operator P: Let $P : H_N \to H_{N/S}$ denote a pooling operator where $N, S \in \mathbb{N}, N/S \in \mathbb{N}$ denote the size of the index set and the "pooling" factor which governs the rate of dimension reduction. As in neural network architectures, pooling operations within scattering networks act to reduce dimensionality. For a module sequence the requirement is the pooling operation be Lipschitz, i.e.. $||Pf - Pg||_2 \leq R||f - g||, \forall f, g \in H_N$ for some scalar $R > 0$. Of particular interest to us is the subsampling pooling operator $(Pf)[k] = f[Sk]$ which has Lipschitz constant 1. This

is the pooling operator we will use in all our scattering experiments (usually with $S = 1$, i.e. no pooling).

These three conditions are all relatively mild and there is no substantial work required to confirm that for reasonable choices of pooling and downsampling operators the CHCDW-12 system is readily admitted into this framework. In a semi-discrete (i.e. continuous) setting the admissibility condition is slightly more complex; however, a number of composite directional systems, including shearlets, have been demonstrated to fit into this framework [146].

A module acts upon a signal x via the operator $U_m[\lambda_m]$

$$U_m[\lambda_m]x := P_m(\rho_m(x \star g_{\lambda_m})), \tag{2.23}$$

that is, one convolves the signal with an element from the filter bank, applies an pointwise nonlinearity, and then optionally downsamples via the pooling operation. The ultimate signal representation is built from cascades of these operations corresponding to different combinations of filtering operations along a "path"

$$q = (\lambda_1, \lambda_2, \ldots, \lambda_m) \in \Lambda_1 \times \Lambda_2 \times \ldots \times \Lambda_m.$$

In particular, the overall operator along such a path is denoted

$$\begin{aligned} U[q]x &= U[(\lambda_1, \lambda_2, \ldots, \lambda_d)]x \\ &= U_d[\lambda_d] \ldots U_2[\lambda_2]U_1[\lambda_1]x. \end{aligned} \tag{2.24}$$

These operators U correspond to nodes in the scattering tree Figure 2.9.

As in Mallat's original scattering transform, these $U[\lambda]$ represent intermediate calculations but not the ultimate representation. Instead, the there is a separate filter χ_m which is responsible for extracting the output representation from each node

$$S[\lambda_1, \ldots, \lambda_m]x := (U[\lambda_1, \ldots, \lambda_m]x) \star \chi_m.$$

Since wavelets are covariant (not translation invariant) part of the role of the feature extraction operator is to add local translation invariance by averaging.

The final signal representation for x from an M-layer scattering, denoted $\Phi_\Omega^m(x)$, is simply the union of the outputs $S[\lambda]$, i.e.

$$\Phi_\Omega^m(x) := \bigcup_{m=0}^{M-1} \{S[q]x\}_{q \in \Lambda_1^m}$$

In return for adopting the above framework and assumptions, one gains a number of benefits. For example, one can obtain a global stability results A global result

Theorem 2.3.1 (Theorem 1 from [148]). *Let the admissible module sequence* $\Omega = ((\Psi_m, \rho_m, P_m))_{0 \leq \leq M-1}$ *have a Bessel bound* $B_m > 0$ *and Lipschitz constants* $L_m > 0, R_m > 0$ *that satisfy*

$$\max_{0 \leq m \leq M-1} \max\{B_m, B_m R_m^2 L_m^2\} \leq 1.$$

Then the feature extractor Φ_Ω *is Lipschitz-continuous with constant* $L_\Omega = 1$*, i.e.*

$$|||\Phi_\Omega(f) - \Phi_\Omega(h)||| \leq ||f - h||_2,$$

for all $f, h \in H_N$ *where the norm in feature space is defined as*

$$|||\Phi_\Omega(f)|||^2 := \sum_{m=0}^{M-1} \sum_{q \in \Lambda_1^m} ||S[q]f||_2^2.$$

Other results include a global energy constraint on the resulting features, a deformation sensitivity result, and more refined results for the space of cartoon-like functions. Since our goal here is not to further refine any of these bounds, we do not elaborate further and instead refer the reader to [146] for complete details and discussion.

2.3.2.3 Empirical Studies

Now we consider empirical performance of a discrete scattering tranform framework equipped with the CHCDW-12 wavelet. Our scattering transform (termed here "HaarNet") uses the CHCDW-12 system at all depths, i.e. for $\Psi_d = \mathcal{A}_{a,\Gamma}\{\Psi\}$ from (2.6). For the pointwise nonlinearity we use $\rho_d(x) = |x|$ and we use a "no pooling" pooling operator $(Pf)[k] = f[k]$ (although we also experimented with dyadic downsampling in other studies not reported here).

In all our experiments we confine the depth of our scattering tree to at most $m = 2$ and we report results on the MNIST data set [103]. Our reason for focusing

on MNIST is twofold: first, it is a well-established baseline and both main scattering frameworks in the literature report classification results using this dataset. Second, it is of sufficiently modest size that it does not require extensive computational resources to evaluate. We also observe that, based on our directional analysis from Section 2.3.1.1, we might expect that the edge-like nature of MNIST puts the CHCDW-12 scattering at a comparative disadvantage. While our experiments will bear this out to some degree, we will also describe settings under which the picture may be somewhat different.

As another caveat, we point out that dimension reduction is a key practical aspect of realizing a scattering transform. In the seminal application paper for Mallat-style scattering the authors leverage the observation that certain paths along the scattering tree tend to carry minimal energy. These "frequency decreasing paths" provide opportunities for on-the-fly dimension reduction, an idea analyzed empirically in detail in [26]. As of this writing we do not have a direct analog of these frequency decreasing paths for the CHCDW-12 framework, and developing this idea further is something we are actively pursuing. Consequently, while we have a few heuristics for helping to control the dimensionality, a more complete dimenstion reduction program is a necessary and logical next step.

Here we will compare our scattering on MNIST with those of the *ScatNet framework* based on Mallat-style scattering [2, 136] as well as *FrameNet* which is made available by the authors of [146]. In order to facilitate comprison, we attempt to eliminate some variables which might confound the results. In particular, we omit post-scattering dimension reduction and we also use only linear SVMs for our

classification experiments. We observe that this setting does not necessarily lead to state-of-the-art performance on MNIST (e.g. [26] obtain slightly better overall results by employing nonlinear SVMs and PCA-based dimension reduction). However, for the purposes of comparing frameworks, such additional sophistication is unhelpful. For the SVM implementation we used the popular LIBSVM package [37] and used standard $k-$fold crossvalidation techniques to select the box constraint hyperparameter. All of our experiments were conducted using the MATLAB software package.

Since many wavelet implementations (including ours) require inputs whose sizes are a power of two, as a preprocessing step we first resized all MNIST digits from their native 28×28 shape to 32×32. We also note that the individual scattering framworks had the ability to resize these images "under the hood"; we do not do so for our CHCDW-12 scattering and instead work at the native resolution of the input. Algorithm 1 summarizes the evaluation procedure.

Table 2.6 shows classification results for the MNIST data set. Reported are overall error rates as a function of number of training examples. All results are based on a linear SVM with no post-scattering dimension reduction. The baseline configurations for Morlet and Shearlet reflect defaults based on existing publications/software. We also attempted a number of different configurations, but empirically observed those reported by the original authors to be superior. In geneaal, all algorithms exhibit reasonable performance on this data set, with the Morlet/ScatNet configuration delivering the best overall performance. This is not surprising given our earlier experiments in directional analysis. Note also that the CHCDW-12 scat-

Algorithm 1 Procedure used to generate results in Tables 2.6 and 2.7.

```
1: procedure CLASSIFY_MNIST(mnist, m, n_feats_max, do_blur)
2:     ▷ Pre-processing
3:     mnist32 = imresize(mnist, [32,32], 'bilinear')
4:     if (do_blur) then
5:         mnist32 = imgaussfilt(mnist32,5)
6:     end if
7:     ▷ Feature Extraction
8:     x_morlet = ScatNet(mnist32.x, m)
9:     x_shearlet = FrameNet(mnist32.x, m)
10:    x_chcdw = HaarNet(mnist32.x, m)
11:    ▷ Classification
12:    for x_algo in {x_morlet, x_shearlet, x_chcdw} do
13:        x = normalizeFeatures(x_algo, [-1,1])
14:        c_svm = selectHypers(trainSubset(x), n_train=300, k=3)
15:        for n_train = 300, 500, ..., 5000 do
16:            model = libsvmTrain(trainSubset(x), n_train, c_svm)
17:            y_hat = libsvmTest(testSubset(x), model, c_svm)
18:        end for
19:    end for
20: end procedure
```

tering transforms are not improving with added depth; this is likely due to a lack of on-the-fly dimension reduction mentioned previously. Addressing this is an area for future research.

It is important to note that FrameNet, when configured with a different (non-shearlet) wavelet and more training examples, has demonstrated state-of-the-art performance on MNIST (described in full detail in [148]). Any relative under-performance here is due to the shearlet configuration (which is not ideally suited for small images like MNIST) . and is not a flaw of the framework itself.

Table 2.7 shows results from a second experiment where we apply a Gaussian blur (with standard deviation $\sigma = 5$) to the MNIST data as a preprocessing step to emulate a particular kind of noise. While the severe blurring has degraded the performance of all three algorithms relative to the original data set, it appears that

# Train	FrameNet		ScatNet		HaarNet	
300	21.2	13.20	7.69	8.67	11.93	13.09
500	13.63	7.59	5.85	3.79	6.59	7.57
700	10.97	5.99	5.03	3.23	5.55	5.99
1000	10.12	5.12	4.42	2.70	4.91	4.63
2000	8.27	4.07	3.08	2.03	3.59	3.30
5000	6.28	2.84	2.11	1.41	2.52	2.41
Wavelet	Shearlet		Morlet		CHCDW-12	
Scattering order	1	2	1	2	1	2
# dimensions	1089	9801	400	3856	12288	20352

Table 2.6: MNIST scattering experiments. Reported are overall error rates as a function of number of training examples. All results are based on a linear SVM with no post-scattering dimension reduction.

for these parameter settings the CHCDW-12 demonstrates slightly more robustness for small training data set sizes. The difference between the performance of Morlet and CHCDW-12 for the Gaussian blurring case is significant and a McNemar's test [132] rejects the null hypothesis that the classifier performances are equivalent at the 5% level. As with our previous studies of directional analysis, this suggests that there may be classes of signals for which the Haar-type wavelet can provide some form of advantage.

2.4 Conclusions

This chapter considered new applications of Haar-type directional wavelets for supervised signal processing problems. Our investigations are completely new in this regard; past investigations of practical applications of this wavelet have been confined to denoising applications. As part of this investigation, we embedded this wavelet into a scattering transform framework and conduct numerical exper-

# Train	Morlet	CHCDW	Shearlet
300	31.52	**27.90**	31.8
500	-	**19.42**	25.03
700	-	**16.55**	21.22
1000	19.05	**14.05**	17.77
2000	11.75	**11.19**	13.47
5000	**8.74**	8.79	10.44
num. dimensions	400	12288	1089

Table 2.7: Blurred MNIST scattering experiments for $m = 1$ and a linear SVM. Reported are overall error rates as a function of number of training examples. Note that baseline configurations for Morlet and Shearlet reflect defaults known to work well for this problem based on publications/software. However, there is a large discrepancy in the number of dimensions used by each approach and additional hyper-parameter tuning of these results could change the outlook.

iments. We demonstrate good, but not best-of-breed, classification performance on the MNIST data set and identify different signal types where the Haar-based scattering may be more effective. We also identify a few potential future research directions where this technique might provide unique advantages. While our focus has been largely upon characterizing performance for a given signal processing task, this work fits into a broader research agenda to develop robust data representations that relate to state-of-the-art algorithms such as deep learning. The importance of a robust representations becomes especially important in large scale data-driven tasks; in Chapter 3 we will explore in detail one particular motivating example.

Chapter 3: Practial Implications of Instability

A key theme of Chapter 2 is the desire to characterize mathematical properties of feature representations used for machine learning. The energy preservation properties of frames and the theoretical stability of scattering transforms are two examples. To help motivate the potential value of such properties, we explore in this chapter a particular class of undesirable behavior which has been observed within data-driven deep neural networks. These so-called *adversarial examples* (AE) [139] demonstrate that modern deep learning algorithms are not robust to certain small magnitude perturbations in the signal input space. We present here a new analysis of AE in the context of remote sensing applications. Unfortunately, the CHCDW wavelets investigated in Chapter 2 are not well-suited for the type of images that arise naturally in this setting; as a result, we do not explore the direct application of CHCDW wavelets in this study. However, adapting the fundamental ideas of stability to this setting remains a fertile direction for future work.

In the rest of this chapter we consider physically-realizable attacks against machine learning algorithms used in remote sensing applications. In particular, we focus on AE in the context of satellite image classification problems. This setting introduces a number of subtle challenges that are not fully addressed by current research

focused on ground-based natural image data. Our research goal is to investigate these unique aspects. Using a recently curated data set and associated classifier, we provide a preliminary analysis of adversarial examples in settings where the targeted classifier is permitted multiple observations of the same location over time. While these experiments are purely digital, the problem setup explicitly incorporates a number of practical considerations that an attacker would need to take into account when physically realizing AE. This chapter makes the following novel contributions: (1) as far as the authors are aware, this is first empirical study of AE for satellite imagery, (2) we propose an approach for digitally designing physically realizable AE by explicitly incorporating remote sensing metadata directly into the optimization process, (3) we consider the implications of attacking signals whose fundamental characteristics are changing over time (e.g. as might occur in land use classification problems), and (4) we empirically demonstrate the importance of physical scale to the attack success rate when perturbations are limited in physical extent. In particular, contributions 2-4 also suggest a number of promising directions for future work.

The content in this chapter also appears in [51, 52], which is joint work with Wojciech Czaja, Neil Fendley, Christopher Ratto, and I-Jeng Wang. This author's own contribution to these works consisted of co-developing the experimental design, conducting the numerical experiments, and co-authoring the resulting publications.

3.1 Background

3.1.1 Adversarial Examples

3.1.1.1 Overview

Many modern deep learning systems exhibit a lack of stability to specially-designed "small" perturbations to the signal input space. While what precisely constitutes a "small" perturbation varies by application, it is generally understood to be a modification that leaves the (human-perceived) signal content unchanged while inducing a fundamental change in the output of the targeted machine learning system. Signals containing perturbations designed in this manner are termed *adversarial examples* (AE) [35, 70, 100, 121, 139].

AE (and their potential real-world implications) have been a topic of substantial recent interest. Despite widespread attention, however, there remain many open questions. For example, while analyses into the mathematical properties of AE have been conducted (e.g. [61, 63, 64]), a complete theoretical understanding of this phenomenon remains elusive. On the more applied end of the spectrum, it is unclear how consistently AE can mislead real-world systems. This question of applicability is broad and depends upon many factors such as *a priori* knowledge of the targeted system, how and where perturbations can be injected in the signal processing chain, what constraints are placed upon the perturbation, and the required level of robustness to variabilities in signal acquisition (variations in viewpoint, lighting, signal preprocessing, etc).

A number of studies of AE in physical settings have recently been performed, including [11, 60, 99]. These experiments involve earth-based sensing in the visible spectrum at relatively close ranges, e.g. within the sensing range of the camera on an autonomous vehicle or a facial recognition system. There continues to be some debate regarding the practicality of fielding such attacks in real-world settings. For example, in [112] the authors concluded that AE are not a concern for autonomous vehicles since many AE generated from a single anticipated perspective did not preserve their adversarial properties when perceived from other viewpoints. However, [11] subsequently demonstrated that, by explicitly accounting for the anticipated distribution of viewpoints (and other variations in the sensing process), it is indeed possible to construct robust AE and therefore the phenomenon merits consideration. Explicit defense against adversarial attacks has also been considered in the literature with varying degrees of success; for a few examples see [10, 12, 35] and references therein.

Developing physically-realizable AE for *remote sensing* (RS) modalities (e.g. satellite imagery, multi-spectral data, LIDAR, and SAR) have not yet been well-studied, despite the fact that deep learning is being considered for a variety of remote sensing applications (e.g. see [150]). Since the digital manifestation of these signals is typically image-like, much of the existing work in AE is applicable. However, fundamental differences between RS and ground-based sensing introduce important, unique considerations when one considers physical realizations of AE. For example, the relatively large physical scales involved in RS suggests that an attacker may be severely limited in terms of how the underlying signal can be perturbed. And

of course it is not simply that images cover a large spatial extent, but that signal processing algorithms exploit this context (see [116] for one example and a discussion of the importance of context). Temporal scale and material properties also provide their own unique challenges in this domain.

AE are designed to mislead a signal processing algorithm (typically a deep learning-based algorithm) and in the remainder we will refer to this as an "attack" on the signal processing system. However, whether the AE itself is being employed in an offensive or defensive capacity in the broader context depends upon one's perspective. From the viewpoint of the entity who is fielding the RS system, an AE can be seen as an attack while, from the perspective of the observed entity, realizing an AE might be viewed as a defensive mechanism. In the remainder we simply refer to an AE as an attack on a signal processing algorithm and remain agnostic regarding the relative roles in any broader context.

3.1.1.2 Techniques

In this section we briefly review techniques for generating adversarial examples. The space of AE techniques is rapidly evolving; here we focus on a few of the more canonical approaches. Additional details related to AE are available in [64] and the associated references. In the remainder we will consider attacks against classification problems for image-like data, however it should be noted that the notion of AE generalizes to other settings as well.

Let $f : \mathbb{R}^d \to \{0, \ldots, k-1\}$ denote a classifier which maps d-dimensional

images to a discrete label set of cardinality k. For a given input $x \in [0,1]^d$ (an image with pixels scaled to be between 0 and 1) and a target label $\ell \in \{0, \ldots, k-1\}$ let $r \in \mathbb{R}^d$ denote a perturbation designed to cause the classifier to predict label ℓ; i.e. $f(x+r) = \ell$. We then refer to r as an *adversarial perturbation* and the resulting signal $x + r$ as an *adversarial example*. An underlying assumption is that $f(x) \neq \ell$ since it makes little sense to speak of a perturbation whose intent is to leave a classifier's original prediction unchanged. The goal to induce a particular prediction $f(x+r) = \ell$ is termed a *targeted attack*. An alternative is a *non-targeted* attack where the adversary's goal is merely to change the original prediction, i.e. $f(x+r) \neq f(x)$. Note that adversarial examples can also be considered in non-classification tasks; however, this is outside the scope of the present discussion. If all details of the classifier f are known by the adversary, this is often referred to as a *white-box* attack. When details of the classifier are not directly known and must be guessed or estimated via query, this is referred to as a *black-box* attack. Generally speaking, white-box attacks tend to be the most difficult to defend against [10].

The seminal paper on adversarial examples proposed the following optimization problem [139]

$$r^* = \min_{r \in \mathbb{R}^d} ||r||_2 \qquad \text{subject to} \quad f(x+r) = \ell, \quad x + r \in [0,1]^d.$$

The second constraint serves to ensure that the adversarial example $x+r$ has admissible pixel values while the $||\cdot||_2$ penalty on r encourages the perturbation to be unobtrusive (the ℓ_2 norm is used as a surrogate for visual subtlety). The same authors

also considered penalty function methods that utilize the classifier's *loss function* $J(x, y, \theta)$, which is a function that penalizes prediction errors. Typically the loss function is used to train a predictor; in the context of AE, it is used as a mechanism to guide the design of a perturbation. In our notation for J, $x \in \mathbb{R}^d$ denotes a signal, $y \in \mathbb{R}$ the corresponding target/desired prediction, and $\theta \in \mathbb{R}^p$ the parameters associated with a function f which produces predictions $\hat{y} = f(x; \theta)$. The specific form of a loss function J depends upon the problem at hand. For example, in regression problems [81] one might use the a loss of the form $J(x, y, \theta) = (y - f(x, \theta))^2$ or $J(x, y, \theta) = |y - f(x, \theta)|$.

For binary classification problems (including neural networks) a popular choice is the cross-entropy loss

$$J(x, y, \theta) = -(y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})),$$

or, in the K-class setting where $y \in \mathbb{R}^K$ [81],

$$J(x, y, \theta) = -\sum_{i=1}^{K} y_i \log(\hat{y}_i).$$

For more details on loss functions and associated classical results in machine learning see [81]; for details on training modern deep learning algorithms see [71]. For a given J, Szegedy and his collaborators proposed another optimization-based approach to

design the adversarial perturbation r

$$r^* = \min_{r \in \mathbb{R}^d} c\|r\|_2 + J(x + r, \ell, \theta) \qquad \text{subject to} \quad x + r \in [0, 1]^d.$$

Here, c is a scalar coefficient used to trade the relative influence of the two soft constraints and ℓ continues to denote the classification label one wishes to induce via the perturbation r.

Subsequently a number of alternative approaches have been explored in the literature. The *fast gradient sign* (FGS) method [70] is a computationally thrifty heuristic for untargeted attacks subject to an ℓ_∞ constraint on the perturbation. It uses the gradient of the classifier loss function J to take a single step of magnitude ϵ in a direction that (locally) leads to the greatest increase in loss

$$r = \epsilon \operatorname{sign}(\nabla_x J(x, y, \theta)).$$

Iterative variants of FGS were presented in [99]. Highly effective attacks for a variety of p-norm constraints upon the perturbation magnitude were presented in [36].

The aforementioned methods all implicitly assume the adversary can manipulate all pixels in the scene. Another interesting source of constraints arises when an attack can only modify a subset of the image support. In [122] the authors use a greedy iterative method to generate sparse perturbations that modify only a subset of pixels in the input space. In [25] the authors consider adding relatively modest-sized "patches" into natural image scenes in order to defeat whole-image

classification algorithms. Analogously, [38, 60] seek to defeat street sign classification and detection by designing physical perturbations whose supports are explicitly constrained to coincide with that of the object in question (in these cases, restricted to the surface of a stop sign). In these physical settings, the aforementioned authors seek to develop attacks that are effective over a range of possible locations, rotations, and scalings. To achieve robustness to these possible variabilities, the authors leverage the idea of optimizing over a collection of potential transformations, an idea introduced by [11] and termed *expectation over transformation* (EOT). In this setting, the idea is to constrain the effective distance δ between adversarial and original inputs over a distribution of transformation functions T

$$\delta = \mathbb{E}_{t\sim T}\left[d(t(x) - t(x+r))\right], \tag{3.1}$$

where d is some suitable distance metric (e.g. a p-norm) and $t \in T$ denotes a specific transformation. The distribution T can capture a variety of transformations, such as scaling, translation or additive noise. Once a suitable T has been determined, adversarial perturbations are generated by solving an optimization problem. The authors of [11] proposed a problem of the form

$$\arg\min_{r\in\mathbb{R}^d} \mathbb{E}_{t\sim T}[-\log \mathbf{P}(y|t(x+r))] + \lambda\,\delta, \tag{3.2}$$

where $\mathbf{P}(y|x)$ is the classifier's probability estimate for label y given image x and λ is a scalar weighting coefficient.

The EOT framework is quite general and therefore applicable to many settings. Of course, for attacks that are ultimately to be realized in the physical world, the question of designing T so that it is sufficiently representative of the physical setting is non-trivial. In most cases, precisely modeling all possible physical phenomena is unrealistic. Thus, each new domain and sensing modality introduces new challenges and opportunities for analysis as one adapts T, the constraints, and the optimization procedure for the specific problem at hand.

3.1.2 Remote Sensing Considerations

In this section we describe salient aspects of the remote sensing problem which must be accounted for when considering a *physical attack*, i.e. the creation of AE by manipulation of the physical world. While scenarios exist under which digital attacks might be possible (e.g. via cyber intrusion into the sensing pipeline), the physical domain is the most readily accessible.

Let $L^s_\lambda(x_1, x_2) : \mathbb{R}^3 \to \mathbb{R}$ denote the *total at-sensor solar radiance* observed by a sensing system for wavelength λ and spatial coordinates (x_1, x_2). For example, the total at-sensor radiance can be modeled as the sum of three sources: energy reflected from Earth's surface and not scattered by the atmosphere (su), energy reflected from Earth's surface and scattered by the atmosphere (sd), and a spatially-invariant molecular scattering term which varies with wavelength (sp) [133]

$$L^s_\lambda(x_1, x_2) = L^{su}_\lambda(x_1, x_2) + L^{sd}_\lambda(x_1, x_2) + L^{sp}_\lambda. \tag{3.3}$$

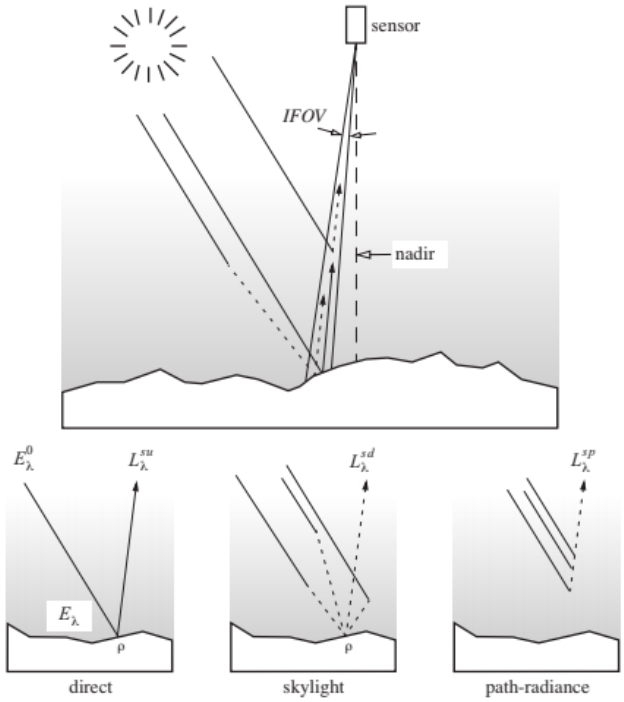


Figure 3.1: At-sensor radiance; Figure 2-3 in [133].

A cartoon representation of these terms appears in Figure 3.1. One natural way to realize a perturbation is by adding or removing materials or changing the surface texture so as to manipulate L^{su}_λ. Of course a RS signal processing algorithm is unlikely to deal directly with raw radiance values, and instead will operate on some digital representation thereof. Figure 3.2 provides one depiction of relevant components in this discretization process. With these processes in mind, we identify a number of ways the RS setting can impact AE design.

1. **Physical Scale**

   The scale of a satellite image is typically orders of magnitude larger than an image taken on the Earth's surface. For example, images taken from the IKONOS satellite can span up to 11.3 km across [56] at sub-meter resolution.

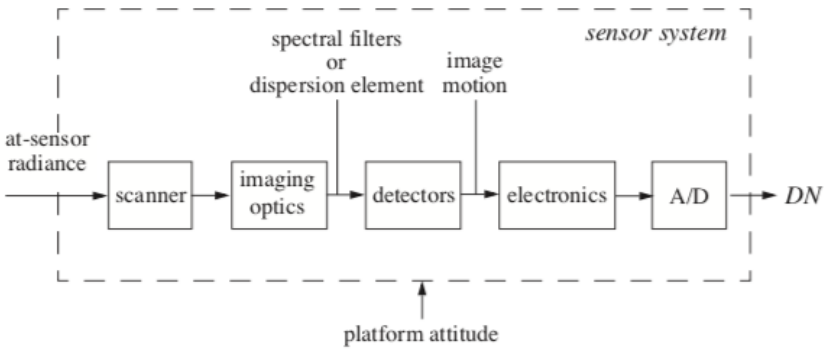


Figure 3.2: Sensing model which maps at-sensor radiance and platform attitude into discrete images (digital numbers (DN)); Figure 3-1 in [133].

The large scale of these scenes imposes a significant challenge upon designing AE with so-called "small" perturbations, especially if they are to be designed in the physical realm. Approaches that subtly perturb entire images are implausible; instead, a practical AE will likely be restricted to perturbing a few regions of modest physical dimension. Constraints on the physical support of AE have been considered already in the context of natural images and local sensing problems (e.g. [25, 60]); however, the potentially vast scale coupled with material constraints adds a subtle but important twist to this consideration.

2. **Viewpoint Geometry**

Remote sensing imagery is influenced by many geometric properties of the sensing process, such as satellite orbit, platform attitude, scanner properties, and earth rotation and shape [133]. While there is some commonality with ground-based sensing modalities, the scale of the RS environment does introduce new considerations. For near-nadir observations from a satellite in

low-Earth orbit, variations in range to various objects in the scene are likely to be modest. Nevertheless, AE robustness to scale is an important issue and we will capture this source of variability in our experiments. While the full three-dimensional geometry of the problem is quite relevant, we will assume for the purposes of this initial study that all objects lie entirely in the ground plane (i.e., the elevation above sea level for each pixel is zero) and that the off-nadir angle is modest. Therefore, any occlusions or shadows imposed by varying elevations of scene elements are ignored. With sufficiently high-resolution metadata providing the satellite ephemeris at the time the imagery was collected, future studies might entertain using an advanced projection/transformation of the images to the ground plane (where each pixel is mapped to a latitude/longitude/elevation).

We observe, however, that the mild off-nadir angle assumption may not be totally egregious since, in many applications, large off-nadir angles may inherently confound image analysis due to the degradation in pixel resolution. Therefore, an adversary may be fundamentally unmotivated to generate attacks for regimes where the targeted algorithm already performs poorly.

3. **Temporal Variability**

Remote sensing data also has a temporal aspect that is distinct from video. An orbiting satellite's ground track will pass through the same spot on the Earth's surface according to a regular schedule (usually several days), enabling it to collect imagery over the same scene. For example, satellite systems such

Figure 3.3: Scene with class label "crop field" exhibits substantial variability over time, due in large part to changes in ground vegetation. Images are from the fMoW data set.

as Sentinel-1 image the entire Earth over a period of days which increases the importance of temporally-aware algorithms [150]. However, the elements of the scene and the environmental conditions surrounding it may change between revisits (e.g. seasonal changes in vegetation, human patterns of life, and weather). For example, Figure 3.3 shows one example of how ground conditions can vary dramatically over time. Algorithms that exploit remote sensing data, including AE, must be robust to changing conditions such as these. In our experiments, we use a data set where the sensing system makes multiple but relatively infrequent passes over a given scene and we measure the effectiveness of an attack across the entire sequence.

4. **Material/Signature Properties**

Material and sensor properties also influence how an adversary may be able to manipulate a scene. For example, in multi/hyperspectral imaging, one may not be able to arbitrarily modify the spectral signature of a given pixel. Instead, material mixture models may determine the admissible set of perturbations that can be realized. Alternatively, in multi-modal settings (e.g.

LIDAR+EO), there are practical constraints upon how a subset of the scene can be modified jointly in each modality. These challenges present opportunities to consider more sophisticated formulations for adversarial attacks. For this foray into AE attacks on remote sensing, we will limit our experiments to the visual spectrum; however, note that that the data set we employ also includes multispectral signatures which could be used in future work.

5. **Atmospheric Effects**

   Unlike machine learning algorithms applied to imagery taken from the surface of the Earth, the performance of algorithms in remote sensing applications is highly susceptible to environmental and atmospheric effects (e.g. illumination, clouds, haze; see Chapter 2 in [133] for more details) and the properties of physically realized AE will be affected by these phenomena as well.

3.2 Methods

3.2.1 Digitally Emulating Physical Attacks

Our goal is to develop adversarial perturbations consisting of opaque material "patches" that, when placed within remotely sensed scenes, degrade the performance of whole image classification algorithms. While our experiments are purely digital, we propose an approach that captures salient aspects of the physical setting where attacks would ultimately be realized.

Inspired by (3.3) we let $\mathcal{P} : \mathbb{R}^3 \to \mathbb{R}$ represent the space of physical signals. An

adversary generates a physical perturbation $r_\lambda^{su}(x_1, x_2)$ and uses a *mixture function* $m : \mathcal{P} \times \mathcal{P} \to \mathcal{P}$ which governs how the perturbation modifies the physical scene

$$L_\lambda^{s'} = m(L_\lambda^{su}, r_\lambda^{su})(x_1, x_2) + L_\lambda^{sd}(x_1, x_2) + L_\lambda^{sp}. \tag{3.4}$$

There are a number of possible models for m; here we will assume that (1) r_λ^{su} is compactly supported on $\Omega \subset \mathbb{R}^2$ and (2) that the physical perturbation completely replaces/obscures the original scene, i.e.

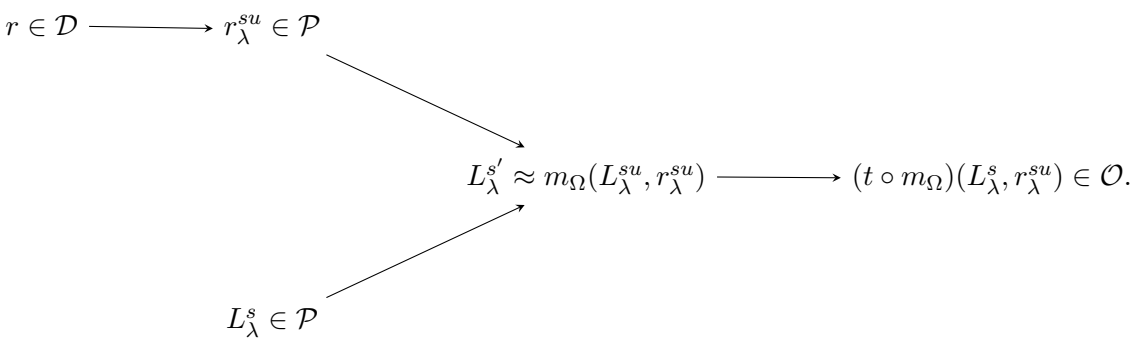
$$m_\Omega(L_\lambda^{su}, r_\lambda^{su})(x_1, x_2) = \begin{cases} r_\lambda^{su}(x_1, x_2) & \text{on } \Omega, \\ L_\lambda^{su}(x_1, x_2) & \text{on } \mathbb{R}^2 \setminus \Omega. \end{cases}$$

One of our first simplifications is to ignore the atmospheric scattering terms for the time being and set

$$L_\lambda^{s'} \approx m_\Omega(L_\lambda^{su}, r_\lambda^{su})(x_1, x_2).$$

We next assume the existence of an *observation function* $t : \mathcal{P} \to \mathcal{O}$, where $\mathcal{O}$ represents the space of digital images, that is the input space of the RS signal processing algorithms. In the remainder we assume RGB images with m_1 rows and m_2 columns, i.e. $\mathcal{O} = \mathbb{R}^{3 \times m_1 \times m_2}$. Note the observation function t is subject to uncertainty since the precise sensing conditions may not be known a-priori; even if they were, the sensing process is sufficiently complicated such that t will typically have some associated modeling error which also contributes to the uncertainty.

Finally, we define $\mathcal{D}$ to be an abstract design space which parameterizes our



$\mathcal{D}$	The space of adversarial designs
$\mathcal{P}$	The space of physical signals
$\mathcal{O}$	The space of digital images
$\Omega \subset \mathbb{R}^2$	The spatial support of the perturbation
m_Ω:	The perturbation "mixing" function $\mathcal{P} \times \mathcal{P} \to \mathcal{P}$
t:	The observation function from $\mathcal{P} \to \mathcal{O}$
r:	The adversarial perturbation design
r_λ^{su}:	Physical realization of the adversarial perturbation
L_λ^s:	The original physical signal; also denoted by x

Figure 3.4: Model for designing physical AE.

adversarial perturbation; keeping with the convention of Section 3.1.1.2 we denote the parameterization of the adversarial perturbation by $r \in \mathcal{D}$. The overall framework is summarized graphically in Figure 3.4. This model is compatible with the spaces assumed in the EOT framework (described in Section 3.1.1.2); however, we take pains to spell it out in order to make clear the approximations being made in our approach.

At this point the setting is fairly general and we now proceed to describe additional assumptions and simplifications we make in order to develop a concrete algorithm. For $\mathcal{D}$, we assume a model whereby perturbations consist of a single opaque, flat, compactly-supported, piecewise-constant perturbation that will be physically placed within the sensed scene. In particular, we assume this perturbation consists

of a square $n \times n$ "checkerboard" of opaque material. We further assume the physical dimensions of each element (i.e. each square in the "checkerboard") are fixed apriori as some number of meteres square. In the current work we focus on the visible spectrum and each design element is associated with a three-channel RGB value that is tuned during optimization. Thus, we take $\mathcal{D} = \mathbb{R}^{3\times n\times n}$. We further assume in our initial approach that the perturbation will be placed in the center of each observed scene; together with the spatial dimensions of the perturbation this defines Ω. The number $(3n^2)$ and the size, in meters, of the perturbation's elements are parameters that are fixed prior to learning the perturbation.

In our experiments, we do not attempt to implement a simulation of t and instead rely on sensed data to design our perturbations. For example, if $x \in \mathcal{P}$ is a source signal to perturb, we will do so by working directly with collected data $t(x) \in \mathcal{O}$. To this end we make a key simplifying assumption that the observed digital signal can be partitioned such that each pixel represents a contribution from either the original signal or the perturbation, i.e.

$$\begin{aligned}(t \circ m_\Omega)(L_\lambda^s, r_\lambda^{su}) &= (t \circ m_\Omega)(0, r_\lambda^{su}) + (t \circ m_\Omega)(L_\lambda^s, 0),\\ &= t(r_\lambda^{su}) + (t \circ m_\Omega)(L_\lambda^s, 0),\\ &=: t(r_\lambda^{su}) + t(x_{\bar{\Omega}}),\end{aligned}$$

where $x_{\bar{\Omega}}$ denotes the subset of the original signal L_λ^s supported on $\mathbb{R}^2 \setminus \Omega$. This is obviously an oversimplification which essentially ignores any possible overlap or edge effects at the interface between the original signal and the perturbation.

Of course, t is subject to uncertainty and, as mentioned in Section 3.1.1.2, the typical approach is to average over a distribution of transformation functions T. A complementary notion, which we explore here, is to use metadata associated with multiple observations $t_1(x), t_2(x), \ldots, t_m(x)$ and approximate T by means of these explicit samples from the true underlying distribution. For example, ground sample distances provide explicit guidance in terms of how a perturbation r (designed in physical coordinates) must be scaled so that $t(r^{su}_{\lambda})$ is dimensionally consistent with $t(L^{su}_{\lambda})$. The quality of this approximation will depend upon complexity of the true sensing process and how much of that complexity the metadata permits one to incorporate into t.

Another challenge with designing physical attacks in this setting is that, due to the variations described in Section 3.1.2, the signal x being attacked may itself change over time (and hence, from one observation to the next). Thus, our approach must accomodate variations both due to the sensing process and the evolution of the targeted signals $\{x_1, \ldots, x_m\} \in \mathcal{P}$. We are further assuming that the attacker is not adapting the perturbation r over time but rather must design a single attack that is effective across all the variabilities manifested in $\{t_1(x_1), \ldots, t_m(x_m)\}$. This is subtly different from most prior work in physical attacks where the dominant source of variability is associated exclusively with the sensing process.

With this in mind we propose the following optimization for digitally designing

targeted attacks against a sequence of observations

$$r^* = \arg\min_{r \in \mathcal{D}} \sum_{i=1}^{m} J\left(m(t_i(x_i), t_i(r_\lambda^{su})), \ell, \theta\right) + \lambda \sum_{i=1}^{m} d(t_i, x_i, r), \tag{3.5}$$

$$\text{subject to} \quad m(t_i(x_i), t_i(r_\lambda^{su})) \in [0,1]^d, \quad i = 1, \ldots, m.$$

Here m is the number of observations available for designing the perturbation, x_i are the physical scenes to perturb, ℓ is the classification label the perturbation desires to elicit, d is a penalty function which encourages visual subtlety, and J, x, θ are properties of the classifier defined in Section 3.1.1.2. Note that the observations $\{t_i(x_i)\}_{i=1}^m$ used to design the perturbation need not consist of all samples present in the data set. For example, one may be interested in exploring how well attacks designed using a subset of samples generalize.

The penalty term d in (3.5) is another important design consideration. In settings where visual subtlety is desired, one could follow the conventional approach of realizing d using a suitable p-norm. In this case, this would be asking for a perturbation that is generally subtle across a range of observational conditions and variations in $\{x_i\}$. The feasibility of visual subtlety is therefore somewhat signal dependent. However, ideally attacks would also be subtle with respect to the 3D geometry of the scene. For example, if the spatial extent of the patch extends beyond the roof of a building or spans regions that are partially obscured/in shadow, it would be preferable that the perturbation r respected these discontinuities. In the absence of a full 3D model of $\mathcal{P}$ another approach would be to segment the scenes $t(x)$ and incorporate this structure into the attack model. In our current work we

loosely approximate this by adding a second term to d which encourages the attack to respect edge-like structure within an image. Let

$$\delta_1 = t(x) - m_\Omega(t(x), t(r_\lambda^{su})),$$

$$\delta_2 = t(x) - m_E(t(x), t(r_\lambda^{su}),$$

where E is the subset of the scene with strong edge-like structure. Then we implement d via

$$d(t, x, r) = \lambda_1 ||\delta_1||_\infty + \lambda_2 ||\delta_2||_\infty. \tag{3.6}$$

Obviously there is ample opportunity for future studies that incorporate more realistic models. Of course, complexity should not be added arbitrarily to these terms since one must also consider impact to the feasibility of solving the resulting optimization problem (e.g. gradient based methods will require reasonably well-behaved functions $\nabla_r t, \nabla_r d$).

3.2.2 Data Set

We base our numerical experiments on the Functional Map of the World (fMoW) data set, a large collection of temporal sequences of satellite imagery designed for evaluating whole image classification problems related to the functional purpose of land use and buildings [42]. The overall fMoW data set contains over a million images from 200 countries; our experiments utilize a subset of the validation

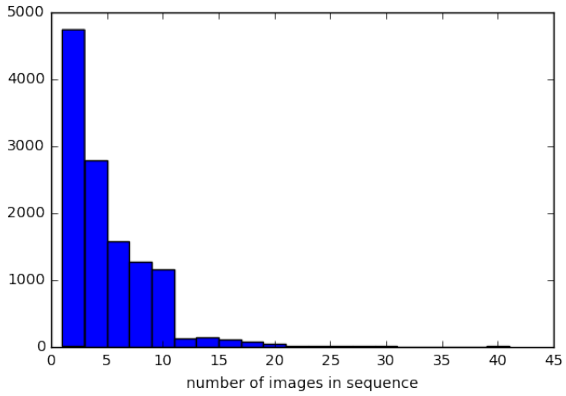


Figure 3.5: Number of (RGB) images per sequence in the fMoW validation split (median=3, maximum=41).

split (the overall validation split contains approximately 53000 images from 12000 unique scenes). Each image contains at least one bounding box labeled with one of 63 possible classes. Furthermore, the data set includes metadata features pertaining to location, time, sun geometry, cloud cover, and physical dimensions of the imaged swath. The fMoW data set has multiple modalities: 4-band or 8-band multispectral imagery as well as high and low resolution RGB imagery. For our experiments, we use the high resolution RGB images but observe that the multispectral domain provides a compelling setting for future study.

We further downsampled the validation subset of fMoW so the resulting sequences satisfy a number of desiderata. First, since our interest is in attacking sequences of nontrivial length, we only consider those with at least 8 views of the same scene. This eliminates a large number of sequences as many in the validation set are quite short (see Figure 3.5). We also only consider images that are correctly classified by the targeted classifier prior to applying any perturbation. Finally, we also limit our study to relatively benign sensing conditions (recall the previous

"Viewpoint Geometry" discussion). We only use images with: mild off-nadir angle (less than 30 degrees), at most modest cloud cover (less than 20 percent of image chip obscured), and sun elevation angles of at least 60 degrees to eliminate more extreme variations in illumination. The resulting experiment includes 66 sequences each having at least 8 admissible frames.

The authors of fMoW also developed and analyzed a number of classification models, some of which use exclusively image data while others make use of metadata and/or exploit the sequential nature of the images by using recurrent networks. Our study considers the one designated "CNN-I", which is a fine-tuned variant of DenseNet with no recurrent structure designed for RGB image data. Our choice of CNN-I is justified in this case since the performance of this network is quite close to that of the recurrent alternatives (see Table 1 in [42] for more details). However, extending the analysis to other networks (e.g. those using metadata or that more explicitly incorporate temporal properties) is another interesting direction for future work.

The fMoW authors also make available their preprocessing algorithm which extracts and resizes bounding boxes into a tensors of dimension 229x229x3 suitable for ingestion by CNN-I/DenseNet. This rescaling has implications for the associated metatdata. In our study we adopt the fMoW preprocessing strategy so that we can use CNN-I off-the-shelf; however, as part of the preprocessing we must also update the metadata so that salient values (e.g. ground sample distance) are still representative following the rescaling.

There remain a few challenges associated with this data set, however. One

substantial issue is that the images for a given scene are not precisely registered. Therefore, while we can use metadata to properly control the scale of a patch, there is no guarantee that the attack will be translated to the precise same location from one observation to the next. Thus, our attack will be subjected to some positional uncertainty.

3.2.3 Software Implementation

Algorithm 2 Procedure used to generate results in Table 3.1.

```
1: procedure GENERATEAE(fMoW, seq_to_attack, θ_CNNI, λ1, λ2, n)
2:     for seq in sequences_to_attack do
3:         χ_e = detectEdges(seq)
4:         for ℓ in {"crop field", "park", "office building", "hospital" } do
5:             r*= minimize (3.5) via SGD
6:         end for
7:     end for
8: end procedure
```

For our experiments, we selected 4 target classes from the fMoW taxonomy: "crop field", "hospital", "office building", and "park". We do not try every possible targeted attack in order to control the computational expense of our experiments; these four classes were selected so as to have some representation of both urban and rural scenery. This selection was made prior to algorithm evaluation and is therefore not cherry-picked for performance.

We implemented our experiments in Python using the TensorFlow [3] machine learning library. We use the typical cross-entropy loss for J and stochastic gradient descent to minimize (3.5). The function t uses metadata available in fMoW (in particular, the ground sample distance) to scale the attack patch appropriately for

	Attack Parameters			Metrics	
experiment	n	m/element	num. frames attacked	attack success rate (%)	total error rate (%)
1	60	0.5	1	11.9	38.1
2	80	0.5	1	15.5	43.2
3	100	0.5	1	19.2	50.8
4	60	0.5	4	31.3	51.7
5	80	0.5	4	44.5	61.5
6	100	0.5	4	53.1	67.3

Table 3.1: Success rates for targeted white-box attacks against the fMoW classifier "CNN-I" for six experiments (id 1-6). Parameters include the number of elements n in each dimension as well as the size, in meters, of each element (see Section 3.1.1.2). "attack success rate" indicates the targeted AE success rate (i.e. $f(x) = \ell$) while "total error rate" indicates how frequently AE caused the classifier to make *any* mistake (i.e. $f(x + r) \neq f(x)$).

each image. As described previously, the penalty function d has two terms: one which encourages a small ℓ_2 distance between the attack and the underlying images while the second term encourages a small ℓ_2 distance between the attack and strong edge-like structure in the scene. These two terms have an associated weighting coefficients of $\lambda_1 = 1\text{e}{-3}, \lambda_2 = 1\text{e}{-1}$. We used the Canny edge detector provided as part of skimage to determine χ_e (using a Gaussian width parameter of 2.0) for all images. Calculations were performed on a GeForce 1080 Ti trained using gradient descent for 1000 epochs at two different learning rates (100 and 20) which took approximately 10 minutes per attack (although our code did not optimize CPU to GPU data transfers).

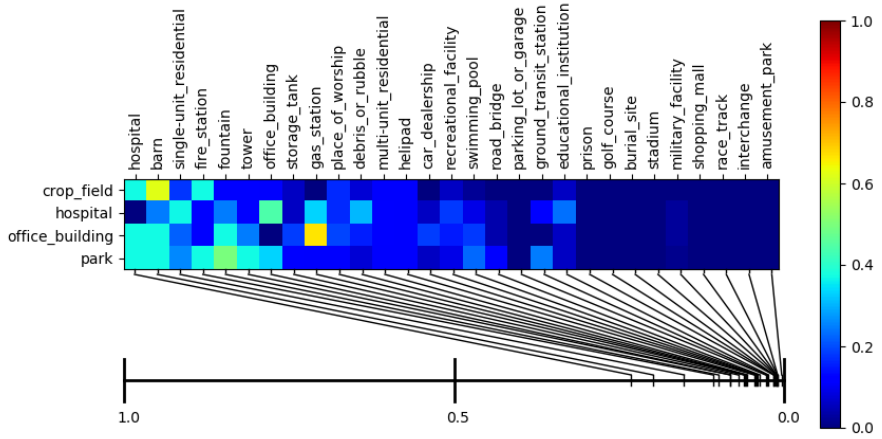


Figure 3.6: Targeted attack success rates for experiment 1. The horizontal axis below the image shows, for each attacked class, the median percentage of the image the physical attack covered. The intensity of the colormap indicates the overall attack success rate.

3.3 Results

Overall attack success rates for six different experiments (covering three different physical attack configurations) are shown in Table 3.1. The table shows the number of parameters used in each dimension of the attack (n) as well as the size of each element within the patch (m/elt.). The first three rows (experiments 1-3) provide a baseline result for when an attack is based solely on the first image in a sequence. Numbers reported are misclassification rates post-attack (note that all images used in this study were correctly classified by CNN-I pre-attack). Experiments 4-6 show how the overall attack rate improves if the attacker is privy to the first four images in each sequence.

Figures 3.6 to 3.11 show targeted attack success rates decomposed by sequence class label and arranged by relative size of scenes. The color of cell $m_{i,j}$ in each heat

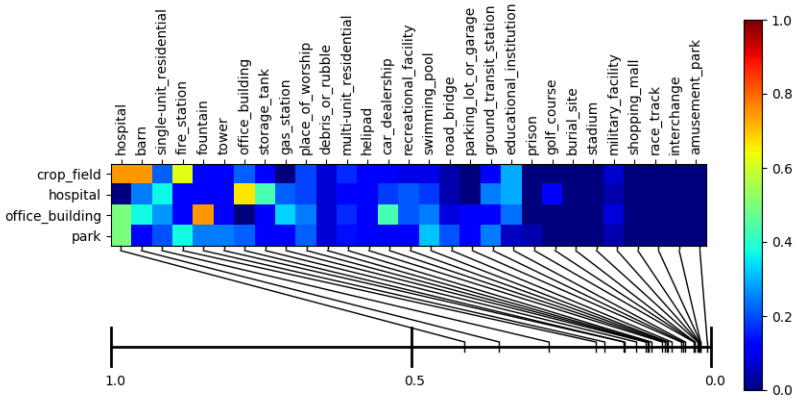


Figure 3.7: Targeted attack success rates for experiment 2.

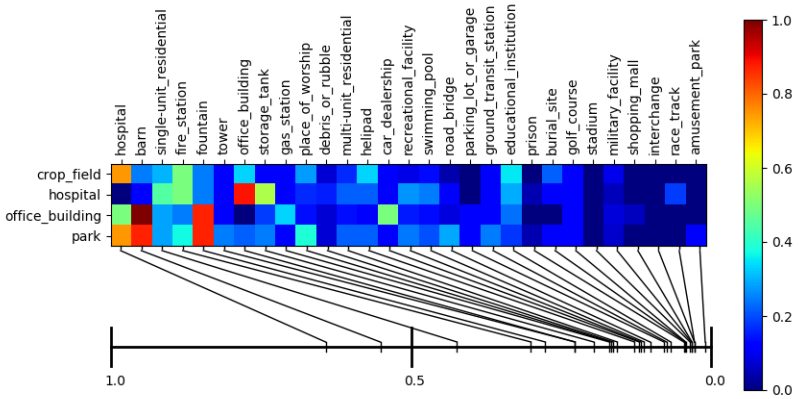


Figure 3.8: Targeted attack success rates for experiment 3.

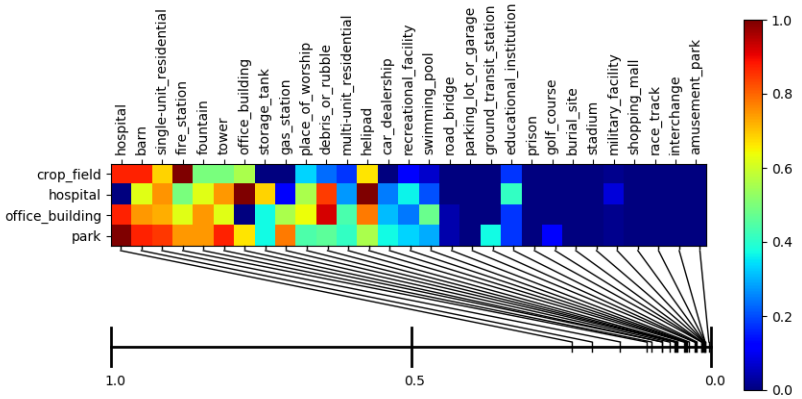


Figure 3.9: Targeted attack success rates for experiment 4.

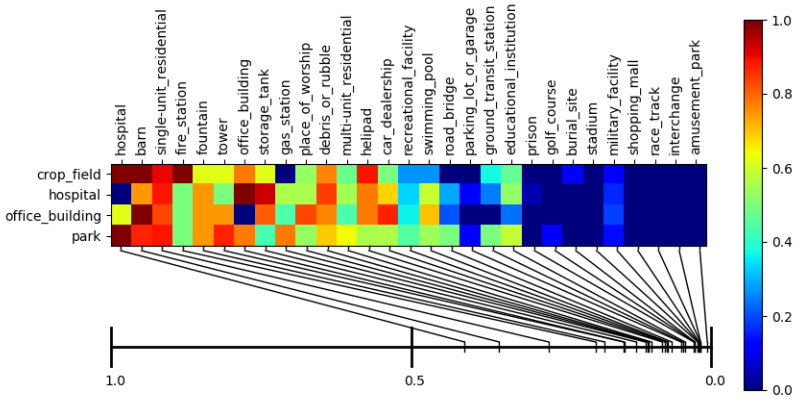


Figure 3.10: Targeted attack success rates for experiment 5.

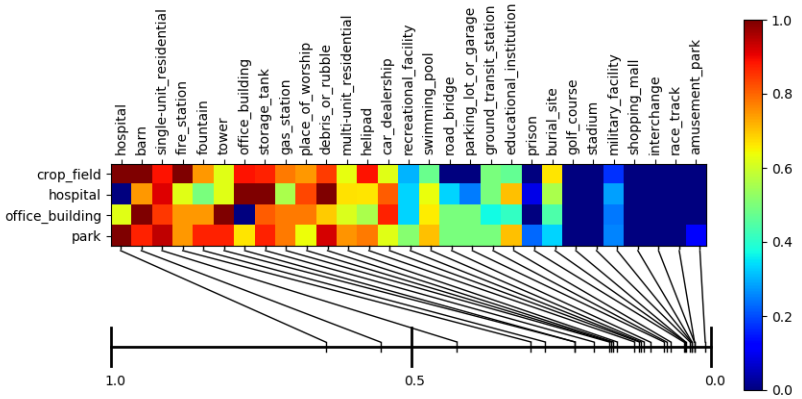


Figure 3.11: Targeted attack success rates for experiment 6.

map indicates the success rate of attacking sequences whose original class is j in scenarios where the target class is i.

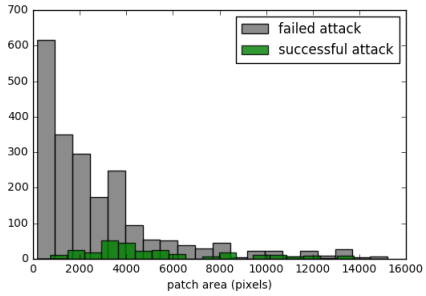
Sequences are ordered by size and the horizontal axis denotes the percentage of the overall scene covered by the fixed-size AE (1.0 indicates the attacker can perturb all pixels in the scene). Since all images are rescaled to a fixed size by the fMoW preprocessing and our attacks are designed in physical coordinates, larger scenes result in fewer available pixels for the attacker to manipulate. A relation-

(a) original image (b) adversarial example

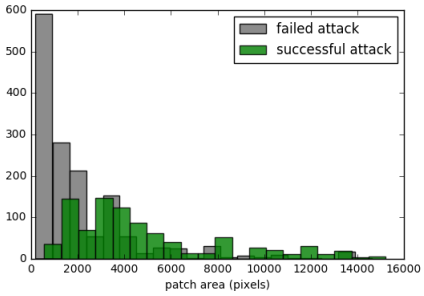
Figure 3.12: Targeted attack causing the classifier to label a "place of worship" as a "hospital".

ship between relative attack size and success rate is evident. The hypothesis that limiting the number of pixels available for an attack may reduce the success rate is further supported by Figure 3.13 where we overlay distributions for successful and unsuccessful attacks as a function of number of pixels manipulated. This suggests that more sophisticated attacks (e.g. multiple, spatially distributed AE) may be necessary in order to successfully attack larger scenes.

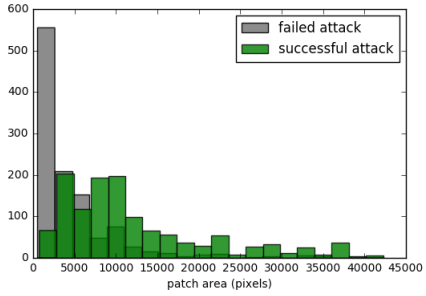
Figure 3.12 provides a visual example of how matching strong edges and shadows is encouraged by our choice of penalty term d (recall (3.5)); also clear, is that this soft constraint does not provide perfect agreement with the 3D geometry of the scene. Note that, if additional metadata were available, the constraints on the spatial support of the attack could be made more realistic. Improving upon these details is a ripe direction for future work.



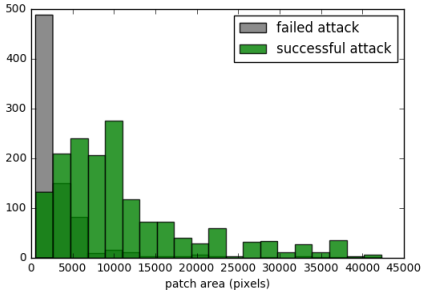
(a) experiment 1, targeted attacks



(b) experiment 1, non-targeted attacks



(c) experiment 6, targeted attacks



(d) experiment 6, non-targeted attacks

Figure 3.13: Distributions of successful and unsuccessful attacks as a function of number of pixels the attack manipulated. Attack success rates increase as the attack is able to manipulate more pixels in the scene (here, corresponds to decreasing ground sample distances).

3.4 Conclusions

This chapter has presented new approaches and experiments for developing adversarial examples in remote sensing applications; in particular, a study of attacks on satellite imagery which takes into account practical physical considerations. We describe an approach for simulating physical attacks in the digital space which is unique in our use of metadata to align the attack with remotely sensed data. These experiments also highlight the importance of physical scale in the AE design process.

This work only begins to scratch the surface of what is possible; additional experimentation (e.g. more realistic simulations of the sensing process, more extensive use of metadata, experimentation on larger sets of data, etc.) are all near-term next steps which would further enrich the results presented here. Additionally, there is abundant room to explore the broader space of attack designs, which could include implementing multiple patch-like attacks (and the corresponding patch location design problem), more directly accounting for 2D and 3D structure of scenes (e.g. by combining attacks with segmentation results), incorporating different notions of visual subtlety, and also exploring the impact of different levels of knowledge about the targeted system on the part of the adversary. More ambitiously, there are interesting questions regarding extensions to other modalities (multi and hyperspectral data, SAR, LIDAR, etc.) and also to settings where multiple sensors are utilized simultaneously. These directions also offer new opportunities for defining relevant constraints, such as exploiting spectral signature databases and material mixture models to ensure physically realizable perturbations in multi/hyperspectral settings.

Attacks to other signal processing algorithms in the remote sensing domain, such as object detection and change detection, also offer interesting opportunities. More generally, as the field of adversarial examples continues to mature, new findings and discoveries may play a role in this setting as well. As mathematical techniques for developing classifiers with provable desirable properties evolve (e.g. [26]) their findings may help inform future experiments of robustness in the remote sensing setting. Finally, there is an obvious need to validate digital experiments physically.

Chapter 4: Biomedical Applications

The previous chapters focused on the importance of stable features for machine learning applications. In this chapter we shift gears somewhat and consider modern applications of deep learning for diagnosing eye-related disorders. Our application involves two medical image processing tasks: a fine-grained segmentation problem and an image classification problem. In the first task, our primary research objective is to evaluate the efficacy of deep learning methods for the automatic fine-grained segmentation of optical coherence tomography (OCT) images of the retina. We compare the performance of our algorithm to that of human annotators and to existing (non-deep learning) algorithms of record. In the second task our primary research objective is to explore methods that combine image-based features with demographic information for detecting age-related macular degeneration (AMD). In both applications, there is an interesting aspect in that the data and/or the desired annotations possess some specialized structure. For the segmentation task there is a smoothness prior associated with the desired estimates while in the latter case the demographic data is of a fundamentally different nature than the tensor-like image data. In both cases we propose methods for addressing these structural properties that involve adding algorithmic components to a deep learning-based feature extrac-

tor. A future goal is to adapt and more directly incorporate the notions of stability and invariance from Chapter 2 for the non-image aspects of these problems. For example, while we presume the Haar-like wavelets of Chapter 2 are not ideally suited for photographic images of the eye, it is possible that multidimensional extensions of these wavelets might provide some interesting capabilities for non-image data. As of this writing, these extensions remain speculative but interesting directions for future work.

Much of the content from this chapter represents extended versions of [125] and [86], which is joint work with Neil Bressler, Philippe Burlina, Delia Cabrera DeBuc, David Freund, Arnaldo Horta, Neil Joshi, Jun Kong, and Katia D Pacheco. With respect to [124] this author was primarily responsible for the algorithm design, developing the numerical experiments (with contributions from N. Joshi on DenseNet), and co-authoring the associated paper. For [86], this author contributed to the design of the numerical experiments, analyzing the results, and co-authoring the paper.

4.1 Background

4.1.1 AMD Overview

Age-related macular degeneration (AMD) is a retinal condition induced by the degeneration of the central area of the retina known as the macula. AMD (together with glaucoma and diabetic retinopathy) is one of the leading causes of blindness and visual impairment, especially among individuals 50 years or older in

the United States [5, 6, 17, 23, 95]. Estimates (circa 2004) put the number people in the United States with some form of advanced state AMD between 1.75 and 3 million [23]. AMD severity is frequently graded using the four category Age Related Eye Diseases Study (AREDS) classification scale:

1. No AMD present.

2. Early stage AMD.

3. Intermediate stage AMD, characterized by large-sized or extensive medium-sized *drusen*, i.e. accumulations of acellular debris present the basement membrane of the retinal pigment epithelium (RPE) and Bruchs membrane.

4. Advanced AMD, characterized by damage to the macula via the wet form (characterized by choroidal neovascularization (CNV) caused by the production of vascular endothelial growth factor (VEGF)) or the dry form (characterized by geographic atrophy (GA) of the RPE, affecting the center of the macula).

Early detection of AMD is desirable as daily intake of certain high-dose vitamins may slow the progression of intermediate stage AMD (or late stage AMD in a single eye) [118]. While examination of the retina by an ophthalmologist is the most effective method of identifying AMD, manually grading fundus images (i.e. images of the back of the eye) is a time-consuming and expensive process. This has motivated a number of researchers to explore how automated methods might be used to expedite retinal image analysis.

While color fundus images are frequently used to diagnose optical disorders, other modalities have proven useful as well. Ultrasound techniques have played a significant role in ophthalmology; two types of devices, known as A-scan and B-scan ultrasound, have been used for diagnostic purposes since the 1950s. Both techniques are useful for identifying certain classes of lesions; B-scans (or "brightness" mode) provides high resolution two-dimensional images that can be used to better differentiate lesions [127]. Optical coherence tomography (OCT) is another complementary technique which uses scattering of near-infrared energy to produce two- and three-dimensional images. OCT imaging is a non-invasive, high-resolution technique capable of capturing micron-scale structure within the human retina. The retina is organized into layers (see Figure 4.1) and abnormalities in this layered structure have been associated with a number of ophthalmic, neurodegenerative, and vascular disorders. For example, studies have shown that advanced AMD lesions correlate with thinning of the outer retina in geographic atrophy as well as underlying choroidal neovascularization [89].

As a part of the central nervous system (CNS), the retina is also subject to a number of specialized immune responses similar to those in the brain and spinal cord; changes in the retinal structure have been associated with CNS disorders such as stroke, multiple sclerosis, Parkinson's disease, and Alzheimer's disease [110]. In particular, thinning of the retinal nerve fiber layer (RNFL) is associated with the aforementioned neurologic disorders and, in some cases, its thickness correlates directly with the progression of neurological impairment. Furthermore, ocular manifestations of CNS disorders can sometimes precede symptoms within the brain

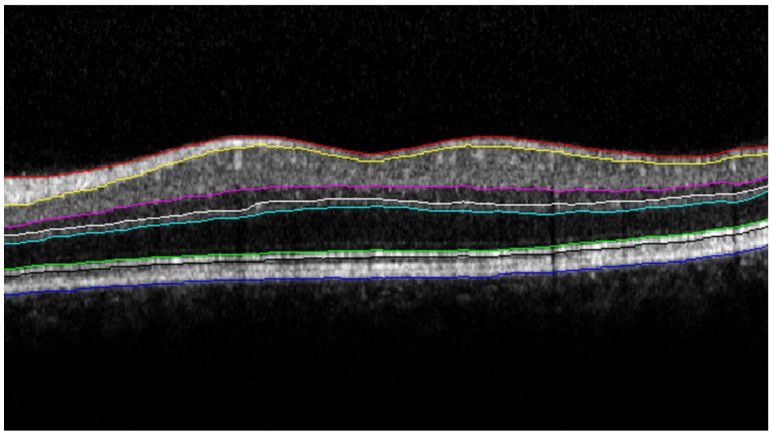


Figure 4.1: OCT B-scan from Spectralis SD-OCT showing the layered retinal structure; figure reproduced with permission from [141]. Note that eight intraretinal layer boundaries are delineated with red, yellow, magenta, white, cyan, green, black and blue solid lines, respectively. The notations are summarized as follows: Red: internal limiting membrane (ILM), yellow: outer boundary of the retinal fiber layer (RNFLo), magenta: inner plexiform layer-inner nuclear layer (IPL-INL), white: inner nuclear layer-outer plexiform layer (INL-OPL), cyan: outer boundary of the outer plexiform layer (OPLo), green: inner segment-outer segment (IS-OS), black: outer segment-retinal pigment epithelium (OS-RPE), and blue: retinal pigment epithelium-choroid (RPE-CH).

itself. Since the retinal structure can be imaged relatively easily via OCT, automated retinal analysis using OCT provides a compelling complement to traditional CNS detection methodologies. Currently, commercial OCT devices provide a map to describe the retinal thickness, typically between the surface of the retina and the retinal pigment epithelial layer of the retina. However, these measurements do not by themselves extract all of the useful information relating to retinal pathology, motivating the use of signal processing techniques.

4.1.2 Prior Work

Work in automated retinal image analysis (ARIA) has steadily progressed in the past two decades as datasets became more plentiful and machine vision and ma-

chine learning techniques have improved (e.g. [27, 30, 65, 66, 85, 143, 144]). A substantial part of this research has been directed towards diabetic retinopathy, another commonly studied retinal disease primarily affecting those with diabetes [45, 75]. Automated detection of AMD remains comparatively less well studied. Most attempts to automate the detection and classification of AMD exploit fundus images, e.g. [30, 85]; however, there has also been recent work utilizing optical coherence tomography (OCT) [144]. Early work in fundus image analysis sought to detect drusen directly using one-class techniques such as support vector data descriptions (SVDD) [30, 66]. The recent dramatic success of deep learning has inspired a number of ARIA applications. A number of authors pursued techniques to automatically detect patients with referable age related macular degeneration from fundus images [27, 28, 29] or OCT [105]. Studies also demonstrate how deep learning techniques can achieve significantly better performance relative to classical techniques [4]. Another recent work uses a generalization of the backpropagation method to generate heatmaps that provide insight into which subset of the image is most responsible for the deep learning algorithms' classification result [126].

Great strides have also been made recently in automatic OCT image segmentation. While initial approaches for segmenting OCT images typically utilized graph-based methods (e.g. [20, 54, 58, 69, 92, 102, 107, 140, 141]) there has been recent interest in applying machine learning techniques as well. For example, in one recent study a 7 layer OCT segmentation using kernel regression-based classification was developed to estimate diabetic macular edema (DME) as well as OCT layer boundaries. These estimates were combined with a graph-based segmentation

algorithm; the overall approach was then validated on 110 B-scans and ten patients with severe DME pathology, yielding DICE coefficient of 0.78 [41]. For extensive reviews of recent state of practice for OCT segmentation see also [20, 140].

Deep convolutional neural networks (CNNs) have also been applied recently for OCT segmentation. In particular, so-called semantic segmentation algorithms, which solve per-pixel classification problems (as opposed to whole image classification), are a natural fit for this setting. A popular approach for implementing semantic segmentation is the U-Net architecture of [130]. One recent study [106] uses U-Nets to delineate macular edema and obtains an F1 score of 0.91, effectively providing performance on par with human annotators. Other deep learning-related studies include [82], which uses U-Nets to demonstrate performance close to that of a classical approach based on random forests, and [62] who uses a hybrid of CNN and graph-based method to identify OCT boundary layers.

Of course, data-driven algorithms rely heavily upon access to representative data sets. Recent efforts at the University of Miami [141] have taken steps to develop publicly available OCT datasets with clinical gold standards for comparing performance among methods, including a number of OCT segmentation algorithms of record. In addition to OCT images and ground truth, the publicly available University of Miami OCT dataset [141] also includes annotations generated by five commonly used OCT segmentation software packages and/or algorithms of record. These reference algorithms/implementations are: Spectralis 6.0 [69], IOWA Reference Algorithm [107], AUtomated retinal analysis tools (AURA) [102], Dufour's (Bern) algorithm [58], and OCTRIMA3D [140]. A complete description of these

algorithms is available in [141].

As indicated above, deep learning is well suited for working with medical images that admit a natural tensor structure. However, architectures that explicitly incorporate less structured information, such as patient demographics, are comparatively less well studied. Typically this auxiliary information has much lower dimensionality relative to the image-like inputs and is often not spatial in nature. In the following we refer to such data as "side channel" inputs; while the term may suggest this data is somehow of secondary importance relative to the image data, this is not necessarily the case. For example, in robotic control problems, side channel information representing current state and/or vehicle goals are critical to producing successful outcomes [149]. Combinations of image-like and side channel information have recently been utilized for fundus image analysis as well. In [67] the authors combine image meta-data (pertaining to the original aspect ratio and field of view prior to pre-processing) with image-based features produced by a deep neural network. Together, these image-like and side channel features comprised the inputs to a decision tree classifier. In this chapter, we will consider similar architectures; however, the side channel information we employ characterizes properties of the patient as opposed to that of the image preprocessing mechanism.

4.2 Application: OCT Image Segmentation

4.2.1 Objective

As described in Section 4.1.1 there is great potential clinical value in being able to accurately estimate the fine-grained structure of retinal layers. Our goal is to explore whether modern deep learning-based semantic segmentation techniques, when coupled with suitable post-processing, can provide an effective mechanism for estimating boundaries between retinal layers. We use an open-source data set for which rich baseline results exist; in particular, we have access to both estimates from "classical" methods (i.e. not based on deep learning) as well as human annotations which admit a characterization of inter-operator error. As far as we are aware, our work is the first attempt to apply deep learning to this particular data set. In addition, we explore whether regression methods can provide added value as a the post-processing step. In particular, we consider regression-based methods that permit one to bring some prior knowledge regarding the smoothness of the retinal surfaces. If successful, this would enable future research efforts whereby clinical priors can more naturally be incorporated into the estimation procedure. A longer-term objective would be to combine these models with mathematical properties of robust network-based features.

4.2.2 Approach

Our approach for estimating retinal surfaces consists of two primary steps. The first solves a per-pixel (or "dense") classification problem of associating each pixel in the image with the most likely corresponding retinal layer. These per-pixel estimates are then post-processed to extract the retinal surfaces (i.e. boundaries between regions). For this post-processing step, we explored two methods: a hand-crafted heuristic and a regression procedure which models retinal surfaces as smooth functions. Note that, while our current experiments involve two-dimensional images, both steps above extend naturally to three dimensions. Thus, our approach is also applicable to settings where labeled volumetric data is available. The overall process is summarized in Algorithm 3; we describe each algorithmic component in further detail below.

Algorithm 3 OCT Segmentation Pipeline

```
1: procedure SEGMENT(bScans, annotations)
2:     for i = 1 . . . nPatients do
3:         data = train_test_split(bScans, annotations, i);
4:         ▷ Semantic Segmentation (Section 4.2.2.1)
5:         cnnModel = densenet_train(data.xTrain, data.yTrain);
6:         yHatRaw = densenet_predict(cnnModel, data.xTest);
7:         ▷ Post-processing: Approach 1 (Section 4.2.2.2)
8:         yHatSeg[i] = heuristic_repair(yHatRaw);
9:         ▷ Post-processing: Approach 2 (Section 4.2.2.2)
10:        yHatRawTrain = densenet_predict(cnnModel, data.xTrain);
11:        gpHypers = select_GP_hypers(yHatRawTrain, data.yTrain);
12:        yHatSegReg[i] = GP_regression(gpHypers, yHatRaw);
13:    end for
14:    return yHatSeg, yHatSegReg
15: end procedure
```

4.2.2.1 Semantic Segmentation

For the first step, we use a CNN to generate the per-pixel layer estimates. Initial studies that generated per-pixel estimates from CNNs utilized a "sliding window" approach whereby the algorithm is presented with a series of overlapping regions extracted from the main image and asked to generate classification labels for the central pixel in each extracted scene (e.g. [44]). However, modern fully convolutional neural networks (FCNs) provide computationally efficient alternatives to sliding window approaches for semantic segmentation problems [111]. FCNs are a subcategory of CNN that take tensor-like data as input and produce class estimates having the same spatial dimensions (i.e. per-pixel or per-voxel labels). Processing the data in this manner avoids unnecessary redundancy in the calculations, which can provide orders of magnitude improvement in runtime. Therefore, we elected to use a FCN architecture for this work; in particular, we adopt the DenseNet architecture [88, 91]. DenseNets are characterized by an extensive use of "skip connections" which permit each layer of the network to directly process the outputs from all previous layers (see Figure 4.4). This construction makes richer sets of features available at each layer of the network while also providing a mechanism to alleviate the vanishing gradient problem which can when training the network via back-propagation. This is in contrast to earlier network architectures where each layer operates solely upon the output of the previous layer. Other FCN architectures, such as U-Nets [130], also include skip connections; however these intra-layer connections are less abundant relative to the DenseNet architecture. While we initially

experimented with U-Nets in this study, we found that the DenseNet architecture ultimately provided superior classification performance (perhaps, in part, as a result of being easier to finetune).

For our experiments we adopted the 103 layer DenseNet-FCN architecture described in [91]; in particular, we used the publicly-available Keras implementation of [128]. We adopted directly the architecture and initial weights of this network; our customizations consisted of adjustments to the loss function and the synthetic data augmentation methodology described below. During training, we minimized the pixel-wise cross entropy loss using the Adam [94] optimizer, with a learning rate of 1e−3. Due to memory constraints we did not load each image into memory at once. Instead, examples consisted of vertical slice of 256×512 pixels that were randomly cropped from the original image (whose dimensions were 496×768; see Section 4.2.3 for full details). These input slices were then processed in small mini-batches of cardinality 2 (again, due to memory considerations). In addition to random cropping, input batches were further augmented with horizontal flipping, image blurring, image sharpening, and brightness adjustments.

Variations in thickness of retinal layers introduces a non-trivial amount of class imbalance (as there are fewer pixels corresponding to the thin, inner retinal layers). To mitigate the impact of this class imbalance in training we increased the weight in the loss penalty for the pixels associated with thin layers by a factor of 10 (roughly corresponding to the level of class imbalance). The model was trained for 500 epochs and model weights were saved whenever performance on the validation set improved. Training the model in this fashion took approximately 24 hours on

an NVIDIA Titan X GPU. Processing at inference time is much more rapid, taking only a few hundred milliseconds to process the entire data set. Runtimes for the classical algorithms of record ranged from 28 to 152 seconds [141]; while a direct comparison of runtimes is not entirely fair due to differences in hardware and the use of GPU acceleration, it is reasonable to say that the deep learning methods are not adversely impacting overall runtime.

For training, we use a 10-fold cross validation to implement a "leave-one-patient-out" evaluation process, where each fold corresponds to the five images associated with a single patient. To evaluate performance on a given patient, we use the 45 images from the other nine patients/folds to train a FNC, holding out the last patient's images to use in test. The patient used for testing is then rotated as in conventional k-fold approaches. Of the nine patients available for training in a given fold, one patient was reserved as a validation set. This stratification allowed us to train the network on representative data while ensuring that the segmented images for a given patient were not a by-product of training on that patient's images.

4.2.2.2 Post-processing

After obtaining per-pixel layer estimates from the FCN, one must then generate the corresponding surface estimates. One approach is to directly extract surfaces from the layer estimates by identifying locations where class estimates change along the axial dimension. However, surfaces are defined as a *unique* location in the axial dimension where the layer estimates change and the raw semantic segmentation

outputs do not satisfy this constraint. For example, the right panel of Figure 4.2 shows a few small regions (indicated by arrows) where monotonicity of class estimates is violated due to the presence of small misclassification regions. We explored implementing custom loss functions that incorporate this monotonicity as a soft constraint; however, even this does not guarantee unambiguous surface estimates.

Another option is to employ post-processing heuristics to address these issues. We explored one heuristic which addresses both spurious and missing estimates. When the classification procedure generates more than one candidate for a layer at a given location, the point which is nearest in Euclidean distance to the prior surface is used (in the case of surface 1, distance to surface 2 is used as the adjudication method). Alternately, if a layer estimate is missing for any given location, an estimate is imputed from the nearest available value for that layer. This particular heuristic coupled with the DenseNet FCN segmentation constitutes a baseline algorithm which we term "SEG".

However, hand-crafted heuristics such as these are rather ad-hoc. As an alternative, we propose to explicitly use our prior knowledge that retinal surfaces (in two-dimensional images) can be modeled as scalar-valued functions with an appropriate level of smoothness and solve a regression problem for each surface. For suitable regression procedures, this approach extends naturally to higher dimensions as well (useful in settings where volumetric data is available).

For this study we employ Gaussian processes (GP) regression with a Radial Basis Function (RBF) kernel [129]. The RBF kernel has two hyper-parameters, a noise variance and a characteristic length scale; we select both using a leave-one-patient-

out cross-validation procedure analogous to what was done when training the CNN. To implement the regression we used the GPy software library [72]. Hyperparameter selection was performed by random search over candidates drawn uniformly at random from a two-dimensional hypercube. For this study we observed that using a single RBF kernel for each patient/region pair produced adequate results; however, in settings where there is substantial non-stationarity in the behavior of a patient or region kernel partitioning methodologies may be of value (e.g. [74]). The GP is characterized by the combination of a mean and a covariance function; the mean function was used as our best estimate for the corresponding surface while the covariance provides some measure of the confidence of the estimate. While these estimates are all one-dimensional functions in the two-dimensional plane, we note that GP regression extends naturally to higher dimensions as well. Other post-processing approaches are of course possible; using GPs provides an interesting option in that (a) there is some ability to bring prior knowledge to bear in the form of a covariance prior and (b) it provides a built-in confidence measure, in the form of the covariance estimate, which may provide some additional diagnostic value. This could be especially compelling in settings where images are less uniform in quality or surfaces are more structurally diverse (as might be anticipated in settings with more severe pathologies). We term this combined FCN and GP approach "SEG+REG".

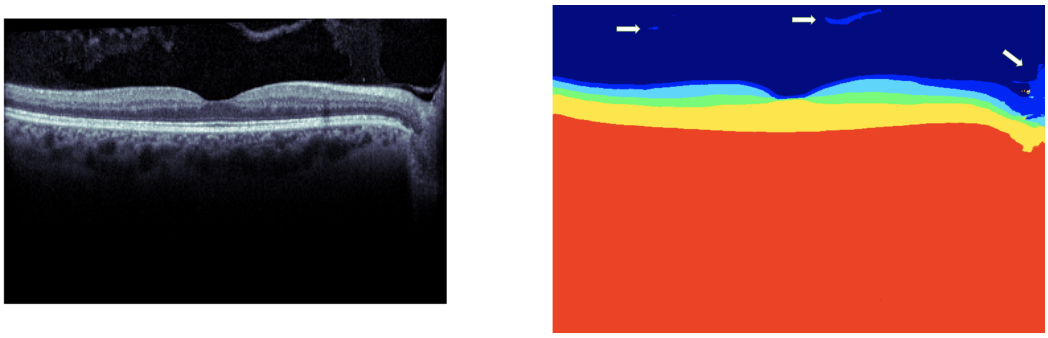


Figure 4.2: Example segmentation; original image (left); neural network segmentation output, before post-processing (right). White arrows denote regions where semantic segmentation layer estimates suffer due to artifacts in the original image.

4.2.3 Data

For this study we utilize the publicly available University of Miami OCT dataset [141]. This data set consists of 50 images spanning 10 different patients with mild, non-proliferative diabetic retinopathy. There are five images associated with each patient: one image of the fovea center, two of the perifovea, and two of the parafovea. Each image consists of 496×768 pixels and the corresponding transversal and axial resolutions are 11.11μm/pixel and 3.867μm/pixel. These images are a subset of volumetric data captured by a Spectralis SD-OCT (Heidelberg Engineering GmbH, Heidelberg, Germany). Two expert graders each independently annotated five retinal surfaces per image, where a "surface" is defined as the boundary between a pair of adjacent retinal layers. The result is a total of 250 annotated surfaces per grader, numbered 1,2,4,6 and 11 (following the convention introduced in [141]). These surfaces and the associated layers are defined in Table 4.1. Following the approach of Tian et al., we use the first grader's annotations as ground truth and the second grader's annotations as a measure of inter-operator agreement.

Surface ID	Upper Layer	Lower Layer
1	Pre-retinal space	Nerve fiber layer
2	Nerve fiber layer	Ganglion cell layer
4	Inner plexiform layer	Inner nuclear layer
6	Outer plexiform layer	Henle's Fiber layer and Outer nuclear layer
11	Bruch's complex	Choriocapillaris

Table 4.1: Annotated surfaces provided by dataset in [141].

Example annotations are shown in Figure 4.3. Magenta lines in the figure denote the estimates generated by human observer #1, which are used as ground truth. Yellow lines depict the estimates produced by one of the algorithms of record (the one generated by the "automated retinal analysis" (AURA) tool suite). Note that, in the case of this image, the AURA estimate does not span the entire horizontal extent of the image. In order to avoid unfairly penalizing any algorithm of record, our metrics (described further below) are computed only for regions where all algorithms produce estimates.

4.2.4 Results

Following the approach in [141], we measure the accuracy of surface estimates by computing the per-pixel differences between the estimate and the ground truth annotations generated by the first manual grader. For a fair comparison, metrics calculations are limited to the regions for which all automated algorithms in the dataset had valid estimates. This unfortunately excludes some remote/lateral regions where cut artifacts are more prevalent (cut artifacts are operator-induced artifacts where the edge of the scan is abnormally truncated, a defect which does not typically affect

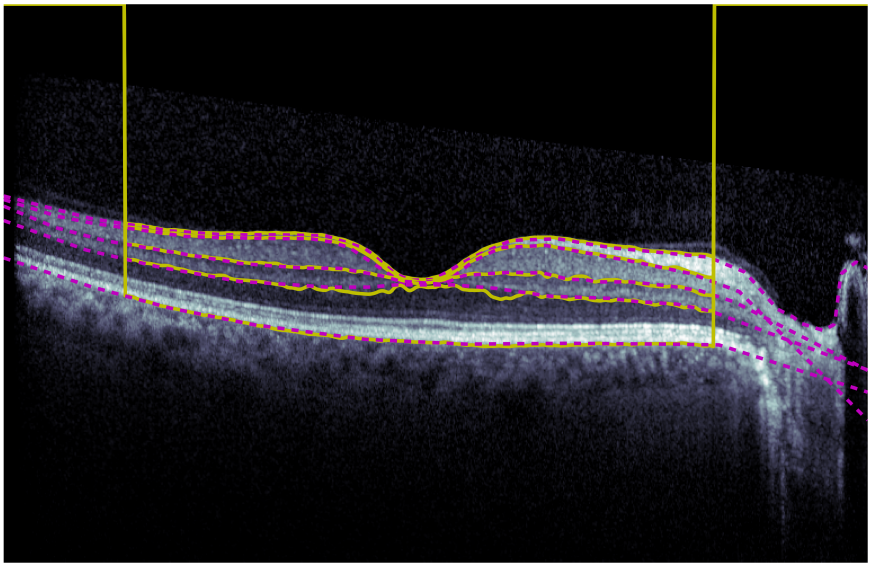


Figure 4.3: Example annotations from the dataset of [141]. Magenta lines correspond to one of the human annotations while yellow lines denote estimates from one algorithm of record (AURA).

central retinal thickness measurements). Thus, a useful direction for future work would be to expand the set of baseline estimates to permit comparisons in these more challenging and dynamic regions. We used mean unsigned errors and mean signed errors as performance metrics for both the proposed algorithms and algorithms of record. For a given surface, the estimate v_{est} and the corresponding ground truth v_{ref} are both vectors (with dimension equal to the width of the evaluation region, in pixels) and the signed error is defined to be

$$e_s = v_{ref} - v_{est};$$

the unsigned error is the absolute value of e_s taken component-wise.

We report the performance of both the SEG and SEG+REG compared with the baseline algorithms. Table 4.2 reports the mean unsigned errors for each algo-

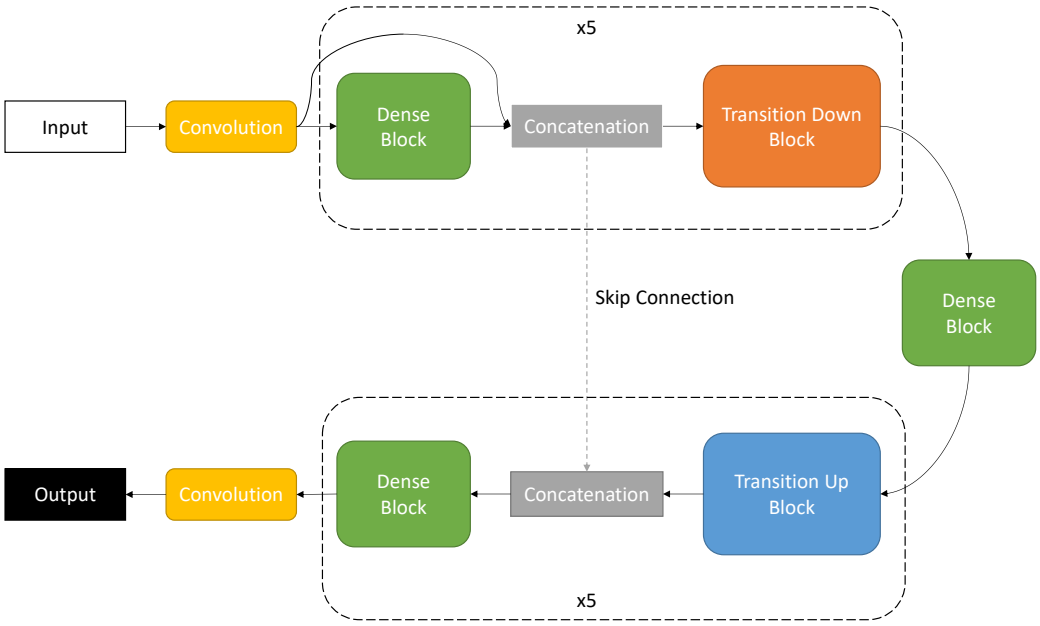


Figure 4.4: Network architecture for the fully convolutional version of DenseNet, summarized in [91].

rithm and surface, and the average and max values across all testing data. Values in bold font indicate when an algorithm performs on-par with human performance (i.e. within the margins of inter-operator error). The table suggests that, in aggregate, our proposed method frequently matches human performance, and performs favorably when compared to other algorithms of record. These results also indicate particularly good performance of the proposed methods on the inner retinal surfaces. Table 4.3 shows the signed errors for the corresponding regions, from which it appears that our method may be slightly overestimating the support of the retinal layers as evidenced by a relatively large positive error on surface 1 and a relatively large negative error on surface 11. Following [141] we also provide the mean unsigned error broken down by ocular regions in Table 4.4 [1].

[1]Note there are some minor differences between these results and table 5 of [141] for the algorithms of record which may be attributed to variations in the extent of the macular region that was evaluated; many of the automated methods tend to exhibit greater variation towards the edges of

Table 4.2: Mean unsigned error aggregated across all eye regions. Values in bold indicate when an algorithm meets or exceeds inter-observer (I-O) performance.

	SEG	SEG+REG	Spectralis	OCTRIMA	AURA	IOWA	Bern	I-O
surface 1	1.13	1.11	1.09	0.95	1.35	2.03	1.71	0.87
surface 2	**1.14**	**1.07**	1.45	1.18	1.19	1.74	2.77	1.14
surface 4	**0.95**	**0.90**	1.92	**0.99**	1.12	1.79	1.60	1.10
surface 6	**1.23**	**1.18**	**1.19**	1.52	1.54	1.51	1.72	1.29
surface 11	**1.06**	**1.02**	**0.99**	1.20	**0.96**	1.22	1.24	1.12
mean	**1.10**	**1.06**	1.33	1.17	1.23	1.66	1.81	1.10
std	0.10	0.10	0.37	0.23	0.22	0.30	0.57	0.15
min	0.95	0.90	0.99	0.95	0.96	1.22	1.24	0.87
max	1.23	1.18	1.92	1.52	1.54	2.03	2.77	1.29

Table 4.3: Mean signed error across all eye regions.

	SEG	SEG+REG	Spectralis	OCTRIMA	AURA	IOWA	Bern	I-O
surface 1	0.90	0.91	-0.82	0.66	1.22	1.99	1.65	0.26
surface 2	-0.12	-0.13	0.76	0.16	0.34	1.47	2.53	0.29
surface 4	0.18	0.18	1.43	0.12	0.41	1.59	1.30	0.29
surface 6	-0.30	-0.29	-0.51	-0.92	-0.51	0.78	1.13	0.09
surface 11	-0.66	-0.65	-0.44	-0.94	-0.58	1.04	0.90	-0.69

Table 4.4: Mean unsigned error for all surfaces and regions.

	SEG	SEG+REG	Spectralis	OCTRIMA	AURA	IOWA	Bern	I-O
surface1 fovea	1.18	1.14	0.90	0.90	0.90	2.14	1.67	0.85
surface1 parafovea	1.12	1.10	1.14	1.00	1.31	1.98	1.81	0.89
surface1 perifovea	1.12	1.10	1.13	0.92	1.62	2.01	1.62	0.86
surface2 fovea	1.34	**1.25**	1.39	**1.15**	**1.29**	2.42	2.02	1.31
surface2 parafovea	1.03	0.98	**0.92**	1.03	**0.92**	1.59	2.45	0.97
surface2 perifovea	**1.15**	**1.09**	2.02	1.35	1.42	1.54	3.47	1.22
surface4 fovea	**1.10**	**1.03**	1.30	**1.12**	1.25	1.81	1.44	1.13
surface4 parafovea	**0.91**	**0.88**	1.32	**0.91**	**1.02**	1.67	1.52	1.08
surface4 perifovea	**0.92**	**0.86**	2.82	**1.00**	1.14	1.89	1.76	1.11
surface6 fovea	**1.45**	**1.38**	1.79	2.75	2.58	1.58	1.86	1.50
surface6 parafovea	**1.26**	**1.22**	**1.10**	**1.36**	1.42	1.50	1.74	1.36
surface6 perifovea	**1.08**	**1.04**	**0.99**	**1.08**	1.14	1.49	1.62	1.11
surface11 fovea	**0.92**	**0.87**	**0.81**	**1.02**	**0.88**	**1.08**	1.23	1.12
surface11 parafovea	**1.07**	**1.03**	**0.98**	1.19	**0.95**	1.14	1.16	1.12
surface11 perifovea	**1.11**	**1.07**	**1.07**	1.31	**1.02**	1.38	1.32	1.11
mean	**1.12**	**1.07**	1.31	1.21	1.26	1.68	1.78	1.12
std	0.15	0.14	0.53	0.45	0.43	0.37	0.57	0.18
min	0.91	0.86	0.81	0.90	0.88	1.08	1.16	0.85
max	1.45	1.38	2.82	2.75	2.58	2.42	3.47	1.50

the scans and we evaluate on the largest common intersection across all algorithms.

4.2.5 Discussion

The results suggest that semantic segmentation using a fully convolutional network using DenseNets together with suitable post-processing using GP is a promising approach to address the problem of fine-grained automated OCT segmentation, a capability with many clinical applications. Our results compare well with existing algorithms of record, often resulting in the smallest mean unsigned errors; overall, performance is largely comparable with human annotation. We note, however, that caution should be exercised when drawing conclusions since the algorithms of record we compare against were developed and optimized using datasets which may not match exactly the University of Miami evaluation dataset (e.g. in aspects such as resolution, noise characteristics, and artifacts).

While our study focused upon estimating OCT surfaces in 2D images, we note that the method extends naturally to 3D volumetric data (e.g. see [43]) and that the semantic segmentation component may also help identify other clinically important structures (such as drusen or other lesions). Furthermore, the GP-based post-processing comes equipped with an uncertainty estimate that could prove useful in some settings. For example, it could be advantageous in situations where we need to provide a plausible range of uncertainty for the prediction when processing time-varying patterns in clinical data acquired longitudinally.

While promising, there are other directions along which this study could be improved moving forward. One potential limitation of our post-processing approach is that by estimating surfaces independently, there is no theoretical guarantee that

the resulting collections of surfaces do not intersect. Another current limitation is the dataset size. The total number of B-scans (50) is not the largest publicly available dataset (e.g. [41] consists of 110 B-scans). Furthermore, the mild nature of the pathologies manifested in our dataset suggests that analysis on more severe cases would be of value. For example, these results may not be representative of other pathologies, such as DME.

It is worth noting that the number of images used in each fold for training, while modest (45), was adequate for performing semantic segmentation. The data volume is comparable with those used in the original U-Net 2D and 3D studies [43, 130] and also aligns with reports by Devalla et al. which discusses the apparent misconception related to the need of a significant training dataset in semantic segmentation [55], a point which could appear at first counterintuitive given that it is a well known fact it takes a much larger number of images in training to perform full image classification. Since each pixel in the image has an associated class label there is substantially more information available for training on a per-image basis. Of course, many of these per-pixel "examples" share context and are therefore highly correlated, there is still a substantial "force multiplier" which arises from dense labels.

Another interesting direction of future work is to investigate performance in more remote/lateral regions of the eye. Other future work involves developing and/or testing additional datasets that are reflective of broader pathologies and permit more comprehensive comparison with other recent methods. This is especially compelling given the variety of neurodegenerative diseases which can manifest

as as abnormalities in the retina.

4.3 Application: Classification Using Hybrid Features

4.3.1 Objective

In this section we consider machine learning methods that exploit a mixture of visual and non-visual features. In particular, we are interested "side channel" information consisting of patient demographic data that may be ordinal or categorical in nature (and thus lacking the metric structure associated with image data). Our motivating application continues to be the automatic detection of age-related macular degeneration (recall Section 4.1.1); however, here we will focus on whole-image classification (i.e. assigning a single classification label to an entire image, in contrast to the per-pixel estimates considered in Section 4.2). This study asks whether random forests (which are well-suited for working with non-metric data) can provide an effective mechanism for jointly leveraging image features together with non-metric demographic information. While none of the individual algorithmic components we consider here are new (deep transfer learning, random forest classification, PCA-based dimension reduction) their joint application to this data set (and the resulting considerations) constitute a novel contribution.

4.3.2 Approach

The key idea behind our approach is to leverage the power of deep transfer learning for extracting features from image-based data together with the ability

Algorithm 4 Classification pipeline; parentheticals specify the dimensions for a single object on the left-hand side.

procedure CLASSIFY(fundusImages, demographics, y, cnnParams, nFolds=5)
    ▷ *Pre-processing (Section 4.3.2.1)*
    xRegion1, xRegion2 = preprocess_images(fundusImages); ▷ $(231 \times 231 \times 2)$
    ▷ *Image Feature Extraction (Section 4.3.2.2)*
    xFeatImg1 = overfeat_features(xRegion1,cnnParams); ▷ (4096×1)
    xFeatImg2 = overfeat_features(xRegion2,cnnParams); ▷ (4096×1)
    xFeatImgBothRegions = concat(xFeatImg1, xFeatImg2); ▷ (8192×1)
    xFeatImg = PCA_dimension_reduction(xFeatImgBothRegions); ▷ (100×1)
    ▷ *Classification (Section 4.3.2.3)*
    xFeat = concat(xFeatImg, demographics); ▷ (112×1)
    groups = k_fold_per_patient(xFeatImg, demographics, nFolds);
    for $i = 1 \ldots$ nFolds **do**
        model = train_rf_classifier(xFeat[groups $\neq i$], y[groups $\neq$i]);
        yHat[i] = predict(model, xFeat[groups == i]);
    end for
    return yHat;
end procedure

of random forests to simultaneously exploit metric and non-metric features. Algorithm 4 summarizes the overall approach; details are provided in subsequent sections.

4.3.2.1 Pre-processing

We preprocessed all raw fundus images by cropping and normalizing each image. This procedure is the same as the preprocessing steps described in [29], which we recapitulate here. The first step is to remove extraneous background pixels unrelated to the structure of the eye. This involved running a simple detector to identify the approximate boundary of the retina and computing a square crop about this region (see Figure 4.5); this removes superflous background pixels unrelated to the structure of the eye. We then extracted two new sub-images corresponding to two overlapping regions - one restricted to the center of the eye and another which includes peripheral areas. This pair of images provides data at two coarse

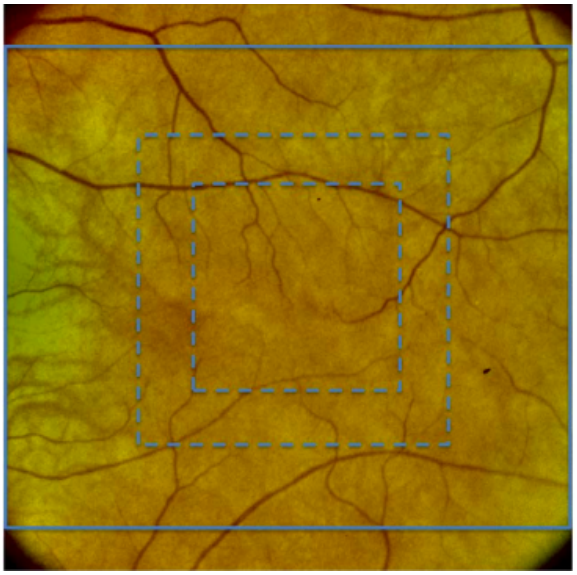


Figure 4.5: Cropping fundus images. The crops used to generate CNN input images are shown in dashed lines. Figure taken from [29].

"scales" to the deep learning feature-extractor. The center of the fundus image typically contains the most useful discriminatory information, while the periphery can provide additional context for more difficult problem instances. These regions are depicted by dashed lines in Figure 4.5.

The cropped image are then run through a intensity normalization step, whereby the image is converted into the Lab color space, and the lightness channel (L) is processed to lower the intensity gradient of the image. This step takes dark shadows and bright highlights and brings them closer to the average intensity of the image. Finally, each image is resized to 231×231 pixels for use in the transfer learning step described below.

4.3.2.2 Image Feature Extraction

To generate image-based features for classification we use the transfer learning methodology described in [27, 29]. The general idea is to augment our somewhat limited collection of images (limited from a deep learning standpoint) by training a classifier on a different problem with similar characteristics for which we have abundant data. Then, we use this classifier to generate features for our problem of interest by extracting activations from intermediate layers of the network. In this case, we use the OverFeat (OF) [135] network that was pre-trained on over 1.2 million general purpose, non-medical images, including classes of various objects, such as animals, edible items, and household objects. Without re-training we provide our pre-processed fundus images to this network and use the output from the first "fully connected" layer (which has dimensions 4096×1) as our feature vectors. Since we have a pair of images for each original fundus image (one for each region) we concatenate these to form a single 8192-dimensional feature vector.

4.3.2.3 Classification

For classification we combine the image-based features with the demographic "side channel" features to train a Random Forest (RF) classifier [21]. Random forests are an ensemble classification strategy where each element in the ensemble is a classification and regression tree (CART). To mitigate the problem of overfitting the ensemble to the training data, each tree is trained on a randomly sampled subset of the feature vectors (a bootstrapping procedure known as *bagging*). In addition,

a random sample of the features is used at each node of the tree to determine the split at that level. These procedures trees that individually tend to be fairly weak predictors; however, in aggregate, the ensemble produces a classifier with good performance.

Note that RF classifiers can easily deal with feature sets that contain a mix of categorical and ranked ordinal features. In addition, one can compensate for missing feature values by introducing the CART notion of surrogate splits [22] which introduces secondary split points at each node of each tree that are utilized if the normal value that would be split on is missing. In the sequel, all RF tree generation was done using such surrogate splits.

The aforementioned variable splitting used in RF is potentially problematic for our application. Normally, one samples on the order of the square root of the number of features at each tree node in order to determine a split point for that node. However, the feature set imbalance between the number of side channel features and the number of retinal image features would cause the side channel features to have a minimal influence in the resulting classifier. In order to correct for this imbalance, we use principal components analysis (PCA) to reduce the number of image-based features to 100 prior to training. The resulting features capture 88.35% of the total variance of the total image features, suggesting that this is not an unreasonable strategy for dimension reduction.

Normally, testing an RF classifier is done using out-of-bag computations; this leverages the fact that each tree in the ensemble is trained on a random subset of the features, allowing the unseen portion of the data to be used as a surrogate

test set for assessing the performance of that particular tree. However, certain properties of the demographic data present challenges. The 4587 patients in the study are associated with multiple records corresponding to multiple observations of each patient over time (see Section 4.3.3). These records will have nearly identical side channel features for a number of fields (e.g. GENDER and ETHNICITY are invariant across visits), which would lead to a partial co-mingling of the testing and training in the out-of-bag exemplars. Even features that should have changed over the course of the study, such as VISUAL_ACUITY, often show little variability over multiple visits for a single patient. In order to mitigate this effect, we customize the k-fold cross-validation approach to ensure that the observations associated with a given patient are confined to a single fold. We then use $k-1$ of those folds to train our classifier and the remaining kth fold to test the classifier. This process is repeated k times, leaving out a different fold for training each time. In this investigation, we used $k = 5$. In addition, to compensate for the random sampling inherent in RF classifiers, we repeat each cross-validation 5 times, and report means and standard deviations of each statistic for which it was appropriate.

4.3.3 Data

The data used in this study are taken from the National Institutes of Health (NIH) Age-Related Eye Disease Study (AREDS) 2014 dataset. The AREDS was a twelve-year, multi-center, prospective cohort study designed to increase understanding of disease progression and risk factors for both AMD and age-related cataracts

as well as develop potential therapies [5, 24]. During enrollment, a wide range of patient characteristics and medical information was obtained for each patient, including demographics (e.g. gender, age, race, education), health history (e.g. blood pressure, cancer, smoking, angina, diabetes), prior eye treatments, vision status, and current supplementation and medication use. In addition, at enrollment, color fundus photographs were taken and graded for severity of AMD. Routine follow-up visits occurred every 6 months, at which time the above medical information was updated for each patient. Approximately every 12 months additional color fundus photographs were taken and manually graded for AMD severity and the scores recorded. In all, images from 4613 patients were obtained along with their corresponding medical data [1].

At each fundus photograph session, four images were typically acquired. These images corresponding to left and right stereo pairs (denoted LS and RS) of both the left and right eyes (denoted LE and RE). Each of these four images was manually graded for AMD severity. However, for some patients, images of only one eye and/or one stereo pair were taken which resulted in an unequal number of images as well as patients in each of the four possible types of retinal images (i.e. LE-LS, LE-RS, RE-LS, and RE-RS). Therefore, in our experiments we used exclusively the 33578 LE-LS images taken from 4587 patients.

The demographic side channel information also has some data availability issues. A number of fields have missing values for a large proportion of the data records. For our study we identified 12 fields that were present in a majority of the records considered. A list of these features, the number of missing entries, and the

Table 4.5: Side channel features used in this study and how frequently they were missing (out of 4587 patients).

Feature	Num. Missing Entries	Categorical?
CORTICAL_OPACITY_FIELD	63	NO
CORTICAL_OPACITY_5MM	63	NO
PSC_OPACITY_5MM	63	NO
SUNLIGHT	286	NO
VISUAL_ACUITY	0	NO
ETHNICITY	0	YES
DIABETES	1166	YES
EDUCATION	0	YES
LASER_PHOTOCOAGULATION	156	YES
GLAUCOMA	1165	YES
CATARACT_SURGERY	59	YES
GENDER	0	YES

nature of the feature (categorical or numeric) are summarized in Table 4.5.

4.3.4 Results

For the first set of experiments, we generated OF features for the left eye, left stereo images in each record and combined them with the side-channel information associated with the record, resulting in 33578 feature vectors, each of which was 112-dimensional. We then looked at a binary classification problem, where class 0 corresponded to AMD severity categories of 1 or 2, and class 1 corresponded to AMD severity categories of 3 or 4. Note that there is a mild class imbalance in the data set. There were 18,798 examples associated with records of class 0, and 14,780 examples associated with record of class 1. In the RF generation, we enforced a uniform prior on sampling from the classes to compensate for this imbalance.

The accuracy, sensitivity, specificity and area under the ROC curve (AUC)

corresponding to both the hybrid forest and the forest using only the image features are given in Table 4.6. As indicated in the table, both the specificity and sensitivity of the classifier are slightly improved by adding the side channel data. In addition, the AUC increase shows more capacity for adjusting the classification thresholds to improve the specificity with little decrease in sensitivity by using the side channel information.

In order to test that the improvements in specificity, sensitivity and the AUC are statistically significant, we used the Kolmagorov-Smirnov (KS) test to determine if the distributions both with and without the side channel information could have been drawn from the same distribution. In all cases, the p-values of the KS test were 0.01 or below, indicating that the differences in the results were statistically significant at the 1% level. However, while the side channel features do provide some benefit to the classifier, the bulk of the classification performance is derived from the image features.

While the benefits of the demographic information in this study are modest, in situations where there is a paucity of training data we hypothesize the side channel information could provide some additional discriminative power. In order to test this, we set up a second experiment. We took the initial data used to generate each classifier and downsampled it, i.e. we trained each random forest on only 10% of the associated training fold and then tested on the entire associated testing fold. The results of this experiment are summarized in Table 4.7. In this case, there is a more noticeable benefit to adding the demographic information,as indicated by the specificity, sensitivity and in the AUC for these classifiers. Repeating the

Table 4.6: Statistics for experiments on full data set

	Specificity	Sensitivity	AUC	Acc.
side	.6973 (8.74e-4)	.5097 (.0029)	.6515 (.0013)	.6146
image	.8821 (.0014)	.6569 (.0006)	.8444 (6.08e-4)	.7834
both	.8895 (4.63e-4)	.6634 (5.40e-4)	.8476 (9.43e-4)	.7904

Table 4.7: Statistics for experiments on reduced data set

	Specificity	Sensitivity	AUC	Acc.
image	.8264 (.0025)	.6128 (.0024)	.7961 (.0014)	.7322
both	.8355 (.0035)	.6257 (.0027)	.8065 (9.71e-4)	.7430

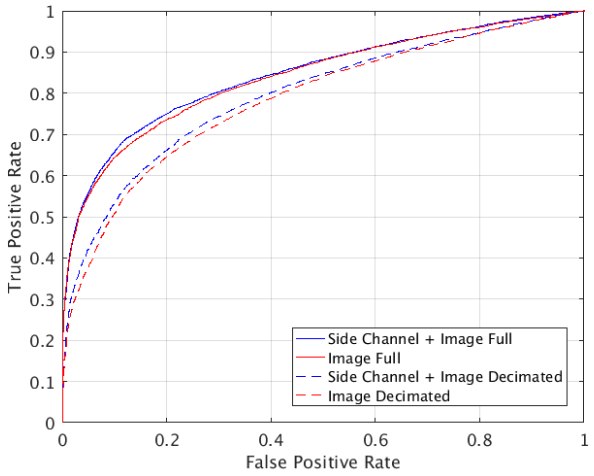


Figure 4.6: ROC curves for generated classifiers.

Kolmogorov-Smirnov test for these samples again results in p-values at or below 0.01. ROC curves for both experiments are provided in Figure 4.6.

4.3.5 Discussion

The proposed hybrid deep-RF architecture shows promise with regard to performance enhancement as there is a measurable improvement which is statistically significant in incorporating both types of information via the proposed approach,

despite some challenges in the dataset that are highlighted below. Also of note is the fact that the addition of side channel information takes on more importance – as might be expected – in the case when training data is less abundant.

One important point is that there are other side channel features present in the AREDS data that could be used to supplement the fundus image data in our experiments. However, the incomplete nature of these features led us to omit them in our initial studies. The smoking features, in particular, are missing in over two thirds of the patient records. One possible future direction is to combine the various smoking features to try to impute a new smoking feature over most of the data set. Past studies have identified smoking as highly correlated with AMD and therefore improved classification performance might be possible by using this demographic feature [134]. Other side channel information related to the vitamin treatments employed during the AREDS study might also be of diagnostic benefit [134].

As indicated earlier, an additional complication in the data is the fact that multiple images were associated with some patients, corresponding to multiple office visits. The fact that the side channel data associated with these multiple visits was nearly identical led to challenges in utilizing standard RF methods for model testing and assessing variable importance. Even the VISUAL_ACUITY feature, which should have shown some change for patients over multiple visits, did not exhibit such changes. Some of these features were likely subject to measurement noise during collection which may also explain the relative diagnostic value of the image data in our studies. Despite the modest improvements provided by these particular features, this study provides a good foundation for future investigations

to incorporate demographic data with visual features for AMD detection.

4.4 Conclusions

In this chapter we considered two medical image processing tasks, both related to the task of automatically identifying ocular disease with a particular focus on AMD. We proposed and analyzed automated segmentation methods which suggest that semantic segmentation using FCNs and DenseNet architectures, coupled with regression-based post-processing using GP, can effectively help address the automated OCT segmentation problem on par with human capabilities in cases of patients with mild retinopathy and improve upon the reference algorithms of record. In addition, we proposed and analyzed a random forest-based approach to incorporating deep visual image features with ordinal or categorical side channel information. We show in this preliminary study that there is a statistically significant gain in the combination of the two heterogeneous types of information through this hybrid approach.

Bibliography

[1] AREDS dbGaP data tables: A users guide version 1.0. URL `https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/GetPdf.cgi?id=phd001552.1`.

[2] Scattering networks in matlab. `https://github.com/scatnet/scatnet`. Accessed: June 2018.

[3] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: a system for large-scale machine learning. In *OSDI*, volume 16, pages 265–283, 2016.

[4] Michael David Abràmoff, Yiyue Lou, Ali Erginay, Warren Clarida, Ryan Amelon, James C Folk, and Meindert Niemeijer. Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning detection of diabetic retinopathy. *Investigative Ophthalmology & Visual Science*, 57(13):5200–5206, 2016.

[5] Age-Related Eye Disease Study Research Group et al. A randomized, placebo-controlled, clinical trial of high-dose supplementation with vitamins C and E, beta carotene, and zinc for age-related macular degeneration and vision loss: AREDS report no. 8. *Archives of ophthalmology*, 119(10):1417, 2001.

[6] Age-Related Eye Disease Study Research Group et al. Potential public health impact of age-related eye disease study results: AREDS report no. 11. *Archives of ophthalmology*, 121(11):1621, 2003.

[7] Joakim Andén, Vincent Lostanlen, and Stéphane Mallat. Joint time-frequency scattering for audio classification. In *Machine Learning for Signal Processing (MLSP), 2015 IEEE 25th International Workshop on*, pages 1–6. IEEE, 2015.

[8] Jean-Pierre Antoine and Romain Murenzi. Two-dimensional directional wavelets and the scale-angle representation. *Signal processing*, 52(3):259–281, 1996.

[9] Jean-Pierre Antoine, Pierre Carrette, R Murenzi, and Bernard Piette. Image analysis with two-dimensional continuous wavelet transform. *Signal processing*, 31(3):241–272, 1993.

[10] Anish Athalye and Nicholas Carlini. On the robustness of the CVPR 2018 white-box adversarial example defenses. *arXiv preprint arXiv:1804.03286*, 2018.

[11] Anish Athalye and Ilya Sutskever. Synthesizing robust adversarial examples. *arXiv preprint arXiv:1707.07397*, 2017.

[12] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. *arXiv preprint arXiv:1802.00420*, 2018.

[13] Radu Balan, Maneesh Singh, and Dongmian Zou. Lipschitz properties for deep convolutional networks. *arXiv preprint arXiv:1701.05217*, 2017.

[14] John J Benedetto. *Wavelets: mathematics and applications*, volume 13. CRC press, 1993.

[15] John J Benedetto and MW Frazier. Wavelets. *Mathematics and Applications, CRC PRESS, Inc., Boca Raton-London-Tokyo*, pages 1–12, 1994.

[16] John J Benedetto, Wojciech Czaja, Przemystaw Gadziński, and Alexander M Powell. The Balian-Low theorem and regularity of Gabor systems. *The Journal of Geometric Analysis*, 13(2):239, 2003.

[17] Alan Bird, Neil Bressler, et al. An international classification and grading system for age-related maculopathy and age-related macular degeneration. *Survey of ophthalmology*, 39(5):367–374, 1995.

[18] Jeffrey D Blanchard. Minimally supported frequency composite dilation wavelets. *Journal of Fourier Analysis and Applications*, 15(6):796, 2009.

[19] Jeffrey D Blanchard and Kyle R Steffen. Crystallographic Haar-type composite dilation wavelets. In *Wavelets and Multiscale Analysis*, pages 83–108. Springer, 2011.

[20] Anna Breger, Martin Ehler, Hrvoje Bogunovic, Sebastian Waldstein, Ana-Maria Philip, Ursula Schmidt-Erfurth, and Bianca Gerendas. Supervised learning and dimension reduction techniques for quantification of retinal fluid in optical coherence tomography images. *Eye*, 2017.

[21] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[22] Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen. *Classification and regression trees*. CRC press, 1984.

[23] Neil M Bressler. Age-related macular degeneration is the leading cause of blindness. *JAMA*, 291(15):1900–1901, 2004.

[24] Neil M Bressler, Tom S Chang, Ivan J Suñer, Jennifer T Fine, Chantal M Dolan, James Ward, Tsontcho Ianchulev, et al. Vision-related function after ranibizumab treatment by better-or worse-seeing eye: clinical trial results from MARINA and ANCHOR. *Ophthalmology*, 117(4):747–756, 2010.

[25] Tom B Brown, Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. Adversarial patch. *arXiv preprint arXiv:1712.09665*, 2017.

[26] Joan Bruna and Stéphane Mallat. Invariant scattering convolution networks. *IEEE transactions on pattern analysis and machine intelligence*, 35(8), 2013.

[27] Philippe Burlina, David E Freund, Neil Joshi, Y Wolfson, and Neil M Bressler. Detection of age-related macular degeneration via deep learning. In *Biomedical Imaging (ISBI), 2016 IEEE 13th International Symposium on*, pages 184–188. IEEE, 2016.

[28] Philippe Burlina, Neil Joshi, Michael Pekala, Katia Pacheco, David E Freund, and Neil M Bressler. Automated grading of age-related macular degeneration from color fundus images using deep convolutional neural networks. *JAMA Ophtalmology*, 135(11):1170–1176, 2017.

[29] Philippe Burlina, Katia D Pacheco, Neil Joshi, David E Freund, and Neil M Bressler. Comparing humans and deep learning performance for grading AMD: A study in using universal deep features and transfer learning for automated AMD analysis. *Computers in Biology and Medicine*, 82:80–86, 2017.

[30] Phillipe Burlina, David Freund, Bénédicte Dupas, and Neil Bressler. Automatic screening of age-related macular degeneration and retinal abnormalities. In *Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE*, pages 3962–3966. IEEE, 2011.

[31] Carlos Cabrelli, Christopher Heil, and Ursula Molter. Accuracy of lattice translates of several multidimensional refinable functions. *Journal of Approximation Theory*, 95(1):5–52, 1998.

[32] Emmanuel J Candès and David L Donoho. Ridgelets: A key to higher-dimensional intermittency? *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 357(1760): 2495–2509, 1999.

[33] Emmanuel J Candès and David L Donoho. Curvelets and curvilinear integrals. *Journal of Approximation Theory*, 113(1):59–90, 2001.

[34] Emmanuel J Candès and David L Donoho. New tight frames of curvelets and optimal representations of objects with piecewise c2 singularities. *Communications on pure and applied mathematics*, 57(2):219–266, 2004.

[35] Nicholas Carlini and David Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pages 3–14. ACM, 2017.

[36] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *Security and Privacy (SP), 2017 IEEE Symposium on*. IEEE, 2017.

[37] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2 (3):27, 2011.

[38] Shang-Tse Chen, Cory Cornelius, Jason Martin, and Duen Horng Chau. Robust physical adversarial attack on faster r-cnn object detector. *arXiv preprint arXiv:1804.05810*, 2018.

[39] Xu Chen, Xiuyuan Cheng, and Stéphane Mallat. Unsupervised deep haar scattering on graphs. In *Advances in Neural Information Processing Systems*, pages 1709–1717, 2014.

[40] Xiuyuan Cheng, Xu Chen, and Stéphane Mallat. Deep haar scattering networks. *Information and Inference: A Journal of the IMA*, 5(2):105–133, 2016.

[41] Stephanie J Chiu, Michael J Allingham, Priyatham S Mettu, Scott W Cousins, Joseph A Izatt, and Sina Farsiu. Kernel regression based segmentation of optical coherence tomography images with diabetic macular edema. *Biomedical optics express*, 6(4):1172–1194, 2015.

[42] Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional map of the world. *arXiv preprint arXiv:1711.07846*, 2017.

[43] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 424–432. Springer, 2016.

[44] Dan Ciresan, Alessandro Giusti, Luca M Gambardella, and Jürgen Schmidhuber. Deep neural networks segment neuronal membranes in electron microscopy images. In *Advances in neural information processing systems*, pages 2843–2851, 2012.

[45] Edouard. Colas, A. Besse, A. Orgogozo, B. Schmauch, N. Meric, and E. Besse. Deep learning approach for diabetic retinopathy screening. *Acta Ophthalmologica*, 94, 2016. ISSN 1755-3768. doi: 10.1111/j.1755-3768.2016.0635.

[46] Wojciech Czaja and Emily J King. Isotropic shearlet analogs for L2 (R k) and localization operators. *Numerical functional analysis and optimization*, 33(7-9):872–905, 2012.

[47] Wojciech Czaja and Emily J King. Anisotropic shearlet transforms for L2. *Mathematische Nachrichten*, 287(8-9):903–916, 2014.

[48] Wojciech Czaja and Weilin Li. Analysis of time-frequency scattering transforms. *Applied and Computational Harmonic Analysis*, 2017.

[49] Wojciech Czaja and Weilin Li. Rotationally invariant time-frequency scattering transforms. *arXiv preprint arXiv:1710.06889*, 2017.

[50] Wojciech Czaja, Julia Dobrosotskaya, and Benjamin Manning. Composite wavelet representations for reconstruction of missing data. In *Independent Component Analyses, Compressive Sampling, Wavelets, Neural Net, Biosystems, and Nanoengineering XI*, volume 8750, page 875003. International Society for Optics and Photonics, 2013.

[51] Wojciech Czaja, Neil Fendley, Michael Pekala, Christopher Ratto, and I-Jeng Wang. Adversarial examples in remote sensing. *arXiv:1805.10997*, 2018.

[52] Wojciech Czaja, Neil Fendley, Michael Pekala, Christopher Ratto, and I-Jeng Wang. Adversarial examples in remote sensing. *to appear the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (ACM SIGSPATIAL 2018)*, 2018.

[53] Wojciech Czaja, Ilya Kavalerov, and Weilin Li. Scattering transforms and classification of hyperspectral images. In *Proc. of SPIE Vol*, volume 10644, pages 106440H–1, 2018.

[54] Delia Cabrera DeBuc. A review of algorithms for segmentation of retinal image data using optical coherence tomography. In *Image Segmentation*. InTech, 2011.

[55] Sripad Krishna Devalla, Khai Sing Chin, Jean-Martial Mari, Tin A Tun, Nicholas G Strouthidis, Tin Aung, Alexandre H Thiéry, and Michaël JA Girard. A deep learning approach to digitally stain optical coherence tomography images of the optic nerve head. *Investigative ophthalmology & visual science*, 59(1):63–74, 2018.

[56] DigitalGlobe. IKONOS: Data sheet, 2013. URL `https://dg-cms-uploads-production.s3.amazonaws.com/uploads/document/file/96/DG_IKONOS_DS.pdf`. [Online; accessed 11-May-2018].

[57] Minh N Do, Yue M Lu, et al. Multidimensional filter banks and multiscale geometric representations. *Foundations and Trends® in Signal Processing*, 5 (3):157–264, 2012.

[58] Pascal A Dufour, Lala Ceklic, Hannan Abdillahi, Simon Schroder, Sandro De Dzanet, Ute Wolf-Schnurrbusch, and Jens Kowal. Graph-based multi-surface segmentation of OCT data using trained hard and soft constraints. *IEEE transactions on medical imaging*, 32(3):531–543, 2013.

[59] Glenn R Easley, Demetrio Labate, and Vishal M Patel. Directional multiscale processing of images using wavelets with composite dilations. *Journal of mathematical imaging and vision*, 48(1):13–34, 2014.

[60] Ivan Evtimov, Kevin Eykholt, Earlence Fernandes, Tadayoshi Kohno, Bo Li, Atul Prakash, Amir Rahmati, and Dawn Song. Robust physical-world attacks on deep learning models. *arXiv preprint arXiv:1707.08945*, 1, 2017.

[61] Fartash Faghri, Ian Goodfellow, Justin Gilmer, Luke Metz, Maithra Raghu, and Sam Schoenholz. Adversarial spheres. 2018.

[62] Leyuan Fang, David Cunefare, Chong Wang, Robyn H Guymer, Shutao Li, and Sina Farsiu. Automatic segmentation of nine retinal layer boundaries in OCT images of non-exudative amd patients using deep learning and graph search. *Biomedical Optics Express*, 8(5):2732–2744, 2017.

[63] Alhussein Fawzi, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. Robustness of classifiers: from adversarial to random noise. In *Advances in Neural Information Processing Systems*, pages 1632–1640, 2016.

[64] Alhussein Fawzi, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. The robustness of deep networks: A geometrical perspective. *IEEE Signal Processing Magazine*, 34(6):50–62, 2017.

[65] Albert K Feeny, Mongkol Tadarati, David E Freund, Neil M Bressler, and Philippe Burlina. Automated segmentation of geographic atrophy of the retinal epithelium via random forests in AREDS color fundus images. *Computers in biology and medicine*, 65:124–136, 2015.

[66] David E Freund, Neil Bressler, and Philippe Burlina. Automated detection of drusen in the macula. In *Biomedical Imaging: From Nano to Macro, 2009. ISBI'09. IEEE International Symposium on*, pages 61–64. IEEE, 2009.

[67] Rishab Gargeya and Theodore Leng. Automated identification of diabetic retinopathy using deep learning. *Ophthalmology*, 2017.

[68] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*, 2015.

[69] Heidelberg Engineering GmbH. Spectralis HRA+OCT user manual software, 2014.

[70] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[71] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.

[72] GPy. GPy: A Gaussian process framework in Python. `http://github.com/SheffieldML/GPy`, since 2012.

[73] Loukas Grafakos and Christopher Sansing. Gabor frames and directional time–frequency analysis. *Applied and Computational Harmonic Analysis*, 25(1): 47–67, 2008.

[74] Robert B Gramacy and Herbert K H Lee. Bayesian treed gaussian process models with an application to computer modeling. *Journal of the American Statistical Association*, 103(483):1119–1130, 2008.

[75] Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, 316(22):2402–2410, 2016. doi: 10.1001/jama.2016.17216.

[76] Kanghui Guo and Demetrio Labate. Optimally sparse multidimensional representation using shearlets. *SIAM journal on mathematical analysis*, 39(1): 298–318, 2007.

[77] Kanghui Guo, Demetrio Labate, Wang-Q Lim, Guido Weiss, and Edward Wilson. Wavelets with composite dilations. *Electronic research announcements of the American Mathematical Society*, 10(9):78–87, 2004.

[78] Kanghui Guo, Gitta Kutyniok, and Demetrio Labate. Sparse multidimensional representations using anisotropic dilation and shear operators, 2006.

[79] Kanghui Guo, Demetrio Labate, Wang-Q Lim, Guido Weiss, and Edward Wilson. The theory of wavelets with composite dilations. In *Harmonic analysis and applications*, pages 231–250. Springer, 2006.

[80] Kanghui Guo, Demetrio Labate, Wang-Q Lim, Guido Weiss, and Edward Wilson. Wavelets with composite dilations and their MRA properties. *Applied and Computational Harmonic Analysis*, 20(2):202–236, 2006.

[81] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning*, volume 1. Springer series in statistics New York, NY, USA:, 2001.

[82] Yufan He, Aaron Carass, Yeyi Yun, Can Zhao, Bruno M Jedynak, Sharon D Solomon, Shiv Saidha, Peter A Calabresi, and Jerry L Prince. Towards topological correct segmentation of macular OCT from cascaded FCNs. In *Fetal, Infant and Ophthalmic Medical Image Analysis*, pages 202–209. Springer, 2017.

[83] Christopher Heil and David F Walnut. *Fundamental Papers in Wavelet Theory*. Princeton University Press, 2006.

[84] Eugenio Hernández and Guido Weiss. *A first course on wavelets*. CRC press, 1996.

[85] Frank G Holz, Erich C Strauss, Steffen Schmitz-Valckenberg, and Menno van Lookeren Campagne. Geographic atrophy: clinical features and potential therapeutic approaches. *Ophthalmology*, 121(5):1079–1091, 2014.

[86] Arnaldo Horta, Neil Joshi, Michael Pekala, Katia D Pacheco, Jun Kong, Neil Bressler, David E Freund, and Philippe Burlina. A hybrid approach for incorporating deep visual features and side channel information with applications to amd detection. In *Machine Learning and Applications (ICMLA), 2017 16th IEEE International Conference on*, pages 716–720. IEEE, 2017.

[87] Robert Houska. The nonexistence of shearlet scaling functions. *Applied and computational harmonic analysis*, 32(1):28–44, 2012.

[88] Gao Huang, Zhuang Liu, Kilian Q Weinberger, and Laurens van der Maaten. Densely connected convolutional networks. *arXiv preprint arXiv:1608.06993*, 2016.

[89] Glenn J Jaffe, Daniel F Martin, Cynthia A Toth, Ebenezer Daniel, Maureen G Maguire, Gui-Shuang Ying, Juan E Grunwald, Jiayan Huang, Comparison of Age-related Macular Degeneration Treatments Trials Research Group, et al. Macular morphology and visual acuity in the comparison of age-related macular degeneration treatments trials. *Ophthalmology*, 120(9):1860–1870, 2013.

[90] Magdalena Jasionowska and Artur Przelaskowski. Wavelet-like selective representations of multidirectional structures: a mammography case. *Pattern Analysis and Applications*, pages 1–10, 2018.

[91] Simon Jégou, Michal Drozdzal, David Vazquez, Adriana Romero, and Yoshua Bengio. The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*, pages 1175–1183. IEEE, 2017.

[92] Radford Juang, Elliot R McVeigh, Beatrice Hoffmann, David Yuh, and Philippe Burlina. Automatic segmentation of the left-ventricular cavity and atrium in 3d ultrasound using graph cuts and the radial symmetry transform. In *Biomedical Imaging: From Nano to Macro, 2011 IEEE International Symposium on*, pages 606–609. IEEE, 2011.

[93] Emily King. *Wavelet and frame theory: frame bound gaps, generalized shearlets, Grassmannian fusion frames, and $P-adic$ wavelets*. PhD thesis, University of Maryland, College Park, 2018.

[94] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. URL `http://arxiv.org/abs/1412.6980`.

[95] Ronald Klein and Barbara EK Klein. The prevalence of age-related eye diseases and visual impairment in aging: Current estimates. *Investigative ophthalmology & visual science*, 54(14), 2013.

[96] Ilya A Krishtal, Benjamin D Robinson, Guido L Weiss, and Edward N Wilson. Some simple Haar-type wavelets in higher dimensions. *The Journal of Geometric Analysis*, 17(1):87–96, 2007.

[97] Jens Krommweh. An orthonormal basis of directional haar wavelets on triangles. *Results in Mathematics*, 53(3-4):323–331, 2009.

[98] Jens Krommweh and Gerlind Plonka. Directional Haar wavelet frames on triangles. *Applied and Computational Harmonic Analysis*, 27(2):215–234, 2009.

[99] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*, 2016.

[100] Alexey Kurakin, Ian Goodfellow, Samy Bengio, Yinpeng Dong, Fangzhou Liao, Ming Liang, Tianyu Pang, Jun Zhu, Xiaolin Hu, Cihang Xie, et al. Adversarial attacks and defences competition. *arXiv preprint arXiv:1804.00097*, 2018.

[101] Gitta Kutyniok and Demetrio Labate. Introduction to shearlets. In *Shearlets*, pages 1–38. Springer, 2012.

[102] Andrew Lang, Aaron Carass, Matthew Hauser, Elias S Sotirchos, Peter A Calabresi, Howard S Ying, and Jerry L Prince. Retinal layer segmentation of macular OCT images using boundary classification. *Biomedical optics express*, 4(7):1133–1152, 2013.

[103] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86 (11):2278–2324, 1998.

[104] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 2015. URL `http://dx.doi.org/10.1038/nature14539`.

[105] Cecilia S Lee, Doug M Baughman, and Aaron Y Lee. Deep learning is effective for classifying normal versus age-related macular degeneration OCT images. *Ophthalmology Retina*, 1(4):322–327, 2017.

[106] Cecilia S Lee, Ariel J Tyring, Nicolaas P Deruyter, Yue Wu, Ariel Rokem, and Aaron Y Lee. Deep-learning based, automated segmentation of macular edema in optical coherence tomography. *bioRxiv*, page 135640, 2017.

[107] Kwon Lee, Michael Abramoff, Mona Garvin, and Milan Sonka. The Iowa reference algorithms (retinal image analysis lab, Iowa institute for biomedical imaging, IA), 2014.

[108] Weilin Li. *Topics in Harmonic Analysis, Sparse Representations, and Data Analysis*. PhD thesis, University of Maryland, College Park, 2018.

[109] Yiran Li. *Feature extraction in image processing and deep learning*. PhD thesis, University of Maryland, College Park, 2018.

[110] Anat London, Inbal Benhar, and Michal Schwartz. The retina as a window to the brain: from eye research to CNS disorders. *Nature Reviews Neurology*, 9 (1):44–53, 2013.

[111] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.

[112] Jiajun Lu, Hussein Sibai, Evan Fabry, and David Forsyth. No need to worry about adversarial examples in object detection in autonomous vehicles. *arXiv preprint arXiv:1707.03501*, 2017.

[113] Stéphane Mallat. *A wavelet tour of signal processing*. Academic press, 1999.

[114] Stéphane Mallat. Group invariant scattering. *Communications on Pure and Applied Mathematics*, 65(10):1331–1398, 2012.

[115] Benjamin Manning. *Composite multi resolution analysis wavelets*. PhD thesis, Washington University, 2012.

[116] Volodymyr Mnih and Geoffrey E Hinton. Learning to detect roads in high-resolution aerial images. In *European Conference on Computer Vision*. Springer, 2010.

[117] James Michael Murphy. *Anisotropic Harmonic Analysis and Integration of Remotely Sensed Data*. PhD thesis, 2015.

[118] NIH. Facts about age-related macular degeneration. `https://nei.nih.gov/health/maculardegen/armd_facts`, 2018. Accessed: September 2018.

[119] Edouard Oyallon and Stéphane Mallat. Deep roto-translation scattering for object classification. *arXiv preprint arXiv:1412.8659*, 2014.

[120] Edouard Oyallon, Eugene Belilovsky, and Sergey Zagoruyko. Scaling the scattering transform: Deep hybrid networks. In *International Conference on Computer Vision (ICCV)*, 2017.

[121] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against deep learning systems using adversarial examples. *arXiv preprint*, 2016.

[122] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *Security and Privacy (EuroS&P), 2016 IEEE European Symposium on*, pages 372–387. IEEE, 2016.

[123] Karas Pavel and Svoboda David. Algorithms for efficient computation of convolution. In *Design and Architectures for Digital Signal Processing*. InTech, 2013.

[124] Michael Pekala, Neil Joshi, David Freund, Neil Bressler, Delia Cabrera DeBuc, and Philippe Burlina. Automated OCT segmentation and comparison with human/machines. *5th MICCAI Workshop on Ophthalmic Medical Image Analysis (COMPAY/OMIA5) 2018*, 2018.

[125] Mike Pekala, Neil Joshi, David E Freund, Neil M Bressler, Delia Cabrera DeBuc, and Philippe M Burlina. Deep learning based retinal OCT segmentation. *arXiv preprint arXiv:1801.09749*, 2018.

[126] Gwenol Quellec, Katia Charrire, Yassine Boudi, Batrice Cochener, and Mathieu Lamard. Deep image mining for diabetic retinopathy screening. *Medical Image Analysis*, 39:178 – 193, 2017. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2017.04.012.

[127] Manzoor A Qureshi and Khalida Laghari. Role of B-scan ultrasonography in pre-operative cataract patients. *International journal of health sciences*, 4(1): 31, 2010.

[128] Fariz Rahman. keras-contrib. `https://github.com/keras-team/keras-contrib/blob/master/keras_contrib/applications/densenet.py`, 2018.

[129] Carl Edward Rasmussen and Christopher KI Williams. *Gaussian processes for machine learning*, volume 1. MIT press Cambridge, 2006.

[130] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015.

[131] Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. Dynamic routing between capsules. In *Advances in Neural Information Processing Systems*, pages 3856–3866, 2017.

[132] Steven L Salzberg. On comparing classifiers: Pitfalls to avoid and a recommended approach. *Data mining and knowledge discovery*, 1(3):317–328, 1997.

[133] Robert A Schowengerdt. *Remote sensing: models and methods for image processing*. Elsevier, 2006.

[134] Johanna M Seddon, Robyn Reynolds, Julian Maller, Jesen A Fagerness, Mark J Daly, and Bernard Rosner. Prediction model for prevalence and incidence of advanced age-related macular degeneration based on genetic, demographic, and environmental variables. *Investigative ophthalmology & visual science*, 50(5):2044–2053, 2009.

[135] Pierre Sermanet, David Eigen, Xiang Zhang, Michael Mathieu, Rob Fergus, and Yann LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. In *International Conference on Learning Representations (ICLR2014)*. CBLS, April 2014.

[136] Laurent Sifre and Stéphane Mallat. Rotation, scaling and deformation invariant scattering for texture discrimination. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 1233–1240. IEEE, 2013.

[137] Laurent Sifre, Michel Kapoko, Edouard Oyallon, and Vincent Lostanlen. Scatnet: a MATLAB toolbox for scattering networks. 2013.

[138] Gilbert Strang and Truong Nguyen. *Wavelets and filter banks*. SIAM, 1996.

[139] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[140] Jing Tian, Boglárka Varga, Gábor Márk Somfai, Wen-Hsiang Lee, William E Smiddy, and Delia Cabrera DeBuc. Real-time automatic segmentation of optical coherence tomography volume data of the macular region. *PloS one*, 10(8):e0133908, 2015.

[141] Jing Tian, Boglarka Varga, Erika Tatrai, Palya Fanni, Gabor Mark Somfai, William E Smiddy, and Delia Cabrera DeBuc. Performance evaluation of automated segmentation software on optical coherence tomography volume data. *Journal of biophotonics*, 9(5):478–489, 2016.

[142] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[143] Emanuele Trucco, Alfredo Ruggeri, Thomas Karnowski, Luca Giancardo, Edward Chaum, Jean Pierre Hubschman, Bashir Al-Diri, Carol Y Cheung, Damon Wong, Michael Abramoff, et al. Validating retinal fundus image analysis algorithms: Issues and a proposal. *Investigative ophthalmology & visual science*, 54(5):3546–3559, 2013.

[144] Freerk G Venhuizen, Bram van Ginneken, Freekje van Asten, Mark JJP van Grinsven, Sascha Fauser, Carel B Hoyng, Thomas Theelen, and Clara I Sánchez. Automated staging of age-related macular degeneration using optical coherence tomographyautomated staging of AMD in OCT. *Investigative Ophthalmology & Visual Science*, 58(4):2318–2328, 2017.

[145] Daniel Weinberg. *Multiscale and directional representations of high-dimensional information content in remotely sensed data*. PhD thesis, University of Maryland, College Park, 2015.

[146] Thomas Wiatowski and Helmut Bölcskei. A mathematical theory of deep convolutional neural networks for feature extraction. *arXiv preprint arXiv:1512.06293*, 2015.

[147] Thomas Wiatowski and Helmut Bölcskei. A mathematical theory of deep convolutional neural networks for feature extraction. *IEEE Transactions on Information Theory*, 64(3):1845–1866, 2018.

[148] Thomas Wiatowski, Michael Tschannen, Aleksandar Stanic, Philipp Grohs, and Helmut Bölcskei. Discrete deep feature extraction: A theory and new architectures. In *International Conference on Machine Learning*, pages 2149–2158, 2016.

[149] Yilun Zhou and Kris Hauser. Incorporating side-channel information into convolutional neural networks for robotic tasks. *IEEE International Conference on Robotics and Automation (ICRA)*, 2017.

[150] Xiao Xiang Zhu, Devis Tuia, Lichao Mou, Gui-Song Xia, Liangpei Zhang, Feng Xu, and Friedrich Fraundorfer. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geoscience and Remote Sensing Magazine*, 5(4):8–36, 2017.