INAUGURAL – DISSERATION

zur Erlangung der Doktorwürde der

Gesamtfakultät für Mathematik, Ingenieur- und Naturwissenschaften

der Ruprecht – Karls – Universität, Heidelberg

vorgelegt von

Schellenberg, Melanie, M.Sc.

aus Pfullendorf, Deutschland

Tag der mündlichen Prüfung: _____

Learning tissue geometries for photoacoustic image analysis

Erstbetreuerin: Prof. Dr. Lena Maier-Hein
Zweitbetreuer: Prof. Dr. Jürgen Hesser

Learning tissue geometries for photoacoustic image analysis

Photoacoustic imaging (PAI) holds great promise as a novel, non-ionizing imaging modality, allowing insight into both morphological and physiological tissue properties, which are of particular importance in the diagnostics and therapy of various diseases, such as cancer and cardiovascular diseases. However, the estimation of physiological tissue properties with PAI requires the solution of two inverse problems, one of which, in particular, presents challenges in the form of inherent high dimensionality, potential ill-posedness, and non-linearity. Deep learning (DL) approaches show great potential to address these challenges but typically rely on simulated training data providing ground truth labels, as there are no gold standard methods to infer physiological properties in vivo. The current domain gap between simulated and real photoacoustic (PA) images results in poor in vivo performance and a lack of reliability of models trained with simulated data. Consequently, the estimates of these models occasionally fail to match clinical expectations.

The work conducted within the scope of this thesis aimed to improve the applicability of DL approaches to PAI-based tissue parameter estimation by systematically exploring novel data-driven methods to enhance the realism of PA simulations (learning-to-simulate). This thesis is part of a larger research effort, where different factors contributing to PA image formation are disentangled and individually approached with data-driven methods. The specific research focus was placed on generating tissue geometries covering a variety of different tissue types and morphologies, which represent a key component in most PA simulation approaches. Based on in vivo PA measurements (N = 288) obtained in a healthy volunteer study, three data-driven methods were investigated leveraging (1) semantic segmentation, (2) Generative Adversarial Networks (GANs), and (3) scene graphs that encode prior knowledge about the general tissue composition of an image, respectively.

The feasibility of all three approaches was successfully demonstrated. First, as a basis for the more advanced approaches, it was shown that tissue geometries can be automatically extracted from PA images through the use of semantic segmentation with two types of discriminative networks and supervised training with manual reference annotations. While this method may replace manual annotation in the future, it does not allow the generation of any number of tis-

sue geometries. In contrast, the GAN-based approach constitutes a generative model that allows the generation of new tissue geometries that closely follow the training data distribution. The plausibility of the generated geometries was successfully demonstrated in a comparative assessment of the performance of a downstream quantification task. A generative model based on scene graphs was developed to gain a deeper understanding of important underlying geometric quantities. Unlike the GAN-based approach, it incorporates prior knowledge about the hierarchical composition of the modeled scene. However, it allowed the generation of plausible tissue geometries and, in parallel, the explicit matching of the distributions of the generated and the target geometric quantities. The training was performed either in analogy to the GAN approach, with target reference annotations, or directly with target PA images, circumventing the need for annotations. While this approach has so far been exclusively conducted in silico, its inherent versatility presents a compelling prospect for the generation of tissue geometries with in vivo reference PA images. In summary, each of the three approaches for generating tissue geometry exhibits distinct strengths and limitations, making their suitability contingent upon the specific application at hand.

By opening a new research direction in the form of learning-to-simulate approaches and significantly improving the realistic modeling of tissue geometries and, thus, ultimately, PA simulations, this work lays a crucial foundation for the future use of DL-based quantitative PAI in the clinical setting.

Erlernen von Gewebegeometrien für die photoakustische Bildanalyse

Die photoakustische Tomographie (PAT) ist ein vielversprechendes, aufstrebendes, nichtionisierendes bildgebendes Verfahren, das sowohl Einblicke in morphologische als auch physiologische Gewebeeigenschaften ermöglicht. Diese Eigenschaften sind von besonderer Bedeutung für die Diagnose und Therapie verschiedener Krankheiten, wie zum Beispiel Krebs und Herz-Kreislauf-Erkrankungen. Die Verwendung von PAT zur Schätzung physiologischer Gewebeeigenschaften erfordert jedoch die Lösung zweier inverser Probleme, von denen vor allem eines aufgrund seiner inhärenten hohen Dimensionalität, potenziell schlechten Problemstellung und Nichtlinearität Herausforderungen darstellt. "Deep Learning" (DL) Ansätze haben ein großes Potenzial, die genannten Herausforderungen zu bewältigen. Jedoch sind diese in der Regel auf simulierte Trainingsdaten angewiesen, die die Grundwahrheiten liefern, da bisher keine Goldstandard-Methoden für die in vivo Bestimmung physiologischer Eigenschaften existieren. Die momentane Diskrepanz zwischen simulierten und gemessenen photoakustischen (PA) Bildern führt bei Modellen, die auf simulierten Daten trainiert wurden, zu einer schwachen Leistung bei in vivo Anwendungen und einer mangelnden Zuverlässigkeit. Dies führt dazu, dass die Schätzungen dieser Modelle gelegentlich klinischen Erwartungen nicht entsprechen. Die im Rahmen dieser Arbeit durchgeführten Herangehensweisen zielten darauf ab, die Anwendbarkeit von DL-Ansätzen zur PAT-basierten Schätzung von Gewebeparametern zu verbessern, indem neuartige, datengetriebene Methoden zur Verbesserung der Realitätsnähe von PA-Simulationen ("learning-to-simulate") systematisch erforscht wurden. Diese Arbeit ist Teil eines größeren Forschungsvorhabens, bei dem verschiedene Faktoren, die zur PA-Bildentstehung beitragen, entschlüsselt und mithilfe von datengetriebenen Methoden einzeln angegangen werden. Der besondere Forschungsschwerpunkt lag auf der Erzeugung von Gewebegeometrien. Diese decken eine Vielzahl verschiedener Gewebetypen und -morphologien ab und stellen eine Schlüsselkomponente der meisten PA-Simulationsansätze dar. Basierend auf in vivo PA-Messungen (N = 288), die in einer Studie mit gesunden Probanden gewonnen wurden, konnten drei datengetriebene Methoden entwickelt werden, die jeweils eines der folgenden Prinzipen ausnutzen: (1) semantische Segmentierung, (2) "Generative Adversarial Networks" (GANs) und (3) Szenegraphen, die Vorwissen über die allgemeine Gewebezusammensetzung

eines Bildes kodieren.

Die Machbarkeit aller drei Ansätze wurde erfolgreich demonstriert. Zunächst wurde als Grundlage für die fortgeschritteneren Ansätze gezeigt, dass Gewebegeometrien durch semantische Segmentierung mit zwei Arten diskriminativer Netzwerke und überwachtem Training mit manuellen Referenzannotationen automatisch aus PA-Bildern extrahiert werden können. Diese Methode könnte in Zukunft ein Ersatz für die manuelle Annotation sein. Die Erzeugung einer beliebigen Anzahl neuer Gewebegeometrien ist damit jedoch nicht möglich. Die GAN-basierte Methode stellt im Gegensatz dazu ein generatives Modell dar. Damit können neue Gewebegeometrien erzeugt werden, die der Verteilung der Trainingsdaten eng folgen. Die Plausibilität der generierten Geometrien wurde in einem Vergleich der Leistung einer nachgelagerten Quantifizierungsaufgabe erfolgreich nachgewiesen. Ein auf Szenengraphen basierender Ansatz wurde entwickelt, um ein tieferes Verständnis wichtiger zugrunde liegender geometrischer Größen zu gewinnen. Im Gegensatz zum GAN-basierten Ansatz erfordert dieser Ansatz Vorwissen über die hierarchische Zusammensetzung der modellierten Szene. Er ermöglichte jedoch die Generierung plausibler Gewebegeometrien und parallel dazu den expliziten Abgleich der Verteilungen der generierten und der gegebenen geometrischen Größen. Das Training erfolgte entweder analog zum GAN-Ansatz mit Ziel-Referenzannotationen oder direkt mit PA-Bildern, wodurch der Bedarf an Referenzannotationen umgangen wurde. Während dieser Ansatz bisher ausschließlich in silico durchgeführt wurde, stellt seine inhärente Vielseitigkeit eine hervorragende Möglichkeit zur Optimierung anhand von in vivo PA-Referenzbildern dar. Abschließend bleibt festzuhalten, dass jeder der drei Ansätze zur Generierung von Gewebegeometrien unterschiedliche Stärken und Grenzen aufweist, sodass ihre Eignung von der jeweiligen Anwendung abhängt.

Durch die Eröffnung einer neuen Forschungsrichtung in Form von ”learning-to-simulate”-Ansätzen und die deutliche Verbesserung der realistischen Modellierung von Gewebegeometrien und damit letztlich der PA-Simulationen legt diese Arbeit eine entscheidende Grundlage für den zukünftigen Einsatz DL-basierter quantitativer PAT im klinischen Umfeld.

Contents

Acknowledgements

An dieser Stelle möchte mich bei einigen Menschen bedanken, die mich bei dieser Promotion kräftig unterstützt haben.

Zuallererst möchte ich mich bei Lena Maier-Hein bedanken, die mir diese großartige Chance gegeben hat, in ihrem Team Intelligente Medizinische Systeme (IMSY) am Deutschen Krebsforschungszentrum (DKFZ) zu promovieren. Vielen Dank Lena für deine Unterstützung in vielerlei Aspekten. Bei fachspezifischen Fragen und Problemen warst du immer eine sehr große Hilfe mit wertvollen Anregungen und Feedback, das Gold wert ist. Zudem hast du eine Begeisterung für die Forschung, die abfärbt. Während meiner Zeit bei IMSY durfte ich viele Einblicke in die Welt der Forschung gewinnen, mein Fachwissen immens erweitern und Einiges fürs Leben lernen. Ich möchte mich ganz herzlich für die hervorragende Betreuung bedanken.

Des Weiteren geht ein großer Dank an meinen Zweitbetreuer Jürgen Hesser, der mich mit lebendigen Diskussionen und Anregungen jederzeit unterstützte.

Vielen Dank Lena und Jürgen auch für die Begutachtung meiner Thesis.

Ein weiteres Danke geht an das ganze IMSY Team und die Mädels vom Office. Obwohl wir so vielseitig, mittlerweile doch ganz schön groß, und meist gut beschäftigt sind, war es nie schwer, jemanden zu finden, der mich mit Rat und Tat unterstützte. Das ist meiner Meinung nach einzigartig und ich bin sehr froh meine Promotion als Teil dieses großartigen Teams durchgeführt haben zu dürfen und sage Danke. Der Spaßfaktor kam dank euch während meiner Promotion auch nicht zu kurz. Vielen Dank für die lustigen Momente, die ich mit euch beispielsweise auf den Retreats, der Retreatplanung mit dir Annika Reinke, den Weihnachtsfeiern, den gemeinsamen Mittagspausen, auf dem Weg zu sowie von den Meetings, und auch während den Meetings haben durfte.

Eine kleine Liebeserklärung geht an das Photoakustik Team mit allen momentanen und ehemaligen Mitgliedern: Kris K. Dreher, Jan-Hinrich Nölke, Niklas Holzwarth, Alexander Seitel, Tom Rix, Marcel Knopp, Christoph Bender, Damjan Kalšan, Fabian Schneider, Janek Gröhl,

Thomas Kirchner, Andrei Cosmin Siea und Patricia Vieten. Ihr seid alle super Menschen, meiner Meinung nach sehr fleißig, und es macht einfach extrem viel Spaß mit euch zu arbeiten. Ich wurde am Anfang meiner Promotion gerade von Janek und Thomas sehr willkommen geheißen und hab mich ab dann immer sehr wohl gefühlt. Vielen Dank für die tolle Unterstützung, die ich von euch in jeglicher Hinsicht erhalten habe, wie beispielsweise das viele wertvolle Feedback, das gemeinsame Brainstorming und die Hilfe bei den Simulationen. Ich erinnere mich gerne an die vielen Erlebnisse, wie die "mal kurzen" Diskussionen im Büro, die Bier-um-vier Zeiten, die (leider einzige) Konferenz mit Kris, Jan, und Niklas in San Franzisco, das iThera-Meeting mit Janek und Thomas in Essen, und vieles mehr. Gerade bei Janek, Kris, Jan, und Niklas möchte ich mich auch für das Zuhören und den Zuspruch bei den "Downs", die es im Laufe einer Promotion doch manchmal gibt, herzlichst bedanken.

Ein besonderes Dankeschön geht an meine Familie und meine Freunde, ohne die diese Arbeit nicht möglich gewesen wäre. Mama und Papa, vielen Dank dafür, dass ihr immer hinter mir steht und mir das Studium, ohne das ich hier nicht wäre, ermöglicht habt. Verena und Kerstin möchte ich danken, dass sie mich gerade in stressigen Zeiten immer gut "relativieren".

Meinen Studienfreunden Vanessa, Alina, und Fuchsi will ich Danke sagen, fürs durch Dick und Dünn gehen. Ich glaube ihr wisst, dass ich ohne euch nicht hier stehen würde. Bei Till und Dat möchte ich mich für das aufs Trapp halten bedanken. Für die immer gute Laune, die überträgt, geht ein Danke an Fabian und Klaus. Zudem bin ich sehr dankbar für Ines, Nadine und Jens, die ich im Laufe der Promotion ins Herzchen geschlossen habe und für lustige Urlaube nicht mehr wegzudenken sind. Und nicht zu vergessen sind meine Mädels von daheim: Christl, Jaci, Anni, und Anika, die einfach immer da sind. Vielen Dank, ihr seid alle Freunde fürs Leben und wart in der ganzen Zeit immens wichtig!

Zuletzt gilt ein ganz besonderer Dank Sebi - there are no words.

Abbreviations

BOLD	Blood-Oxygen-Level-Dependent
BVF	Blood Volume Fraction
CE	Cross-Entropy
CEST	Chemical Exchange Saturation Transfer
CNN	Convolutional Neural Network
CSI	Chemical Shift Imaging
CT	Computed Tomography
DL	Deep Learning
DSC	Dice Similarity Coefficient
FCNN	Fully-Connected Neural Network
FFT	Fast Fourier Transformation
FNO	Fourier Neural Operator
GAN	Generative Adversarial Network
GB	Gigabyte
GELU	Gaussian Error Linear Unit
GMMN	Generative Moment Matching Network
GNN	Graph Neural Network
GPU	Graphics Processing Unit
GT	Ground Truth
LeakyReLU	Leaky Rectified Linear Unit
MC	Monte Carlo
MITK	Medical Imaging Interaction Toolkit

ML	Machine Learning
MMD	Maximum Mean Discrepancy
MRI	Magnetic Resonance Imaging
MSE	Mean Squarred Error
MSOT	Multi-Spectral Optoacoustic Tomography
NSD	Normalized Surface Distance
PA	Photoacoustic
PAI	Photoacoustic Imaging
PAUS	a combination of PA and US
PDE	Partial Differential Equation
PET	Positron Emission Tomography
qPAI	quantitative Photoacoustic Imaging
RAM	Random Access Memory
ReLU	Rectified Linear Unit
RQ	Research Question
RTE	Radiative Transfer Equation
SIMPA	Simulation and Image Processing for Photonics and Acoustics
SNR	Signal-to-Noise Ratio
sO_2	Oxygen Saturation
SSIM	Structural Similarity Index Measure
TanH	Tangens Hyperbolicus
US	Ultrasound
VAE	Variational Autoencoder

List of Figures

List of Tables

List of Algorithms

1. Introduction

In this chapter, the relevance of learning tissue geometries for the analysis of Photoacoustic (PA) images is motivated both clinically (cf. Section 1.1) and technically (cf. Section 1.2). Furthermore, the overall approach is presented together with the objectives defined in this thesis (cf. Section 1.3).

1.1. Clinical Motivation

Molecular imaging is of tremendous importance in medical research due to its ability to non-invasively visualize, characterize, and quantify biological processes in vivo at cellular and molecular levels [Schober et al., 2020]. This means that, unlike conventional imaging techniques that purely represent anatomy, molecular imaging techniques can complement the understanding of various diseases [Cassidy et al., 2005]. They allow the study of fundamental biological processes and, therefore, the direct detection of molecular abnormalities rather than the consequence of an accumulation of multiple alterations [Schober et al., 2020]. This holds enormous potential for precision medicine, particularly in detecting and treating cancer, neurological, and cardiovascular diseases.

Various technologies exist that enable molecular imaging, such as contrast-agent enhanced Ultrasound (US) imaging, Chemical Shift Imaging (CSI), Chemical Exchange Saturation Transfer (CEST), Blood-Oxygen-Level-Dependent (BOLD) Magnetic Resonance Imaging (MRI), dynamic contrast-enhanced Computed Tomography (CT), and optical imaging, to name a few examples. They are based on different physical principles, providing complementary tissue information, each having different advantages and disadvantages.
A comparatively novel non-ionizing modality is Photoacoustic Imaging (PAI), which combines optical with US imaging and hence has the potential to deliver not only morphological but also physiological tissue properties in penetration depths of several centimeters with sub-millimeter spatial resolutions [Beard, 2011].

The image formation is based on the photoacoustic effect. Specifically, nanosecond laser pulses with wavelengths in the near-infrared range illuminate the tissue and penetrate a few centimeters deep due to the overall high scattering of the tissue. Different molecules, referred to as chromophores, such as endogenous oxy- and deoxyhemoglobin, melanin, lipids, and water, or exogenous injected contrast agents, absorb the laser energy following their characteristic wavelength-dependent absorption spectra and convert it into heat, causing a brief temperature rise in the surrounding tissue, typically less than 0.1 K. This, in turn, leads to a thermo-elastic expansion, resulting in a local pressure increase, typically less than 10 kPa. The resulting initial pressure rise generates low-amplitude broadband acoustic waves covering a range of frequencies between 0.1 and 100 MHz. These waves travel through the tissue and can be detected as time series data through the piezoelectric effect at the surface of the tissue.
In other words, PAI detects sound waves generated solely by the initial pressure rise that is proportional to the absorption of near-infrared light from various chromophores and the light fluence, which, in turn, depends on absorbers and scatterers in the tissue (light in - sound out principle).
Thus, the source of image contrast is fundamentally different from US imaging, where the detected signals are also acoustic time series. In US imaging, it is mainly the reflections of transmitted sound waves at tissue interfaces that are received by the transducer, depending on impedance changes, and hence mechanical and elastic properties of the tissue are provided (sound in - sound out principle). The image contrast of PAI is more similar to optical imaging techniques based on light-tissue interactions, such as diffuse optical imaging, or classical ballistic optical imaging, such as optical coherence tomography. In comparison to these techniques, PAI generally offers high resolution and, at the same time, large penetration depths [Wang et al., 2012]. Compared to typical imaging modalities such as MRI and CT, PAI does not offer a comparably large field of view but is generally more affordable and might be easier to include in clinical practice, especially for imaging during surgery.

PAI is of great interest in various clinical applications, especially due to its unique potential to estimate chromophore concentrations and derivable physiological properties in a spatially resolved manner. In fact, there is no reliable and non-invasive gold standard method for this purpose, highlighting the great potential of PAI.
The clinical relevance is primarily linked to the key hallmark of PAI, the high intrinsic vascular contrast. This contrast is based on the absorption of oxy- and deoxyhemoglobin in blood and allows the inference of associated physiological properties such as total blood volume and Oxygen Saturation (sO_2), defined as the ratio of the signal contributions of oxy- and deoxyhemoglobin. For example, these properties are particularly relevant for the detection and

therapy response monitoring of hallmarks of cancer, such as angiogenesis and hypoxia [Laufer et al., 2012, Lin et al., 2022, Mallidi et al., 2011]. In addition, cardiovascular diseases such as venous thrombosis or atheromatous arterial stenosis [Chlis et al., 2020] can be monitored. Hemodynamic changes and sO_2 estimations can also stage inflammatory diseases such as Crohn's disease [Knieling et al., 2017, Gröhl et al., 2021b].

1.2. Technical Challenges

However, the derivation of clinically relevant biomarkers from recorded PA time series data is non-trivial, as it involves the solution of two inverse problems.
One first needs to solve the *acoustic inverse problem*, which involves the reconstruction of the acoustic time series data into a PA image (initial pressure distribution), which can then be analyzed in detail in a second step to obtain morphological or physiological tissue information. This inversion depends on acoustic parameters in the tissue, such as speed of sound and acoustic attenuation, which are typically unknown. Therefore, the reconstruction is usually approximated by conventional US-specific reconstruction algorithms [Matrone et al., 2014, Xu et al., 2002, Xu et al., 2005, Gröhl et al., 2021b], which make assumptions about these parameters. Moreover, in some settings, the inversion can be ill-posed and hampered by various detector-specific factors. For example, typical handheld PA detectors often have a limited detection bandwidth and a limited view, resulting in undersampling [Gröhl et al., 2021b].
For the quantification of physiological parameters, the optical image formation process requires inversion. In general, the goal is to determine the spatially resolved spectral behavior of the absorption, which, together with known literature spectra, allows the unmixing of the concentration of different chromophores and the derivation of physiological parameters. However, this so-called *optical inverse problem* is non-linear and ill-posed. The reason for the non-linearity is that the initial pressure distribution not only depends on the absorption but also on the light fluence, which in turn depends on the optical parameters in the tissue, such as absorption and scattering. The ill-posedness is based on the fact that there might be ambiguous solutions of underlying optical properties to the same initial pressure distribution.

Another hurdle in quantitative Photoacoustic Imaging (qPAI) is the lack of established gold standard methods that reliably provide Ground Truth (GT) references about chromophore concentrations or physiological tissue properties in vivo. This poses a *chicken-and-egg* dilemma in developing and validating a novel PA-based reference quantification method. Numerous attempts have been reported to solve the optical inverse problem of PAI using various techniques

ranging from model-based approaches under different assumptions [Cox et al., 2006, Cox et al., 2007, Shao et al., 2011, Pulkkinen et al., 2014, Brochu et al., 2016] to fully data-driven methods [Gröhl et al., 2021b].
As shown in a recent literature review [Gröhl et al., 2021b], supervised Deep Learning (DL) methods have become an essential tool to tackle the quantification problem, among other challenges in PA image analysis. Unlike model-based approaches, these methods are typically fast and require little domain-specific prior knowledge. They usually rely on a large number of simulated PA data with known underlying parameters to cope with the lack of GT labels. Note that for some applications other than quantification, the training data can be obtained from real measurements. For example, semantically segmented reference annotations are often derived from PA images and an additional source of information, such as co-registered US images, which provide complementary structural tissue information.

The current main limitation preventing the success of DL-based quantification models is their poor generalizability of performance to real data. This phenomenon is referred to as a domain gap between the data distributions of the in silico training and the real test data.
Numerous factors contribute to the domain gap and require consideration for realistic image simulation, such as the tissue digital twin, the device digital twin, the photon propagation, the acoustic wave propagation, and the noise characteristics. Figure 1.2.1 shows the tissue and device digital twins that serve as input components to the simulation model. The digital twin of the tissue can be further divided into three classes that define the morphology of the different tissue types, further referred to as *tissue geometries*, as well as optical and acoustic parameters.

This work focuses on automatic modeling of realistic tissue geometries (as highlighted in pink in Figure 1.2.1). This step is essential to the realism of the simulations for two main reasons. First, the tissue geometries serve as the basis for all subsequent simulation steps. Second, the nonlinear light propagation in the tissue, and thus the distribution of image values, strongly depends on the tissue geometries. Consequently, realistic tissue geometries are crucial to match the data distributions of digital and real images for DL-based models.
The realistic modeling of these geometries also allows a profound understanding of the morphological tissue structures. This knowledge can be important for various PA image analyses. For example, it can support the analysis of clinically relevant measures for different diseases that are often related to morphological abnormalities. Furthermore, knowledge about tissue geometries can provide a clearer and more comprehensible presentation of quantitative assessments and reduce the dimensionality of inverse problems, for example, by incorporating local consistencies of underlying tissue properties specific to different morphological structures as regularization.

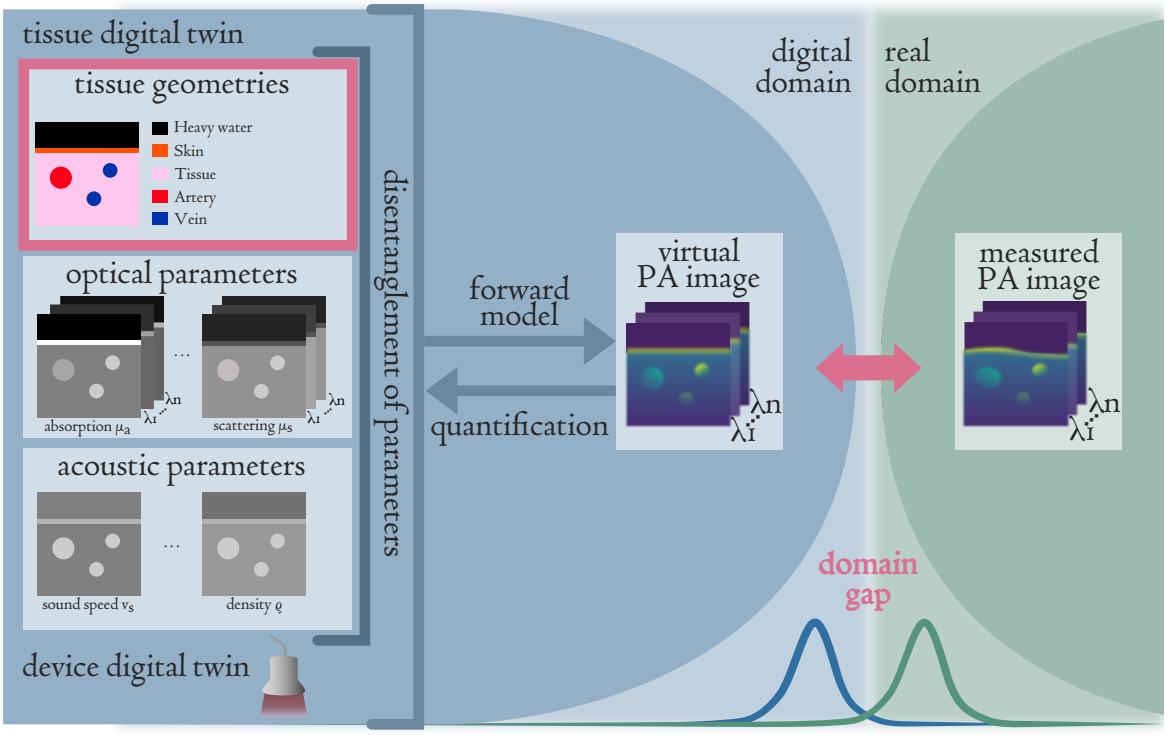

Figure 1.2.1.: Contribution of this thesis. To enable data-driven quantitative Photoacoustic Imaging (PAI), the domain gap between digital and real Photoacoustic (PA) images must be overcome. To this end, various steps of image simulation require careful and realistic modeling. This work proposes the disentanglement of the different factors contributing to image formation. In particular, the focus is on the data-driven generation of tissue geometries (highlighted in pink) to reduce the domain gap. Two factors contributing to image formation are shown: the tissue and device digital twins, which serve as inputs of a PA forward model. The tissue digital twin is further disentangled into three categories, namely tissue geometries that encode morphological information, optical (e.g., the absorption coefficient μ_a and scattering coefficient μ_s at wavelength λ) and acoustic (e.g., the sound speed v_s and the density ρ) tissue parameters, as shown in 2D.

Generally and in contrast to other research areas, such as computer vision or MRI and CT imaging, the research field of PAI has not yet focused on the automated generation of realistic PA images. Specific to tissue geometry generation, the following gaps in the literature have been identified that this thesis addressed:

1. Tissue geometries for PA simulations have been approached mainly by conventional methods. In particular, basic geometries, numerical pattern phantoms, simple model-based approaches, or manual image segmentation from other imaging modalities were primarily applied. However, these approaches usually result in simplified, often unrealistic, or poorly scalable tissue shapes and compositions.

2. A closely related topic, semantic segmentation, has previously been addressed in the PAI community using both traditional and DL-based methods. However, until the start of this thesis, there has been no work investigating DL for multi-label semantic segmentation of tomographic PA images.

3. DL-based tissue generation for PA images has not yet been addressed, for example, with generative neural networks that became popular and powerful in other research fields.

1.3. Approach and Objectives

This thesis is part of a larger European Research Council (ERC)-funded concept (Neural Spectral Image Decoding, Grant agreement ID: 101002198) that aims to advance the realism of digital PA images to enable data-driven qPAI in the long run. In this concept, the quantification is formulated as a DL-based decoding task, allowing for pixel-wise estimation of underlying physiological parameters or chromophore concentrations of PA images. The major innovation of the concept lies in the data-driven modeling of PA images, where the core idea is to disentangle and address separately the different components involved in the image formation that contribute to the domain gap (cf. Figure 1.2.1). In this way, the influence of single components on the quantification can be analyzed, and a learning-based optimization of each component can be performed. In a broader context, this concept contrasts with previous learning-from-simulations qPAI approaches, as PA image synthesis and quantification are considered as one joint framework, which is also referred to as a *learning-to-simulate* approach.

As part of this larger concept, this thesis aimed to automatically model realistic tissue geometries with DL. In this context, *tissue* refers to biological tissues, such as skin and muscle tissue. But also other biological and non-biological macroscale structures visible in PA images are included in this term, such as vessels and device-specific transducer fluid and transducer membrane. *Geometry* refers to these structures' number, shape, and position in 2D.

Three methods were investigated, each addressing one of the ensuing Research Questions (RQs). All of them followed one key principle, which was to leverage acquired PA images or patterns derived from them. As shown in Figure 1.3.1, the methods based their training and/or inference on real PA images, US images, and/or their corresponding manual reference annotations to generate realistic tissue geometries in a data-driven manner.

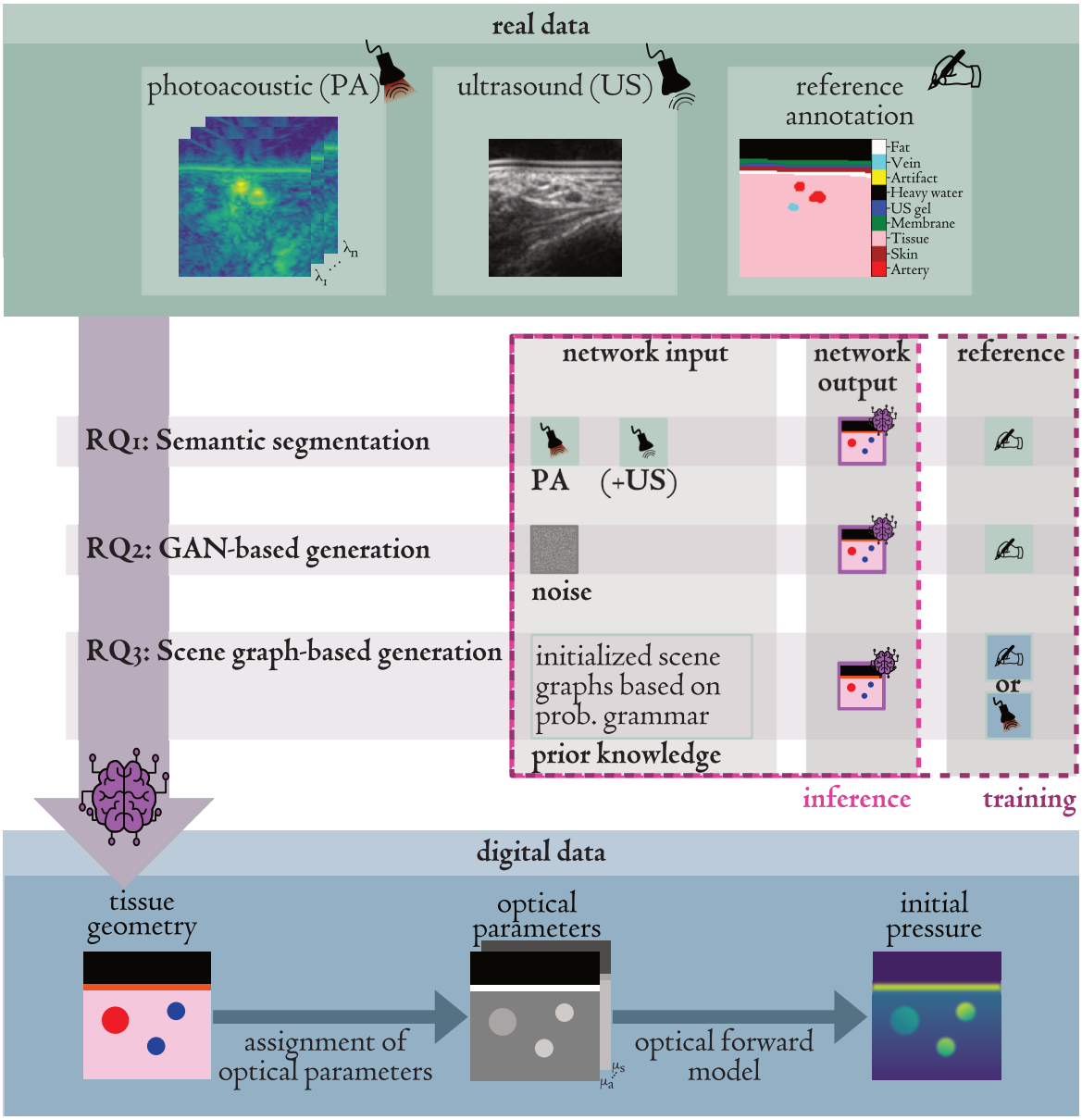

Figure 1.3.1.: The three approaches linked to the three Research Questions (RQs)(RQ1 - RQ3) to generate tissue geometries with Deep Learning (DL). The key principle is to leverage acquired Photoacoustic (PA) images, Ultrasound (US) images, and/or their associated manual annotations for network training (dashed purple box) and inference (dashed pink box), respectively. After the assignment of optical parameters to the tissue geometries, such as the absorption coefficient μ_a and scattering coefficient μ_s, initial pressure distributions can be simulated. Note that a green background color denotes real data, a purple one characterizes data-driven techniques, and a blue one represents virtual/digital data.

Research Questions

To investigate the feasibility of data-driven methods for realistic tissue geometry modeling, three RQs that build on each other were investigated:

Research Question RQ1:

Can discriminative neural networks be leveraged to extract tissue geometries from real PA images via automatic semantic segmentation?

RQ1 is considered the basis of the work and intended to fundamentally investigate whether tissue geometries can be automatically extracted from PA images. It involves estimating tissue geometries directly from multi-spectral PA images via semantic segmentation with supervised training and manual reference annotations. This question aims to show the feasibility of this approach, which is challenging given the limited number of available data and which has not been demonstrated before. Additionally, the potential added benefit of leveraging co-registered US images along with PA images shall be investigated.

Research Question RQ2:

Can Generative Adversarial Networks (GANs) be leveraged for the generation of plausible tissue geometries?

RQ2 is intended to go one step further than RQ1 and generate completely new tissue geometries. The question studies GANs in terms of augmenting a set of manual reference annotations and thus generating a required number of new tissue geometries. A challenge for the training of GANs is the limited number of available PA-based reference annotations. Furthermore, the validation of this approach is challenging since there is no GT information available for the generated geometries.

Research Question RQ3:

Can scene graphs be leveraged for the generation of plausible tissue geometries?

RQ3 is intended to provide an understanding of key geometric quantities in parallel to the generation of new tissue geometries by leveraging scene graphs that encode prior knowledge about tissue composition. In contrast to RQ2, this approach is more constrained by the encoded prior knowledge but offers the possibility to explicitly learn geometric quantities, such as the position of different tissue geometries. In this question, the goal is to generate geometries that either resemble a set of target reference annotations directly (analogous to RQ2) or target PA images after performing a forward simulation. The second case thus bypasses the need for reference annotations and goes beyond the previous methods. However, the inclusion of a differentiable simulation makes the optimization more complex. This concept has been shown to be feasible in the computer vision domain, but it is by no means clear whether it can be applied to PA images since the image properties are inherently different. For example, the PA image intensity decreases with image depth due to light absorption. In addition, there is the challenge of dealing with the limited amount of PA data available.

1.4. Outline

This thesis starts with an *Introduction* chapter 1 that explains the clinical and technical motivation for why learning tissue geometries is relevant for PA image analysis. The overall approach and objectives, along with the three research questions, are included at the end of this chapter.

Chapter 2 is about the *Fundamentals* relevant to the thesis. It is separated into a section on photoacoustic imaging and a section on deep learning. In the photoacoustic imaging section, the physical background is given on light-tissue interactions and the photoacoustic image formation. The deep learning section elaborates on key principles with regard to the optimization of deep learning approaches, supervised and unsupervised learning, as well as discriminative and generative models. An additional subsection explains the neural network types applied in this thesis.

The *Related Work* chapter 3 summarizes relevant literature related to the content of the thesis. First, pertinent approaches for semantic segmentation of medical and photoacoustic images are presented. Key principles and outstanding works for image simulation and synthesis in the fields of computer vision, medical, and photoacoustic imaging follow.

Chapter 4 covers the *Contributions* of the thesis. It consists of four sections. The first section explains the photoacoustic data, including its acquisition, processing, and annotation. The

following three sections each address one of the three research questions. First, the work on *Tissue Geometry Estimation with Neural Networks* (RQ1) is presented. Then, the approach for *Tissue Geometry Generation with Generative Adversarial Networks* (RQ2) follows. In the last section, the approach for *Tissue Geometry Generation with Scene Graphs* (RQ3) is covered. The structure for each of these three sections is the same. First, the concept overview of the approach is explained. Then, relevant material and methods are presented. The experiments and experimental conditions follow. The results are given in the subsequent subsection. The end of each of these sections is a discussion of the results.

A general *Discussion* is given in chapter 5. Its purpose is to consider and discuss the three approaches developed jointly.

The *Summary* of the thesis follows in chapter 6. It includes a summary of contributions and a final conclusion of the findings of the thesis. A list of publications authored during the time of this thesis can be found at the end of this chapter.

The final chapter 7 presents the *Supplemental Material*. Four sections are presented. The first one gives further details on the photoacoustic data. Additional results of the approach for *Tissue Geometry Estimation with Neural Networks* (RQ1) follow. The next section gives further insights into the approach for *Tissue Geometry Generation with Generative Adversarial Networks* (RQ2). Finally, the list of publications is given, along with the categorization of the respective approaches to tissue modeling that were used for the literature review.

2. Fundamentals

In this chapter, the basic principles relevant to this thesis are presented. First, in Section 2.1, the basic principles of light-tissue interactions and PA image formation are given. This is followed by a section on DL, including the key concepts of optimization, super- and unsupervised learning, discriminative and generative models, and neural network types (cf. Section 2.2).

2.1. Photoacoustic Imaging

PAI is an emerging imaging modality that belongs to the field of biophotonics. It combines the advantages of optical and US imaging, which makes PAI promising for various applications. In more detail, the contrast of optical imaging and the generally high spatial resolution deep in the tissue of US imaging are combined in PAI. This section summarizes the main principles relevant to this thesis. More specifically, fundamental principles of light-tissue interaction (cf. Section 2.1.1) and the PA image formation that is based on it (cf. Section 2.1.2) are presented. For a deeper understanding of the fundamentals, the reader is referred to the books by Wang et al., 2012 and by Keiser, 2016 as well as the website by Prahl et al., 2017. These sources also served as the basis for the following section.

2.1.1. Light-Tissue Interaction

Light-tissue interactions are the fundamental principles of biophotonics, which plays an important role in many fields. For example, it is indispensable in basic research of life sciences as well as in the diagnosis, therapy, and surgery of various diseases [Keiser, 2016]. In general, the concept of biophotonics is to use light in the range from mid-ultraviolet ($\sim$ 190 nm) to mid-infrared ($\sim$ 1060 nm) to obtain information about biological tissue. Light interacts with biomedical tissue in different ways and depending on its wavelength. In the typical wavelength

range of biophotonics, the dominating physical processes are reflection, refraction, absorption, and scattering [Keiser, 2016, Wang et al., 2012], as schematically shown in Figure 2.1.1.

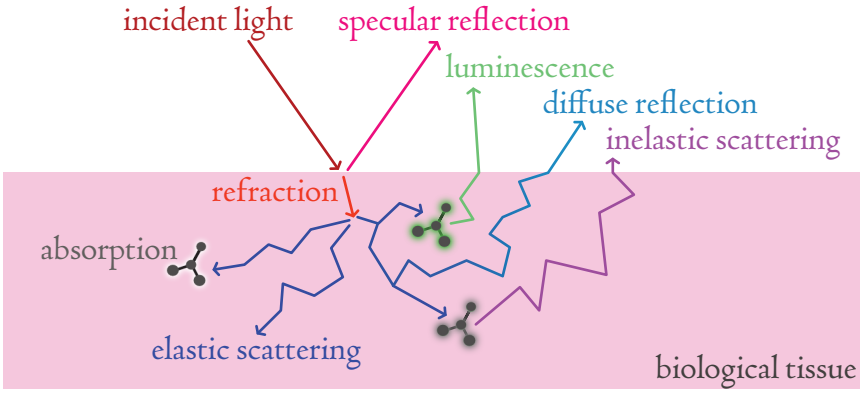

Figure 2.1.1.: Dominant light-tissue interactions in biophotonics. An incident light beam can get reflected at the surface between different tissue types or enter the tissue via refraction. Molecules in the tissue can interact with the light photon and elevate an electron from a ground state to an excited state, for example, via absorption. The excited electron can relax by nonradiative transitions, such as heat, or by emitting a photon, as in the case of luminescence. Virtual excitation states are related to elastic and inelastic scattering, such as Rayleigh and Raman scattering, respectively. Multiple scattering events can eventually lead to light escaping the tissue, which is referred to as diffuse reflectance [Wang et al., 2012].

Reflection describes the phenomenon in which an incident beam of light at the interface of two different dielectric tissue media is reflected back into the original medium. The process by which an incident beam of light is refracted and enters the second tissue medium is called refraction and depends on the angle between the interface and the incident beam. Reflection and refraction are a consequence of the different speeds of light in different media, defined by the material-specific refractive index. Absorption is a result of the interaction of light with molecules, which causes an electron to be lifted from a ground state to an excited state. In other words, the energy levels of a molecule are quantized, and a photon's energy can be converted into an electron transition from one energy level to a higher one. The electron relaxes to the ground state through a non-radiative transition, such as heat, or by emitting a photon, which is referred to as fluorescence and phosphorescence, dependent on the lifetime of the excited electron. Scattering occurs when the refractive index of a tissue structure differs from that of the surrounding tissue. In contrast to absorption, scattering excites the molecule into a very short-lived higher virtual state, which re-emits the photon during relaxation into a different direction [Wang et al., 2012].

In PAI, light in the visual and near-infrared spectral window ranging from ~ 400 nm to ~ 1300 nm is typically used because of the relatively large penetration depth of light in this range. The main light-tissue interactions in this window are scattering and absorption, with scattering dominating. Therefore, biological tissue in this context is also called turbid medium. The high level of scattering ensures that photons propagate diffusely through the tissue. It is thus very likely that they are absorbed at some point during scattering processes. Broadly speaking, the near-infrared window allows insights into tissue information linked to the absorption of light by various molecules, also known as chromophores, since scattering is overall less dependent on wavelengths than the absorption of different molecules. Note that absorption is the most relevant property in PAI since it allows the quantification of the concentration of different chromophores and, thus, the inference of physiological parameters.
The ensuing parts summarize the key physical principles of absorption, scattering, and anisotropy.

Absorption

According to the theory of quantum mechanics, a chromophore absorbs light only if the photon's energy matches the one needed to excite an electron from one of the discrete energy levels to another. The photon's energy E is dependent on its wavelength λ and can be calculated by:

$$E = \frac{hc}{\lambda}, \tag{2.1}$$

with Planck's constant $h = 6.63 \cdot 10^{-34}$ [J s] and the speed of light in vacuum $c = 3 \cdot 10^{8}$ [m s^{-1}]. There are different types of transitions between energy levels, which accordingly require different amounts of photon energy. A distinction is made between electronic transitions (potential energy) and vibrational, rotational, and translational transitions (kinetic energy). For photon energies corresponding to wavelengths in the near-infrared range, vibrational transitions are the most common.

The probability of a medium absorbing a photon per unit path length is defined by the absorption coefficient μ_a [Wang et al., 2012]. More specifically, it is defined as:

$$\mu_a = N_a \sigma_a = N_a Q_a \sigma_g, \tag{2.2}$$

with N_a being the number density of absorbers in a medium, σ_a the absorption cross-section, Q_a the absorption efficiency, and σ_g the geometric cross-sectional area. A typical value in biological tissue is $\mu_a = 0.1$ cm^{-1}, corresponding to an *absorption mean free path* of $\frac{1}{\mu_a} \sim 1$ cm.

Following the definition of the absorption coefficient, light attenuates with increasing distance in an absorbing tissue. The behavior of the light intensity with increasing path length x in an absorbing-only tissue is described by the Beer-Lambert law:

$$I(x) = I_o e^{-\mu_a x}, \tag{2.3}$$

where I_0 denotes the initial light intensity. In accordance, the probability of a photon not being absorbed in the tissue is defined as the internal transmittance T_i:

$$T_i(x) = \frac{I(x)}{I_0} = e^{-\mu_a x}. \tag{2.4}$$

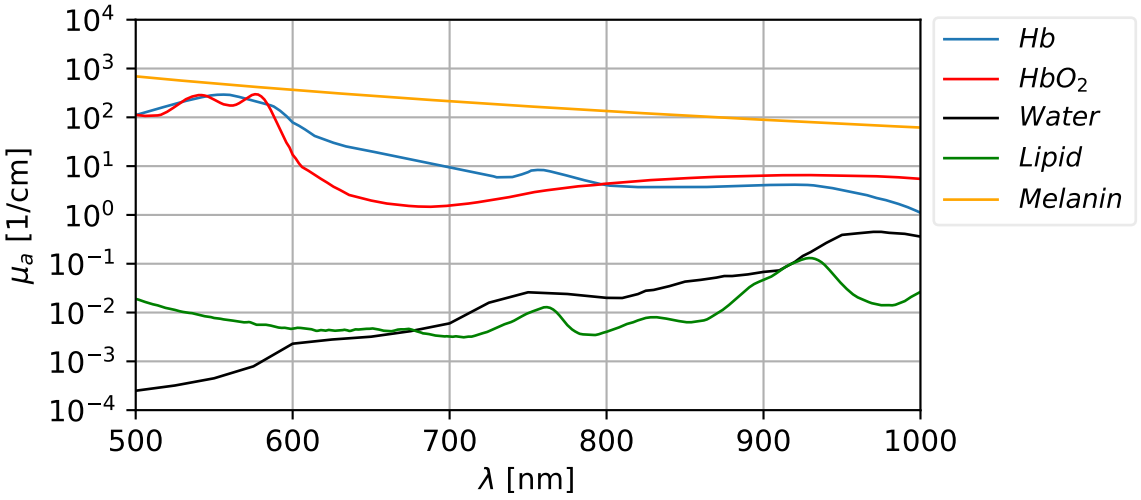

Figure 2.1.2.: Spectral behavior of the absorption coefficient (μ_a) of main chromophores in human biological tissue for optical and near-infrared wavelengths (λ). The spectral data was downloaded from Jacques, 2015. While the absorption spectra were directly available or explicitly described for water, lipid, and melanin, the spectra for hemoglobin (Hb) and oxyhemoglobin (HbO_2) were calculated from the molar extinction coefficients, assuming 150 g L^{-1} of Hb.

The main chromophores in biological tissue are melanin, hemoglobin (Hb), oxygenated hemoglobin, referred to as oxyhemoglobin (HbO_2), water, and lipid. Their absorption spectra are shown in Figure 2.1.2. Usually, several chromophores coexist in a region of interest in biological tissue. The total absorption coefficient of different chromophores is a linear combination of

the individual absorption coefficients weighted by the chromophores's concentration. More specifically, the total absorption coefficient can be written as:

$$\mu_a = \sum_i C_i \mu_{a,i}, \tag{2.5}$$

where C_i describes the concentrations of a specific chromophore i. To extract the concentration of different chromophores from a measured total absorption coefficient, a system of linear equations needs to be solved. Note that for the solution of this so-called *linear unmixing*, the number of measured wavelengths needs to be at least as high as the number of chromophores to be determined. For clinical applications, the concentrations of the chromophores Hb and HbO_2 are of particular interest since their ratio determines the important physiological property, *tissue oxygen saturation* sO_2. The tissue oxygen saturation is defined as:

$$sO_2 = \frac{C_{HbO_2}}{C_{HbO_2} + C_{Hb}}. \tag{2.6}$$

Scattering

Scattering describes the change of direction of a (straight) photon's trajectory after interacting with biological structures, such as cell membranes, lysosomes, mitochondria, and whole cells. In contrast to absorption, the scattering photon's energy excites the molecule into a very short-lived higher virtual state, which re-emits the photon during relaxation into a different direction [Wang et al., 2012]. There are two types of scattering, elastic and inelastic, which are defined according to whether the energy of the photon is conserved or not. In biological tissue, elastic scattering is dominating. It is strongest for structures similar (Mie theory) or smaller (Rayleigh theory) than the wavelength of the photon and whose refractive indices differ from the surrounding tissue.

In analogy to the absorption coefficient, the probability of a medium scattering a photon per unit path length is defined as the scattering coefficient:

$$\mu_s = N_s \sigma_s = N_s Q_s \sigma_g, \tag{2.7}$$

where N_s defines the number density of scatterers in a medium, σ_s the scattering cross-section, Q_s the scattering efficiency, and σ_g the geometric cross-sectional area. A typical value for biological tissue is μ_s = 100 cm^{-1}, correpsonding to the scattering mean free path of $\frac{1}{\mu_s} \sim 0.01$ cm.

In other words, the probability of a photon not being scattered with increasing distance is described by the ballistic transmittance T_b:

$$T_b(x) = e^{-\mu_s x}. \tag{2.8}$$

Anisotropy

The dimensionless measure anisotropy, g, defines the mean deflection angle projected on the original photon trajectory of a scattering event. As schematically shown in Figure 2.1.3, a photon scattered by a structure changes its direction with a deflection angle θ. The anisotropy defines the expectation value of the $\cos(\theta)$, defined by:

$$g = \int_0^\pi p(\theta) \cos(\theta) 2\pi \sin(\theta)\, \mathrm{d}\theta = \mathbb{E}(\cos(\theta)), \tag{2.9}$$

with $p(\theta)$ being the ratio of the light scattered into the angle θ and $\int_0^\pi p(\theta) 2\pi \sin(\theta)\, \mathrm{d}\theta = 1$. A typical value for biomedical tissue is $g = 0.9$.

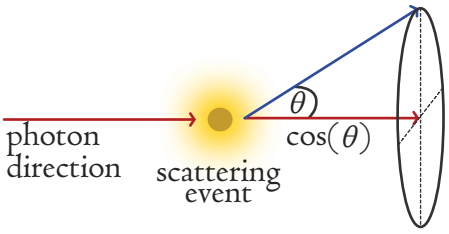

Figure 2.1.3.: A photon scattered by a structure changes its direction with a deflection angle θ [Prahl et al., 2017].

2.1.2. Photoacoustic Image Formation

Several steps are required to form a PA image, and the typical acquisition procedure is the following. First, a short light pulse of a wavelength in the visual and near-infrared range is sent into the tissue. The light propagates through the tissue as described by the Radiative Transfer Equation (RTE). Assuming certain conditions are met, it follows the phenomenon that gives the modality its name: the photoacoustic effect.

The photoacoustic effect is schematically visualized in Figure 2.1.4 and was first described by Alexander Graham Bell in 1880. Endogenous or exogenous chromophores in the tissue absorb the light's energy and convert it into heat, leading to a thermo-elastic expansion of surrounding tissue. The associated local pressure rise generates acoustic sound waves that, after propagating to the tissue's surface, can be measured with an US detector. These detected time series data can be reconstructed into a PA image, for example, by conventional US imaging reconstruction algorithms. Typically, multispectral PA images are acquired by repeating the described image acquisition procedure with light pulses of different wavelengths.
The ensuing parts provide details about the RTE and the PA signal formation.

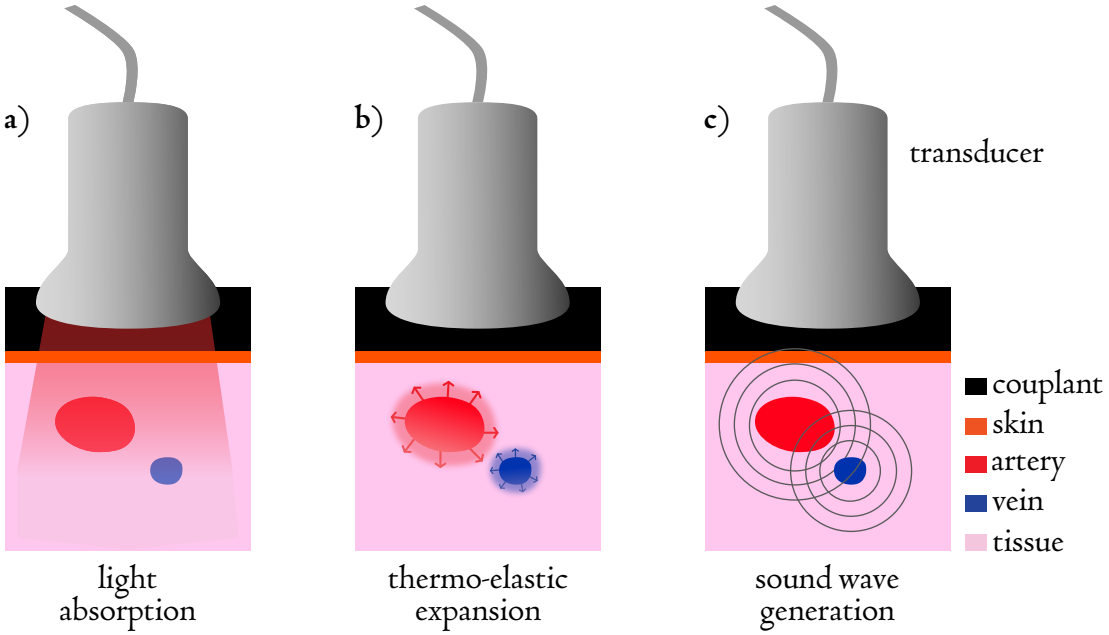

Figure 2.1.4.: The photoacoustic effect in biomedical imaging. (**a**) A light pulse in the visual or near-infrared range is directed into the tissue. Chromophores in the tissue absorb the light's energy and convert it into heat. (**b**) This leads to thermo-elastic expansion of surrounding tissue. (**c**) The resulting local pressure rise generates acoustic sound waves that propagate through the tissue, and that can be measured with an US detector.

Radiative Transfer Equation

The RTE describes the photon transport in biological tissue analytically. It is derived from the principle of conservation of energy and defined by five energy terms [Wang et al., 2012]:

$$\begin{underset{(1)}{\mathrm{d}P} = \underset{(2)}{-\,\mathrm{d}P_{\mathrm{div}}} \underset{(3)}{-\,\mathrm{d}P_{\mathrm{ext}}} + \underset{(4)}{\mathrm{d}P_{\mathrm{sca}}} + \underset{(5)}{\mathrm{d}P_{\mathrm{src}}}\,.} \tag{2.10}$$

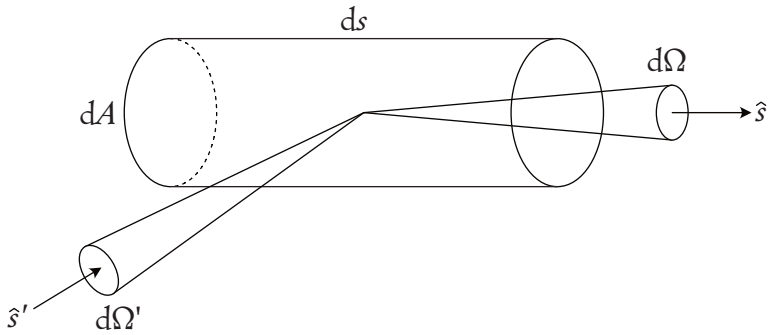

Figure 2.1.5.: Stationary differential cylindrical volume element with differential length $\mathrm{d}s$ and differential area element $\mathrm{d}A$. $\hat{s}$ and $\hat{s}'$ are the photon propagation directions and $\mathrm{d}\Omega$ and $\mathrm{d}\Omega'$ are the corresponding differential solid angle elements [Wang et al., 2012].

To simplify the explanation of these five terms, a stationary differential cylindrical volume element is considered following the derivation by Wang et al., 2012 (cf. Figure 2.1.5). The cylindrical volume element has a differential length element $\mathrm{d}s$ and a differential area element $\mathrm{d}A$ that is perpendicular to $\mathrm{d}s$. The photon propagation direction along $\mathrm{d}s$ is denoted by $\hat{s}$, and $\mathrm{d}\Omega$ is the corresponding differential solid angle element. $\hat{s}'$ and $\mathrm{d}\Omega'$ are another photon propagation direction and associated differential solid angle element. The five terms can be explained as:

(1) $\mathrm{d}P$ defines the *overall change in energy* in the volume element $\mathrm{d}V = \mathrm{d}A\,\mathrm{d}s$ within the solid angle element $\mathrm{d}\Omega$ per unit time t.

(2) $\mathrm{d}P_{\mathrm{div}}$ is the energy diverging out from the volume element or the solid angle element per unit time, if the photon beam is not collimated.

(3) $\mathrm{d}P_{\mathrm{ext}}$ defines the energy that is extinct due to absorption and scattering in the volume element within the solid angle element per unit time.

(4) $\mathrm{d}P_{\text{sca}}$ is the energy per unit time that is gained from surrounding tissue from any direction $\hat{s}'$ and scattered into the solid angle element.

(5) $\mathrm{d}P_{\text{src}}$ describes the produced energy by the light source entering the volume element within the solid angle element per unit time.

Note that other factors, such as coherence, polarization, and non-linearity, are neglected in this formulation of the RTE. In more detail, Equation 2.10 is defined as [Wang et al., 2012]:

$$\underset{(1)}{\frac{\partial L(\vec{r},\hat{s},t)/c}{\partial t}} = \underset{(2)}{-\hat{s}\nabla L(\vec{r},\hat{s},t)} \underset{(3)}{-\mu_t L(\vec{r},\hat{s},t)} + \underset{(4)}{\mu_s \int_{4\pi} L(\vec{r},\hat{s}',t)P(\hat{s}'\cdot\hat{s})\,\mathrm{d}\Omega'} + \underset{(5)}{S(\vec{r},\hat{s},t)}. \tag{2.11}$$

L is the radiance [$\mathrm{Wm}^{-2}\mathrm{sr}^{-1}$], that defines the energy flow per unit normal area per unit solid angle, c represents the speed of light, L/c defines the propagating energy per unit volume per unit solid angle, $\mu_t = \mu_a + \mu_s$ the extinction coefficient, and S the light source's energy. A detailed derivation of all parts of the equation can be found in the book by Wang et al., 2012.

The RTE having six independent variables is generally difficult to solve. Typically, the assumption that biomedical tissue is highly scattering ($\mu_a \ll \mu_s$) and that the scattering medium is nearly isotropic after sufficient scattering events holds. Note that the isotropic photon propagation can be described by the reduced scattering coefficient $\mu_s' = \mu_s(1-g)$. With these assumptions, the RTE can be simplified by the diffusion approximation.
A more accurate approximation of the RTE can be performed with numerical Monte Carlo (MC) methods. These methods rely on stochastic processes whose expected value of a random variable corresponds to a physical quantity of interest. Unlike the absorption of actual photons, which is a binary process, the absorption of virtual photons for MC-based photon propagation is typically modeled by tracking the absorption probability on every path segment [Fang et al., 2009]. In other words, virtual photons, so-called photon packets, propagate step-wise through the medium and lose a fraction of their absorption probability, which is referred to as a packet weight W_p. First, a photon packet with a specified packet weight W_p is launched at the light source in a specific direction. For each pixel/voxel along that direction, the packet weight is decreased by the absorption coefficient along that step size. After the scattering length is reached, a new direction is calculated. This process continues until the photon packet leaves the medium or is terminated by the *Russian Roulette* technique. The latter technique is applied when the packet weight falls below a certain threshold. This light-weighted photon packet adds little information for continued photon propagation but must be properly terminated to conserve

energy. The technique of Russian Roulette determines with the probability $1/m$ whether a photon packet with an updated weight package $m \cdot W_p$ survives. Both diffusion approximation and Monte Carlo methods are derived and explained in detail in the book by Wang et al., 2012.

Photoacoustic signal formation

Two confinements need to be fulfilled to allow a PA signal to be formed. The light pulse that is used to illuminate the tissue needs to be much shorter than the thermal and stress relaxation times, τ_{th} and τ_{st}, such that heat and stress conduction is negligible during the pulse duration. The thermal relaxation time describes the thermal diffusion and is defined as:

$$\tau_{\text{th}} = \frac{d_c^2}{\alpha_{\text{th}}}, \tag{2.12}$$

with d_c [m] being the characteristic dimension of the heated region and α_{th} [m^2/s] the thermal diffusivity. The stress relaxation time characterizes the pressure propagation and is defined as:

$$\tau_{\text{st}} = \frac{d_c}{v_s}, \tag{2.13}$$

with the speed of sound v_s [$\text{m}\,\text{s}^{-1}$].

Given the light pulse fulfills the confinements, for example, by using a laser pulse of a few nanoseconds, the absorption of the light's energy by the chromophores leads to a local rise in pressure p_0 [Pa]. This initial pressure is defined by:

$$p_0 = \frac{\beta}{\kappa} T = \frac{\beta}{\kappa} \frac{\eta_{\text{th}} A_e}{\rho C_v}, \tag{2.14}$$

where β [K^{-1}] represents the thermal coefficient of volume expansion, T [K] the changes in temperature, and κ [Pa^{-1}] the isothermal compressibility. The temperature change T can be reformulated with the percentage that is converted into heat η_{th}, the specific optical absorption A_e [J/m^3], the mass density ρ [kg/m^3], and the specific heat capacity at constant volume C_v [J/kgK].

With the use of the dimensionless Grüneisenparameter defined as

$$\Gamma = \frac{\beta}{\kappa \rho C_v}, \tag{2.15}$$

Equation 2.14 can be rewritten as:

$$p_0 = \Gamma \eta_{\text{th}} A_e = \Gamma \eta_{\text{th}} \mu_a \Phi, \tag{2.16}$$

where Φ represents the optical fluence [J/cm^{-2}]. A typical value of the temperature-dependent Grüneisenparameter is $\Gamma(37°) = 0.2$, and initial pressure values are usually in the order of $\sim$ 10 kPa [Treeby et al., 2010].

The initial pressure generates an US wave that propagates through the tissue, which can be described by linear acoustics [Treeby et al., 2010]. Assuming a lossless medium, the respective equations of motion, continuity, and state can be represented by [Morse et al., 1969]:

$$\begin{aligned} \frac{\partial u}{\partial t} &= -\frac{1}{\rho_0} \nabla p, \\ \frac{\partial \rho}{\partial t} &= -\rho_0 \nabla u, \\ p &= v_s^2 \rho, \end{aligned} \tag{2.17}$$

with the initial conditions of $p_0 = \Gamma \eta_{\text{th}} \mu_a \Phi$ and $\frac{\partial p_0}{\partial t} = 0$. This means the time evolution of the pressure p depends, among others, on the acoustic particle velocity u, the ambient density ρ_0, and the acoustic density ρ. The combination of these equations [Cox et al., 2005] leads to the photoacoustic wave equation:

$$\left(\nabla^2 - \frac{1}{v_s^2} \frac{\partial^2}{\partial t^2} \right) p(\vec{r}, t) = -\frac{\Gamma}{v_s^2} \frac{\partial H(\vec{r}, t)}{\partial t}, \tag{2.18}$$

where $p(\vec{r}, t)$ represents the acoustic pressure at location $(\vec{r})$ and time t. $H(\vec{r}, t)$ [W/m^3] is the heating function defined as:

$$H(\vec{r}, t) = \rho C_v \frac{\partial T(\vec{r}, t)}{\partial t}. \tag{2.19}$$

2.2. Deep Learning

Deep Learning is undeniably a field of active research and shows innovations in many practical applications [Goodfellow et al., 2016]. Various domains have been revolutionized by DL, such as the ones related to object recognition, object detection, speech recognition, and natural language understanding [LeCun et al., 2015]. In general, DL is a powerful and versatile subfield of Machine Learning (ML), whose systems aim to automatically learn and improve with experience and data to gain knowledge about an environment. More specifically, these systems recognize patterns from a set of data. However, automatic pattern recognition strongly depends on the way the data is represented. The concept of DL relies on learning representations of data on multiple hierarchical levels with a more abstract level always building on a simpler one [Goodfellow et al., 2016]. In other words, the data is transformed by simple, non-linear computing operations into an abstract representation, which is transformed into a more abstract representation by another transformation, and so on.

In this section, the key concepts of DL and related components used in this thesis are presented. First, an overview of the typical optimization procedure in DL (cf. Section2.2.1) is given. Then, a section highlighting the differences between super- and unsupervised learning follows (cf. Section 2.2.2). Another important distinction in DL to discriminative and generative models is described in the subsequent section (cf. Section2.2.3). Lastly, neural network types applied in this work are given in Section 2.2.4. For further fundamentals and details on DL, the books by Goodfellow et al., 2016 and Aggarwal, 2023 are strongly recommended. These books also served as the basis for the ensuing section.

2.2.1. Deep Learning Optimization

Following the book by Goodfellow et al., 2016, the aim in DL is typically to solve a task for some data, which is referred to as the test data (~ 20 % of data). The result of the task is commonly quantified by a performance measure P. For this purpose, the goal is to optimize a DL-based model f parametrized by Θ to achieve a high performance measured with P while relying solely on data similar to the test data, referred to as the training data (~ 60 % of data). In other words, in contrast to a conventional optimization algorithm, P is optimized indirectly in DL. For the optimization, a cost function $J(\Theta)$ is typically minimized based on the given training data x.

It is described as the average of the individual loss terms L of the model estimations $f(x, \Theta)$ and, if given, a target y:

$$J(\Theta) = \mathbb{E}(x, y)_{\sim p_{\text{data}}} L(f(x; \Theta), y), \tag{2.20}$$

where p_{data} denotes the empirical data distribution. Note that this *empirical risk minimization* problem would describe a conventional optimization problem if the data distribution would follow *the true data-generating* distribution $p_{\text{true, data}}$.

The optimization of the model's parameters Θ is typically performed iteratively with gradient-based methods, similar to conventional optimization algorithms. However, in DL, most optimization algorithms, also referred to as *optimizers*, follow minibatch stochastic gradient descent methods [Goodfellow et al., 2016]. This means the cost function is calculated on a subset of the training data, a so-called minibatch. The computed gradients of the corresponding cost function with respect to the algorithm's parameters are backpropagated to perform an update of the parameters Θ. In this context, the *minibatch size* defines the number of samples in a minibatch, and an *epoch* defines a cycle of updates in which all training data was used once. The optimization depends on different hyperparameters, such as the minibatch size or the learning rate of the optimizer. During training, the model is typically applied to another data set, referred to as the validation set ($\sim$ 20 % of data), in order to optimize these hyperparameters.

2.2.2. Supervised and Unsupervised Learning

Generally, the optimization schemes in DL/ML can be distinct into two classes depending on the data the algorithms *experience* [Goodfellow et al., 2016, Johnson, 2019]: *supervised* and *unsupervised learning*. Figure 2.2.1 gives an illustrative example.

Supervised learning is the most commonly applied training scheme in DL/ML [LeCun et al., 2015] where the algorithms experience labeled data. Typically, a mapping function from the data x to the label y via discriminative models (cf. Section 2.2.3) is achieved with supervised learning.

Unsupervised algorithms leverage data without any additional information, such as labels. These algorithms aim to find patterns in the underlying structure of the data, for example, to cluster similar samples of the data or to model the probability distribution that generated the data. The latter is of high interest in DL and can be learned explicitly, as in density estimation, or implicitly. A popular example of implicit learning of the probability distribution of the

data is the use of generative models for image synthesis (cf. Section 2.2.3), for instance with GANs [Goodfellow et al., 2016].

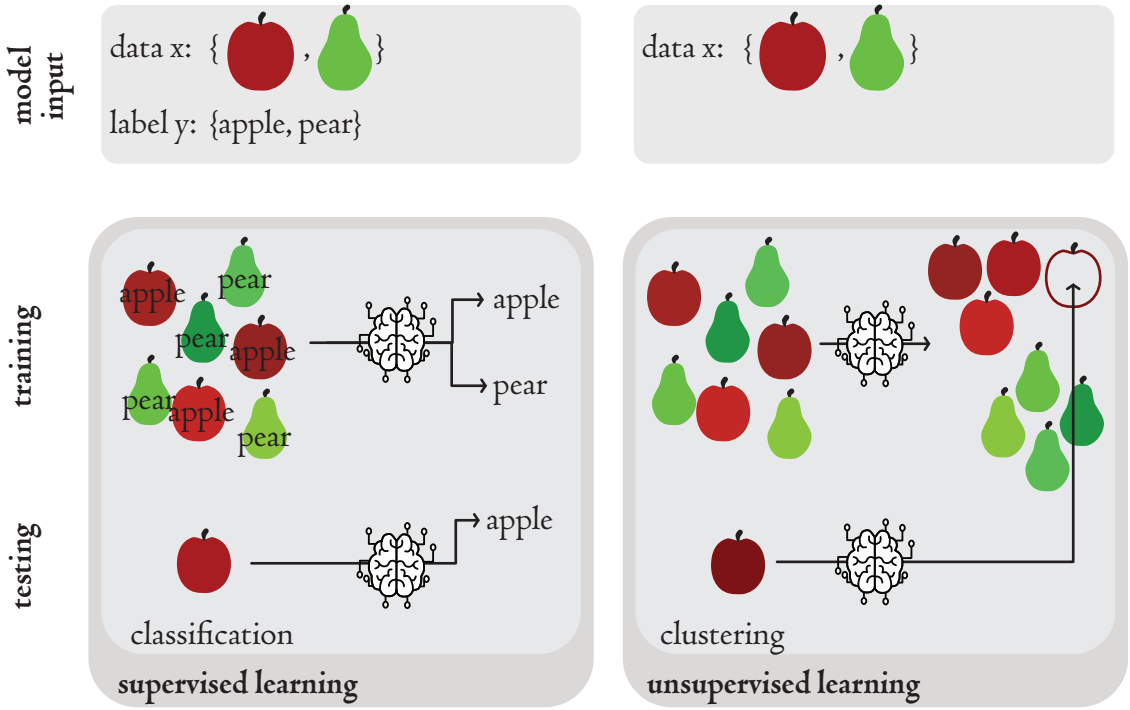

Figure 2.2.1.: Schematic and simplified example of supervised and unsupervised learning. In supervised learning, the models experience data x along with labels y, for example, for classification tasks. In the illustrated classification example, the classifier learns to map one of the labels "apple" or "pear" to the data. At inference, the test apple is categorized into the correct discrete label "apple". In unsupervised learning, data x without any further information is provided, and the model learns to find patterns in the underlying structure of the data. The clustering example shows a model that learns to understand features typical for apples and pears such that two respective clusters are created, and an unseen test apple is assigned to the correct "apple cluster".

2.2.3. Discriminative and Generative Models

There is a second major distinction in DL/ML into *discriminative* and *generative* models [Johnson, 2019]. The separation is based on different types of probability distributions that the models fit to the training data during learning.

The concepts and typical tasks for discriminative and generative models, along with loss functions and assessment metrics used in the thesis, are presented in the following.

Discriminative models

Discriminative models learn to estimate the probability distribution $p(y|x)$ that defines the probability of a label from y given an input sample of the training data x. They are closely aligned with supervised learning [Johnson, 2019]. The two most common tasks for supervised discriminative models are (a) classification and (b) regression.

(a) Classification task

In classification, the model learns to predict a categorical variable, the label y, for a data sample x. The categorical variable is discrete and one out of a finite number of different categorical variables [Lindholm et al., 2022]. Specifically, in this thesis, the classification task tackled was semantic segmentation of images. In semantic segmentation, each pixel of an image is classified into one categorical variable. Two types of loss functions can be distinguished, which take into account the information of either a single pixel or of a full image.

Single pixel-based loss The two loss functions applied in this thesis considering single pixel information for image classification tasks are the Cross-Entropy (CE) and soft margin loss. The CE is defined as [Goodfellow et al., 2016][1]:

$$L_{\text{CE}}(\hat{y}, y) = \frac{1}{N} \sum_{n=1}^{N} \sum_{c=1}^{C} y_{n,c} \log \frac{\exp(\hat{y}_{n,c})}{\sum_{c'=1}^{C} \exp(\hat{y}_{n,c'})}, \tag{2.21}$$

where N is the minibatch size, and C is the number of classes. $\hat{y}_n$ is the estimated vector of the single pixel n with length C, and y_n is the corresponding reference one-hot encoded vector with an entry of one for the correct class.
The soft margin loss is defined as [Chatterji et al., 2021][2]:

$$L_{\text{SoftMargin}}(\hat{y}, y) = \frac{1}{N} \sum_{n=1}^{N} \sum_{c=1}^{C} \frac{\log\big(1 + \exp(-\hat{y}_{n,c} \cdot y_{n,c})\big)}{C}, \tag{2.22}$$

where $\hat{y}_n$ is the estimated vector of length C and y is the corresponding one-hot encoded reference vector that contains values of -1 and 1 for negative and positive examples, respectively.

[1] `https://pytorch.org/docs/stable/generated/torch.nn.CrossEntropyLoss.html`
[2] `https://pytorch.org/docs/stable/generated/torch.nn.SoftMarginLoss.html`

Image-based loss A typical loss function for semantic segmentation that includes the estimations of an entire image is the Dice loss [Drozdzal et al., 2016]. The work by Isensee et al., 2021 adapted this loss function by adding a smoothing factor and a constant value in the denominator for numerical stability. The adapted loss is referred to as Soft Dice loss and is defined as:

$$L_{\text{SoftDice}}(\hat{Y}, Y) = -\frac{1}{C}\sum_{c=1}^{C}\left(\frac{2|\hat{Y}_c \cap Y_c| + \epsilon_{\text{smooth}}}{|\hat{Y}_c| + |Y_c| + \epsilon_{\text{smooth}} + 1e^{-8}}\right). \tag{2.23}$$

Here, $\hat{Y}$ and Y are the estimated and reference images, respectively, and ϵ_{smooth} is the smoothing factor, chosen here as $1e^{-5}$. The constant value is set to $1e^{-8}$. Note that in this thesis, the Soft Dice loss is calculated per minibatch.

Assessment metrics According to the work by Maier-Hein et al., 2022, problem-aware selection of metrics is an important topic for validation of ML algorithms. Based on this work, the following overlap- and contour-based metrics were used in this thesis to assess the performance of the DL-based algorithms for semantic segmentation of images: the Dice Similarity Coefficient (DSC) and the Normalized Surface Distance (NSD).
The DSC [Dice, 1945] is defined as:

$$\text{DSC}_c = \frac{2|\hat{Y}_c \cap Y_c|}{|\hat{Y}_c| + |Y_c|}, \tag{2.24}$$

where $\hat{Y}_c$ and Y_c are the estimated and reference images corresponding to class c, respectively.
The NSD [Nikolov et al., 2021] is defined as:

$$\text{NSD}_c = \frac{|S_{\hat{y}_c} \cap B^{\tau}_{y_c}| + |S_{y_c} \cap B^{\tau}_{\hat{y}_c}|}{|S_{\hat{y}_c}| + |S_{y_c}|}, \tag{2.25}$$

with the tolerance τ and the surfaces $S_{\hat{y}_c}$ and S_{y_c} and the border regions $B_{\hat{y}_c}$ and B_{y_c} of the estimation and the reference for class c, respectively.

(b) Regression task

Regression tasks are similar to classification tasks. However, for regression, data with continuous labels instead of categorical labels are available. Thus, the format of the model's output is a continuous value instead of a discrete one [Goodfellow et al., 2016]. Although the regression tasks in this thesis are on the image level, the cost function is calculated as the average of loss values computed for each single pixel, as described in the following.

Single pixel-based loss A typical loss function for regression is the Mean Squarred Error (MSE). In this thesis, it was calculated as the average over single pixels of an image and the minibatch:

$$\text{MSE}(\hat{Y}, Y) = \frac{1}{N \cdot M} \sum_{n=1}^{N} \sum_{m=1}^{M} (\hat{y}_{n,m} - y_{n,m})^2, \tag{2.26}$$

where N is the minibatch size, M the number of pixels of an image, $\hat{y}$ and y the single-pixel estimation and reference, respectively.

Assessment metrics The performance of an image-level regression task is assessed with the relative error (RE), absolute error (AE), and the Structural Similarity Index Measure (SSIM) [Wang et al., 2004] in this thesis. The relative and absolutes errors for a single pixel estimation and reference, $\hat{y}$ and y, are defined as:

$$\text{RE} = \frac{|\hat{y} - y|}{y} \text{ and } \text{AE} = |\hat{y} - y|. \tag{2.27}$$

The SSIM compares two images with respect to luminescence, contrast, and structure. It is defined as:

$$\text{SSIM}(\hat{Y}, Y) = \frac{1}{M} \sum_{m=1}^{M} \left(\frac{(2\mu_{\hat{y}}\mu_y + (K_1 L)^2)(2\sigma_{\hat{y},y} + (K_2 L)^2)}{(\mu_{\hat{y}}^2 + \mu_y^2 + (K_1 L)^2)(\sigma_{\hat{y}}^2 + \sigma_y^2 + (K_2 L)^2)} \right), \tag{2.28}$$

with M being the number of pixels per image, μ the mean intensity, σ the standard deviation, and L the dynamic range of the pixel values. K_1 is usually set to 0.01 and K_2 set to 0.03 [Wang et al., 2004].

Generative models

Generative models aim to model the probability distribution $p(x)$ of example samples x or some properties of that distribution [Goodfellow et al., 2016]. Different approaches exist, such as explicit modeling techniques, methods that approximate $p(x)$, and methods that allow sampling from $p(x)$ without explicitly modeling it (implicit modeling). These models often go hand in hand with unsupervised models. However, there are also conditional generative models, which learn the conditional probability distribution $p(x|y)$ and, therefore, require labels [Johnson, 2019].

In this thesis, differentiable generative models are investigated. These models typically use differentiable neural networks to model $p(x)$ by transforming samples of a latent variable z into samples x or into distributions over samples x [Goodfellow et al., 2016]. This neural network is also called a *generator* $g(z; \Theta^{(g)})$, and the corresponding parameters are indicated by the superscript g. Popular differentiable generative models are Variational Autoencoders (VAEs) and GANs. In the following, the principles of (a) a GAN involving a discriminator network for training and (b) a Generative Moment Matching Network (GMMN) that trains the generator in isolation [Goodfellow et al., 2016] are explained as they are applied in this thesis.

(a) Generative Adversarial Network

Since the introduction of GANs [Goodfellow et al., 2014], this type of neural network has been successfully applied in many domains, generating, for example, natural images of high resolution and high perceived image quality across a variety of data sets [Karras et al., 2020a, Zhu et al., 2023]. The principle of a GAN is based on game theory, where a generator competes against an adversary discriminator. More specifically, the generator learns to generate samples $x = g(z; \Theta^g)$, and the discriminator $d(x; \Theta^d)$ learns to distinguish these "fake" generated samples from "real" training samples [Goodfellow et al., 2016]. Note that the parameters of the discriminator are indicated by the superscript d. Typically, the discriminator is a supervised discriminative classifier. As the generator, and thus the generated samples, become more realistic, it is more difficult for the discriminator to classify these samples as "fake" with high probability. Conversely, as the discriminator improves, it is more difficult for the generator to fool it.

The adversarial training is typically performed by optimizing the following min-max loss function:

$$\min_g \max_d L(\Theta^g, \Theta^d) = \mathbb{E}_{x \sim p_{\text{data}}}[\log d(x)] + \mathbb{E}_{z \sim p_{\text{model}}}[\log(1 - d(g(z)))], \tag{2.29}$$

with Θ^g and Θ^d being the parameters of the generator and discriminator, respectively [Goodfellow et al., 2014, Goodfellow et al., 2016]. The optimization of g and d is typically performed in an alternating fashion until the generated and real samples are indistinguishable, which results in a discriminator output of 1/2. At inference, only the generator is used [Goodfellow et al., 2014].

It is mathematically proven that once the optimal discriminator for any generator is found, the overall global minimum is reached when the probability distributions of the data and the model are exactly the same: $p_{\text{data}} = p_{\text{model}}$ [Goodfellow et al., 2014].

Although GANs are powerful tools to generate new samples that follow the training data distributions without the need to explicitly model $p(x)$, the training is often cumbersome. For example, mode collapse and diminishing gradients are common issues [Arjovsky et al., 2017]. However, active research is ongoing to stabilize GAN training, and the use of deep convolutional layers for image generation and the inclusion of dropout layers in the discriminator are just two example improvements [Goodfellow et al., 2016].

(b) Generative Moment Matching Networks

GMMNs [Li et al., 2015, Dziugaite et al., 2015] are another class of differentiable generative methods. Like GANs, the generator aims to generate samples that resemble the training data distribution. However, instead of using a discriminator during training, the concept approximates adversarial learning by relying on *moment matching*. In other words, the generator is trained with a loss function that compares all orders of statistics computed for the generated samples with the ones computed for the training samples. As soon as the statistics coincide, the generated samples are likely to follow the training data distribution [Li et al., 2015].

A *moment* is defined as the expectation of different powers of a random variable. For example, the first-order moment defines the mean; the second-order moment is the mean of squared values, and so on [Goodfellow et al., 2016]. Because the computation of these moments becomes computationally expensive when x contains many samples, GMMNs use a statistical hypothesis test [Li et al., 2015] based on the Maximum Mean Discrepancy (MMD) [Gretton et al., 2006].

The MMD is defined as [Dziugaite et al., 2015]:

$$\text{MMD} = \sup_{f \sim \mathbb{F}} (\mathbb{E} f(x) - \mathbb{E} f(y)) \tag{2.30}$$

where f is a function chosen from the function class $\mathbb{F}$, x are samples from the training distribution, and y are generated samples.

If $\mathbb{F}$ is the unit ball in a reproducing kernel Hilbert space with kernel k, the kernel trick can be applied, and the MMD can be rewritten as [Gretton et al., 2012]:

$$\text{MMD}^2 = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{i'=1}^{N} k(x_i, x_{i'}) - \frac{2}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} k(x_i, y_j) + \frac{1}{M^2}\sum_{j=1}^{M}\sum_{j'=1}^{M} k(y_j, y_{j'}), \quad (2.31)$$

where N and M denote the total number of training and generated samples, respectively.

In other words, the MMD implicitly maps the samples into an infinite-dimensional feature space by the kernel function, and first-order moments of the training and generated distributions in that feature space are computed. The loss function is zero if and only if the distributions are exactly the same.

To calculate the MMD, a kernel needs to be chosen [Li et al., 2015]. In this thesis, a multiscale polynomial kernel[3] is used, which is defined as:

$$k(x_i, x_j) = a^b \cdot \left(\frac{a + 0.5 \cdot x_i x_j}{b}\right)^{-b}, \quad (2.32)$$

with a and b to be set.

Since the MMD highly depends on the choice of kernel and its parameters and to ensure an appropriate selection of kernel parameters, one often uses a range of kernel bandwidths (corresponding to different values for a and b) [Li et al., 2015], and the final MMD is calculated as the sum of the individual results with different kernel bandwidths as:

$$L(x_i, x_j) = \sum_a \sum_b \text{MMD}^2_{a,b}(x_i, x_j). \quad (2.33)$$

2.2.4. Neural Network Types

Neural networks are a combination of simple units called artificial neurons. An artificial neuron is inspired by the function of a biological neuron [Aggarwal, 2023, Gurney, 1997]. In simple terms, a biological neuron processes input signals determining the membrane potential. Depending on whether the membrane's potential exceeds a certain threshold, an action potential with an

[3] `https://github.com/vislearn/analyzing_inverse_problems/blob/master/inverse_problems_science/losses.py`

"all-or-nothing character" is generated that propagates along an axon [Gurney, 1997]. Artificial neurons non-linearly transform $n+1$ input nodes x into a one-dimensional output node y in a similar fashion, as defined by:

$$y = g\left(\sum_{i=0}^{n} x_i w_i\right) = g\left(\sum_{i=1}^{n} x_i w_i + x_0 \cdot w_0\right). \tag{2.34}$$

The first operation is a weighted sum of the input with corresponding $n+1$ weights. Note that the first input usually equals one and, therefore, the multiplication $x_0 w_0$ defines the bias. A non-linear activation function g then transfers the sum into the output y [Aggarwal, 2023, Gurney, 1997]. Note that the use of nonlinear activation functions in DL is crucial, as they alone introduce nonlinearity into a model that allows complex functions to be approximated. A schematic is shown in Figure 2.2.2.

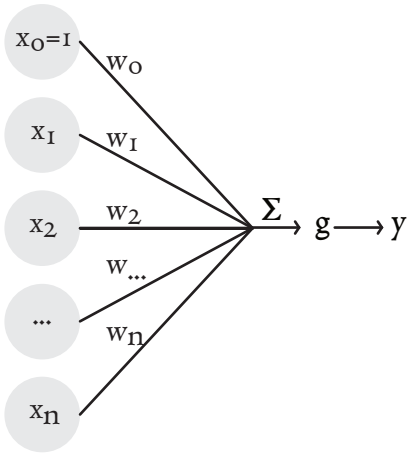

Figure 2.2.2.: Schematic of an artificial neuron. A weighted sum is applied on the $n+1$ dimensional input x with corresponding weights w. The first input usually equals 1, and its multiplication with w_0 defines the bias. A non-linear activation function g transfers the sum to the output y.

Popular activation functions that are used in this thesis are the Leaky Rectified Linear Unit (LeakyReLU), the Gaussian Error Linear Unit (GELU), the Tangens Hyperbolicus (TanH), the Sigmoid, and the Softmax functions. An overview of these functions, including their definitions, is shown in Figure 2.2.3.

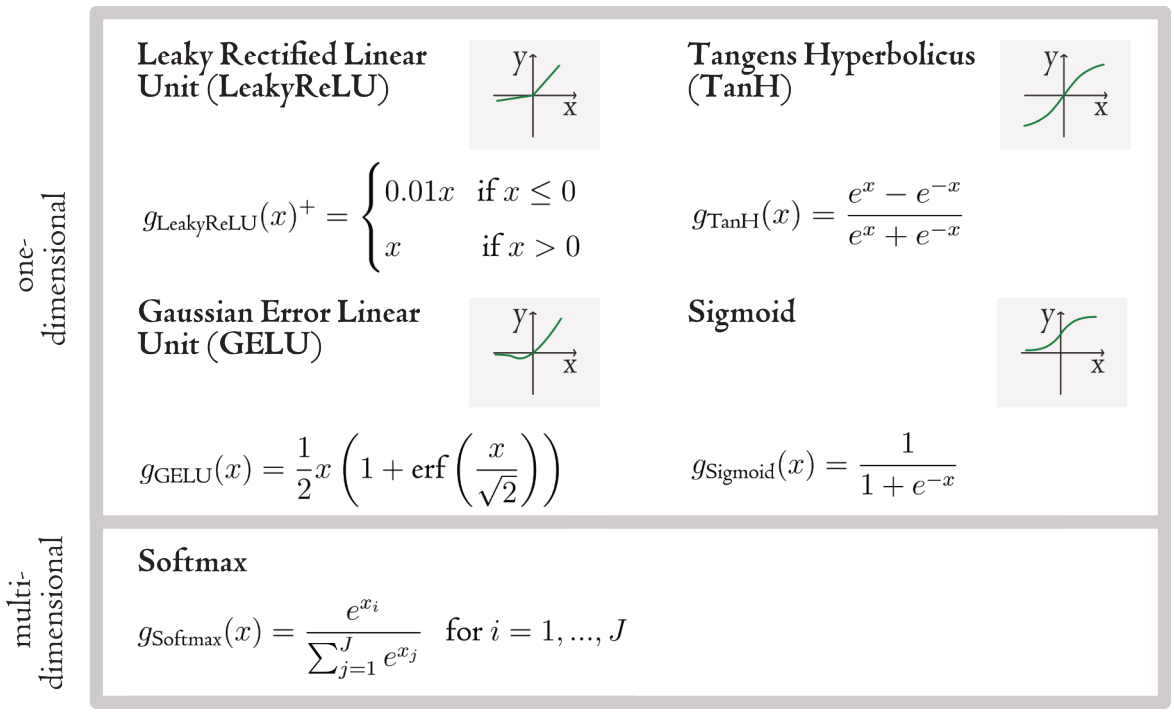

Figure 2.2.3.: Activation functions g used in this thesis. Four activation functions, namely the Leaky Rectified Linear Unit (LeakyReLU), Tangens Hyperbolicus (TanH), Gaussian Error Linear Unit (GELU), and Sigmoid functions, were applied to a one-dimensional input x. For each of the functions, an example curve is sketched. Note that the scales of the x- and y-axes are not the same for the individual plots. Specifically for the GELU function, the Gaussian error function is indicated by erf. The Softmax function was applied to handle multi-dimensional input x. The definitions and illustrative example plots are given.

One of the units combining the weighted sum and an activation is also referred to as a computational layer. A neural network uses multiple layers, where the output of one layer is used as the input of another one, and so on. The intermediate layers are also known as *hidden* layers. During training, the weights/parameters of the different units are optimized. The fundamental type of neural network is a feed-forward neural network, where the input information is forwarded to the output only in one direction [Aggarwal, 2023, Goodfellow et al., 2016]. In this thesis, only feed-forward networks are considered.

Fully-Connected Neural Network

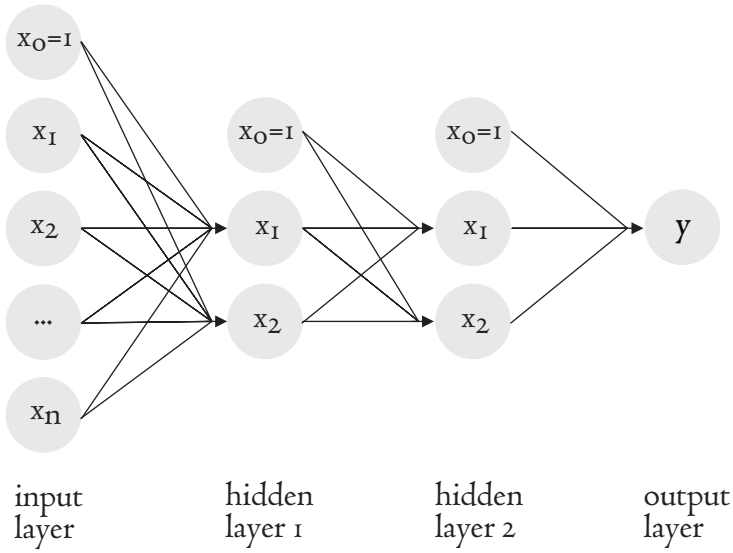

Figure 2.2.4.: Example of a Fully-Connected Neural Network (FCNN) with two hidden layers. An input of $n + 1$ dimensions is transformed into a (hidden) output of two dimensions by a fully-connected transition, including an activation function. One input node x_0 is set to one and serves to learn a bias. The (hidden) output and another additional node x_0 are the input of another fully-connected transition with the same (hidden) output dimensions. The one-dimensional output y is reached after a final fully-connected transition.

The basic feed-forward network architecture is a Fully-Connected Neural Network (FCNN) where every output node of a layer is a combination of all input nodes of the previous layer. An example of a FCNN with two hidden layers is shown in Figure 2.2.4. FCNNs have a simple and verstile design. However, the number of parameters grows quickly with increasing number of layers and sizes of layers and input, respectively. Therefore, FCNNs are generally less suitable for analyzing high-dimensional inputs like images.

Convolutional Neural Networks

In computer vision, Convolutional Neural Networks (CNNs) are most often used and show remarkable success stories across a variety of domains. For example, CNNs achieve human performance for some recognition and detection tasks [LeCun et al., 2015]. In contrast to FCNNs, CNNs are based on convolutional kernels and particularly suited for an input consisting of multiple arrays, such as images, which have a two-dimensional grid structure of pixels and, moreover, often possess multiple channels, such as the red-green-blue (RGB) channels.

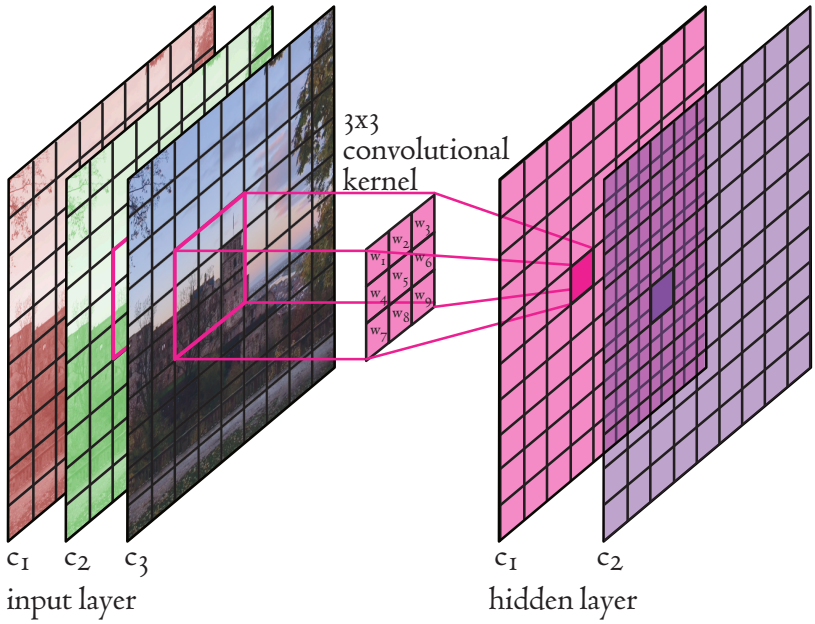

Figure 2.2.5.: Example of a convolutional layer of a Convolutional Neural Network (CNN). An input image with three channels c is transformed into a two-channel output by applying two convolutional kernels.

In a convolutional layer, a convolutional kernel is stepped across the whole multi-channel array, and the results of each of the convolutional computations are written in an output feature map. This means that all channels from the previous layer are considered for one output layer. Applying another kernel corresponds to an additional output layer, and so on (cf. Figure 2.2.5). During training, the individual kernel weights are optimized for the task at hand. The motivation for using convolutional kernels is three-fold. First, neighboring pixels in images are usually highly correlated, and second, most image structures are spatially invariant and can appear at any location - both properties that a convolution takes into account by design. Third, a convolutional layer has a significantly lower number of learnable parameters in comparison to a layer of a FCNN. In this thesis, two specific types of CNNs are implemented, the U-Net and the Fourier Neural Operator (FNO).

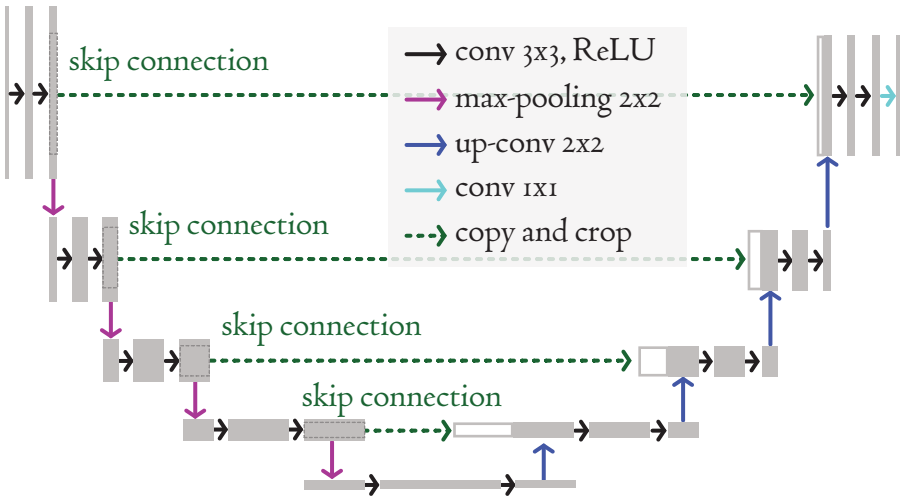

Figure 2.2.6.: The original U-Net architecture proposed by Ronneberger et al., 2015. A two-dimensional input is processed by two convolutional layers (kernel size 3 x 3 px and including a Rectified Linear Unit (ReLU) activation) and then downscaled by a max-pooling layer (kernel size 2 x 2 px). This process is repeated four times, followed by two convolutional layers. The output is then upscaled by an up-convolutional layer (kernel size 2 x 2 px) whose output, together with a copy of the convolutional layer of the same hierarchical level (skip connections), is the input of another two convolutional layers. This upscaling process is also repeated four times. At the end, a final convolutional layer (kernel size 1 x 1 px) is applied. Note that the ReLU is equal to the Leaky Rectified Linear Unit (LeakyReLU), except that all values for $x < 0$ are set to zero.

U-Net The U-Net [Ronneberger et al., 2015] is a CNN-based network architecture originally developed for the segmentation of biomedical images. One concept behind a U-Net is that convolutional layers are applied on different hierarchical levels, which is generally typical for CNNs. However, the architecture consists not only of a downsampling part but also an upsampling part, which explains the "U" in the name. For the downsampling, pooling layers are important because they merge semantically similar features [LeCun et al., 2015]. For example, an image can be downscaled with *max-pooling* by keeping only the maximum value within a local neighborhood of several pixels. The upsampling part is usually realized with transposed convolutions. In addition, the U-Net uses residual skip connections that connect the original or downsampled representation with the upsampled one, which emphasizes detailed features that are more present in the first representations compared to the last ones. The original U-Net implementation is shown in Figure 2.2.6.

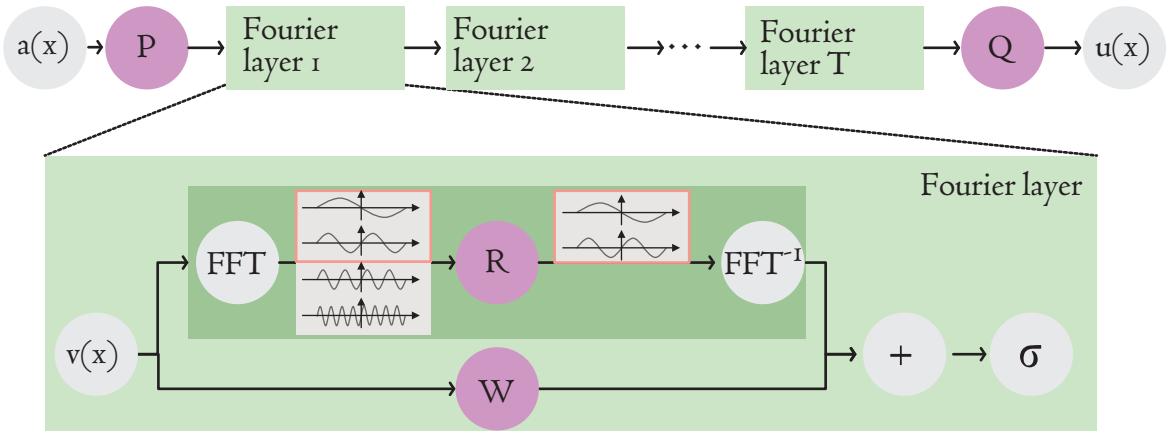

Figure 2.2.7.: Fourier Neural Operator (FNO) network. The input function a discretized at x is upscaled by a linear layer P. T Fourier layers follow until another linear layer Q is applied for downscaling, which results in the output $u(x)$. A Fourier layer consists of two paths. First, the input $v(x)$ is transformed by a Fast Fourier Transformation (FFT) into the Fourier domain. A specific number of Fourier modes is multiplied by the linear transformation R, and an inverse FFT follows, which represents the final step of the first path. Second, $v(x)$ is linearly transformed by W, and the output is summed up with the results of the first path. At the end, an activation function σ is applied. Note that the transformations highlighted in purple are learnable.

Fourier Neural Operator network The FNO network has its origin in the challenge of solving Partial Differential Equations (PDEs) [Li et al., 2020c]. They are based on neural operators [Li et al., 2020d] and enable their efficient implementation. In general, neural operators learn a mapping between mesh-free, infinite-dimensional function spaces. In other words, the input a and the output u of the network are continuous functions with potentially different discretizations, and the network learns the mapping between these two domains by exploiting some examples $a(x)$ and $u(x)$. Therefore, neural operators provide solutions that are independent of the discretization. Their success has been shown in the context of PDEs in comparison to classical methods such as (neural) finite element methods. The concept is based on the combination of linear, global integral operators and non-linear, local activation functions. In more detail, neural operators use iterative updates defined as:

$$v_{t+1}(x) = g\left(Wv_t(x) + (Kv_t)(x)\right) \tag{2.35}$$

$$\text{with} \quad (Kv_t)(x) = \int_D k(x,y)v_t(x)\,\mathrm{d}y\,, \tag{2.36}$$

where g denotes a non-linear activation function, W a (usually learned) linear transformation, $v_t(x)$ a representation at step t, and K a kernel integral transformation with the kernel function k that is also learned. In FNOs, the integral operator is formulated as a (global) convolution and implemented by a FFT to decrease the complexity from $\mathcal{O}(N^2)$ to $\mathcal{O}(N \cdot \log(N))$. This means the iterative update can be written as:

$$v_{t+1}(x) = g\left(W \cdot v_t(x) + \text{FFT}^{-1}(R \cdot \text{FFT}(v_t))(x)\right), \tag{2.37}$$

where R is a (simple) linear transformation to be learned. The bias term W is introduced to keep non-periodic features. In practice, FNO networks are typically implemented by first upscaling the input into a latent space with a linear transformation. Then, several blocks of FNOs follow. Note that R is commonly only applied to a specific number of lowest Fourier modes as a form of regularization (low-pass filtering). Finally, the last representation is downscaled to provide the output. A schematic visualization of a FNO network is shown in Figure 2.2.7. In this thesis, FNO networks are applied for learning photon propagation, which is based on the continuous RTE. The work by Rix et al., 2023 has successfully shown the application of FNO networks for this purpose.

Graph Neural Networks

Graph Neural Networks (GNNs) are a class of neural networks that deal with graph data which shows complex relationships and interdependency between objects [Wu et al., 2020]. GNNs fall under the umbrella of geometric DL that aims to develop DL techniques for non-Eucledian domains [Bronstein et al., 2017]. The ground-breaking performances of GNNs have been shown for different tasks [Zhou et al., 2020]. For example, GNNs have brought remarkable innovations in biology [Jumper et al., 2021] and natural language processing [Devlin et al., 2018]. In comparison to CNNs that are especially suited for grid-structured data, GNNs handle more complex data since graphs can be of arbitrary size and topology. Generally speaking, CNNs can also be seen as a subgroup of GNNs.

A graph consists of nodes that are connected by edges, and the graph structure is typically represented by the adjacency matrix A [Aggarwal, 2023]. In this matrix, the entry ij defines the weight of the connection between nodes i and j. If the edges are undirected, the adjacency matrix is symmetric. For some graphs or graph representations, the nodes and edges contain features. In general, two types of GNN learning exist that are separated with respect to node-centric or graph-centric predictions. While the first aims to analyze individual nodes, the second analyzes entire graphs. In this thesis, GNNs are used to analyze the nodes, or more

specifically the node features, of undirected graphs with edges that do neither contain features nor weights. While there are lots of different GNN architectures, the GNN of this thesis is based on transformer convolutions. These, in turn, are based on graph convolutions, [Kipf et al., 2016].

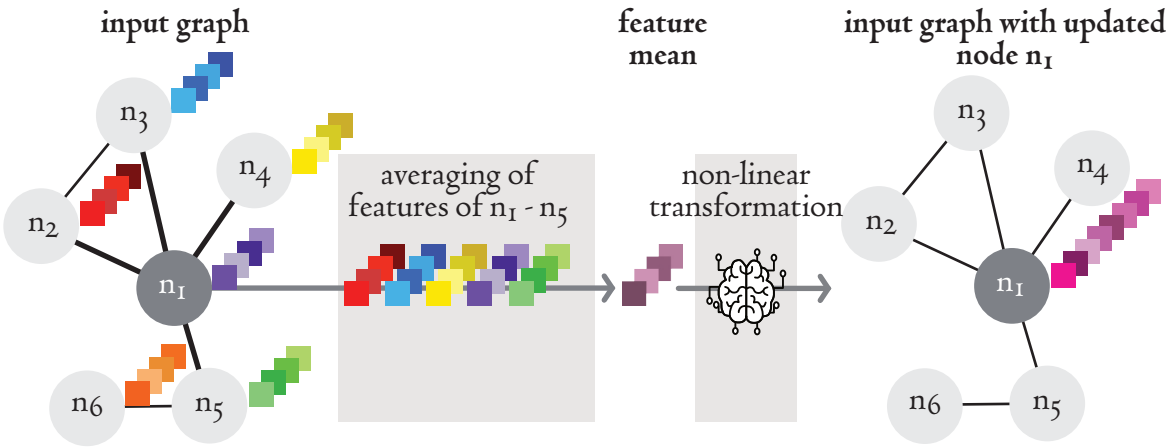

Figure 2.2.8.: Example of a graph convolutional for one target node n_1. First, the node features of the target node n_1 (here, four-dimensional) are averaged with the features of the target node's neighbors n_2 - n_5 that are indicated with a thicker edge width. This feature mean is used as the input of a neural network that non-linearly transforms it into a new representation (here, eight-dimensional). This procedure is repeated for all nodes of the input graph.

Graph convolutions are similar to convolutional kernels applied to images. In analogy to conventional kernels applied to a single pixel of an image, a graph convolution is applied to a single node of a graph. The analogy of channels of an image are the features of the nodes. However, other than for grid structures, the number of neighbors varies for graph nodes. In graph convolutions, the features of the target node $h_i^{(k)}$, as well as the neighboring features $h_j^{(k)}$ are first averaged and then non-linearly transformed into a new representation $h_i^{(k+1)}$. More specifically, an update of node features is defined as:

$$h_i^{(k+1)} = g\left(W^{(k)} \sum_{j \epsilon A_i \cup i} \frac{h_j^{(k)}}{\sqrt{|A_i \cdot A_j|}}\right) \tag{2.38}$$

where $W^{(k)}$ is a learnable linear transformation and g is the activation function. Note that the number of features in the previous layer k can be different from the number of the $k+1$-th layer. The denominator is a normalization, which is proportional to the degrees A (total number of edges) of the nodes [Aggarwal, 2023]. As usual for CNNs, several of these graph convolutions

follow each other. The more layers, the deeper the network, and the more distant neighboring features are considered for a target node. A schematic of a single update of a target node is given in Figure 2.2.8.

The specific type of transformer convolution does not simply average features of the target and neighboring nodes but learns a weighting of how the neighboring features shall be considered for the node at hand. This concept is popularly known as attention [Vaswani et al., 2017]. In this thesis, the implementation of transformer convolutions introduced by Shi et al., 2020 is applied, which includes a multi-head attention mechanism. The feature update of the target node h_i is defined as:

$$h_i^{(k+1)} = W_{\text{residual}}^{(k)} h_i^{(k)} + \frac{1}{C} \sum_{c=1}^{C} \left(\sum_{j \epsilon A(i)} \alpha_{c,\text{ij}}^{(k)} W_{c,\text{value}}^{(k)} h_j^{(k)} \right) \tag{2.39}$$

$$\alpha_{c,\text{ij}} = g \left(\frac{W_{c,\text{query}}^{(k)} h_i^{(k)} W_{c,\text{key}}^{(k)} h_j^{(k)}}{\sqrt{d}} \right). \tag{2.40}$$

In this formula, four learnable linear transformations W are included. The first one is referred to as residual W_{residual} that linearly transforms the features of the target node. The second one transforms the features of the neighboring nodes $h_j^{(k)}$ into *value* vectors by the multiplication with $W_{c,\text{value}}^{(k)}$ for each of the c attention heads. The output is multiplied with the attention coefficients $\alpha_{c,\text{ij}}^{(k)}$. These are computed by a Softmax activation g applied on the multiplication of the *query* vector and the *key* vector divided by the hidden size d of each head. The query and key vectors are computed in analogy to the value vector by multiplying the third and fourth transformations, $W_{c,\text{query}}^{(k)}$ and $W_{c,\text{key}}^{(k)}$, with the features $h_i^{(k)}$ and $h_j^{(k)}$, respectively.

To summarize this section, the type of neural network needs to be chosen appropriately, depending on the task. Furthermore, it is necessary to specify several additional parameters during implementation, which have not been included in this section. For example, the *kernel size*, the *stride*, and the *padding* need to be set for CNNs. But also other strategies that improve training need consideration. For example, the choice of optimizer is crucial, and it might be important to include normalization layers, dropout layers, augmentations strategies, label smoothing, and regulations. For further details regarding these methods, the books by Aggarwal, 2023 and Goodfellow et al., 2016 are recommended.

3. Related Work

This chapter presents work related to the thesis. In particular, the basic concepts and outstanding individual publications on semantic segmentation in the field of medical imaging and specifically in PAI are explained. (cf. Section 3.1). Additionally, concepts for generating content in the field of computer vision and, in analogy, anatomies in medical imaging are presented (cf. Section 3.2). Beyond that, the main pillars for virtual PA image generation and existing methods for tissue geometry generation in the PAI field are described.

3.1. Semantic Segmentation

Semantic segmentation is one of four classification tasks (cf. Figure 3.1.1). Generally, an image can be classified into categorical target variables at the image, object, and/or pixel level [Maier-Hein et al., 2022]. A classification that considers an entire image is defined as *image-level classification*. Predicting distinct objects of different classes in an image is called *object detection*. *Semantic segmentation* is the classification of each pixel of an image into one of the categorical variables. The combination of pixel-wise and object-wise classification is referred to as *instance segmentation*. The four classification tasks are illustrated using a PA image in Figure 3.1.1.

This section presents related work regarding automatic semantic segmentation in the field of medical imaging and PAI. In the field of medical imaging, the focus is on DL-based methods, given that these data-driven methods generally outperform traditional techniques and are therefore considered state of the art. In contrast, for PAI both traditional and data-driven approaches that have emerged in recent years are covered.

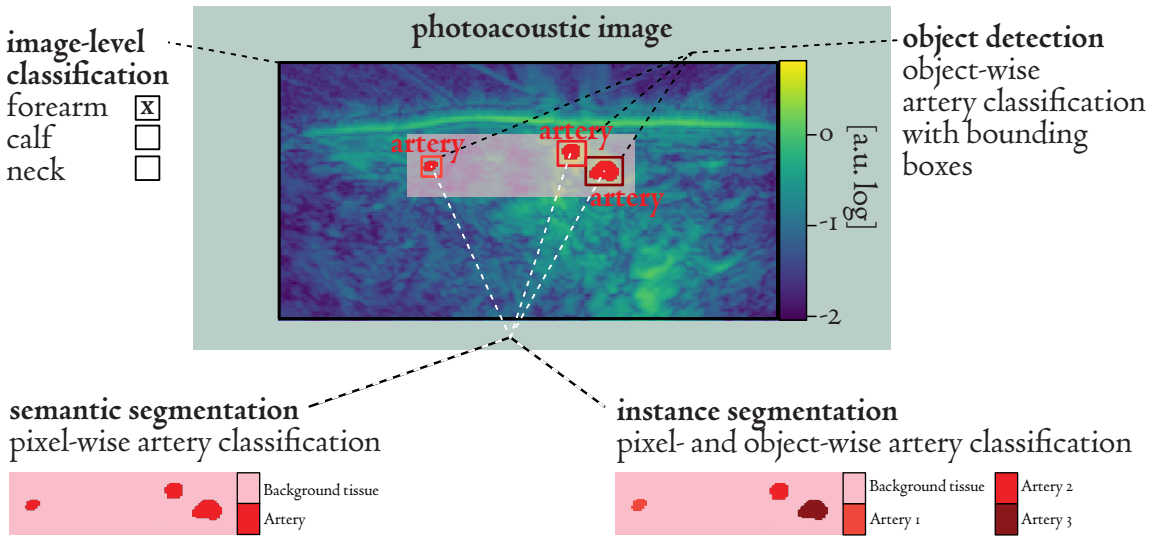

Figure 3.1.1.: Four classification tasks performed on a Photoacoustic (PA) image. The classification of the entire PA image as a *forearm* image is an example of *image-level classification*. Detecting different arteries belongs to the *object detection* category. The pixel-wise classification of arteries is an example of *semantic segmentation*. Differentiating these semantically segmented arterial pixels into different object categories is an example of *instance segmentation*. Note that for semantic and instance segmentation, only a cutout of the actual PA image is shown, which is highlighted in the PA image by the semitransparent area.

3.1.1. Semantic Segmentation in Medical Imaging

Semantic segmentation in medical imaging is considered one of the most challenging tasks in medical image analysis [Hesamian et al., 2019]. At the same time, semantic segmentation is shown to be important to assist radiological experts in a variety of clinical applications, such as image-guided interventions, diagnosis, and radiotherapy planning across different imaging modalities, including MRI, Positron Emission Tomography (PET), CT, US, and visible-light imaging [Asgari Taghanaki et al., 2021].

There exist traditional methods, such as thresholding and boundary extraction, as well as data-driven methods. Since traditional methods often require domain knowledge and do not scale well with heterogeneity and an increasing number of medical data [Du et al., 2020, Kar et al., 2021], DL methods that generally show higher accuracy and robustness are emerging [Kar et al., 2021].

Many directions to improve DL-based biomedical semantic segmentation have been explored in recent years. For example, for supervised and weakly-supervised training schemes, key innovations are summarized in the reviews by Hesamian et al., 2019 and Wang et al., 2022a. The review paper by Wang et al., 2022a describes the recent progress for supervised methods with improvements related to network architectures and loss functions. For weakly supervised approaches, it reveals that developments in data augmentation, transfer learning, and interactive segmentation were mainly addressed. According to Wang et al., 2022a, most works in medical image segmentation are on designing the network structure, with the encoder-decoder-based U-Net and its adaptations being the most commonly applied architectures. For example, deepening network blocks by residual layers, optimizing skip connections, increasing receptive fields by atrous convolutions, enabling learning on different scales by pyramid schemes, targeting feature extraction by attention mechanism, decreasing the number of parameters by designing different convolutions, and increasing the efficiency using graph convolutions are main advances from the last years.

Even though these innovations can improve the segmentation performance of a specific task, their applicability to a new task is often limited, as considered by Isensee et al., 2021. This work hypothesizes that not only the network architecture itself but also expert design choices, such as preprocessing of the data, data augmentation, and tuning of hyperparameters with regard to hardware conditions, training scheme, and postprocessing, highly influence the segmentation performance. Therefore, they proposed the *no new U-Net* (nnU-Net), which self-configures the mentioned design choices automatically given any new task or data without the need for expert knowledge. The design choices are implemented as three classes, namely, a set of fixed parameters, interdependent rules, and empirical decisions. The underlying network architecture is a vanilla U-Net, or more specifically, a 2D, 3D [Çiçek et al., 2016], and cascade version [Isensee et al., 2021] of it. The nnU-Net has shown excellent segmentation performance across a variety of biomedical segmentation challenges and is, therefore, applied in this thesis.

3.1.2. Semantic Segmentation in Photoacoustic Imaging

Semantic segmentation in PAI can be beneficial for numerous applications. For example, segmentation enables a refined image analysis in corresponding regions of interest and the inference of morphological parameters as shown for tumor vessels [Sun et al., 2020]. Additionally, optimizing optical and acoustic properties for different segmented tissue structures was shown to improve the reconstructed image quality [Liang et al., 2022, Lutzweiler et al., 2015].

Various publications, which can be divided into traditional and data-driven approaches [Le et al., 2022a], investigated methods for semantic segmentation of PA images. Traditional approaches are typically based on thresholding, image filtering, edge detection, and time series raw data analysis. To give some examples, automatic thresholding was demonstrated in the context of PA breast imaging [Zhang et al., 2018a] and PA microscopy in mice [Raumonen et al., 2018, Sun et al., 2020, Liang et al., 2022]. Through image filtering, such as vessel enhancement filtering [Frangi et al., 1998] and filtering-based edge detection, the vasculature of PA microscopy images [Yang et al., 2014, Mai et al., 2021] and body contours of tomographic PA images [Mandal et al., 2015] in mice could be segmented. The frequency-domain analysis enabled the spatial separation of different chromophores [Cao et al., 2017] and structures of different sizes [Moore et al., 2019] using Fourier transformations. In addition, time series data could be analyzed to segment main compartments of constant but different speeds of sound, which allowed improved PA image reconstructions [Lutzweiler et al., 2015]. Some works combined different traditional approaches [Sun et al., 2020] or directly compared them for their use cases [Yuan et al., 2020]. There are also methodologically more advanced approaches, such as the dynamic programming-based strategy for skin surface segmentation of PA microscopy images [Nitkunanantharajah et al., 2019].

Similar to the medical imaging domain, traditional approaches generally do not scale well with large numbers of heterogeneous images. In contrast, data-driven approaches learn to understand the inherent image features and properties of different segmentation classes from a set of training data. Both ML, such as k-nearest neighbor [Gonzalez et al., 2021] and random forest [Moustakidis et al., 2019] classifiers, and DL are on the rise for PA image segmentation [Gröhl et al., 2021b]. For DL, the U-Net [Ronneberger et al., 2015] is the most commonly used network architecture and applied, for example, for mouse contour segmentation [Lafci et al., 2020b], and vessel segmentation in simulation [Luke et al., 2019] and in vivo [Chlis et al., 2020] tomographic studies. Another network architecture used is a FCNN that was compared [Gerl et al., 2020] and combined with a U-Net [Yuan et al., 2020] in the context of PA microscopy images.

While there are more sophisticated methods, such as the approach proposed by Boink et al., 2019 that allows for both reconstruction and segmentation of vessels of tomographic PA images based on a learned primary-dual method, there has been no work to the time of this thesis that provides semantic multi-label segmentation of tomographic PA images in humans using DL.

3.2. Image Simulation and Synthesis

This section presents related work on image simulation and synthesis, particularly emphasizing content generation in computer vision and anatomy generation in medical imaging and PAI. It should be noted that the definitions for image simulation and synthesis by Frangi et al., 2018 are used within the thesis. Therefore, simulation refers to image/data generation based on modeled prior knowledge. For example, physical principles or organ physiology can serve as prior knowledge. Synthesis, on the other hand, refers to image/data generation based on learning phenomenologic models, for instance, by pattern recognition from a set of representative examples. While both definitions imply visually realistic and quantitatively accurate virtual images, simulations tend to be easier to control, and syntheses are frequently faster.

The first part of this section summarizes relevant related work concerning content generation in computer vision and, in analogy, tissue geometry generation in medical imaging. The second part explains the fundamental principles of image generation (cf. Figure 3.2.1) and the different concepts of tissue geometry generation in PAI.

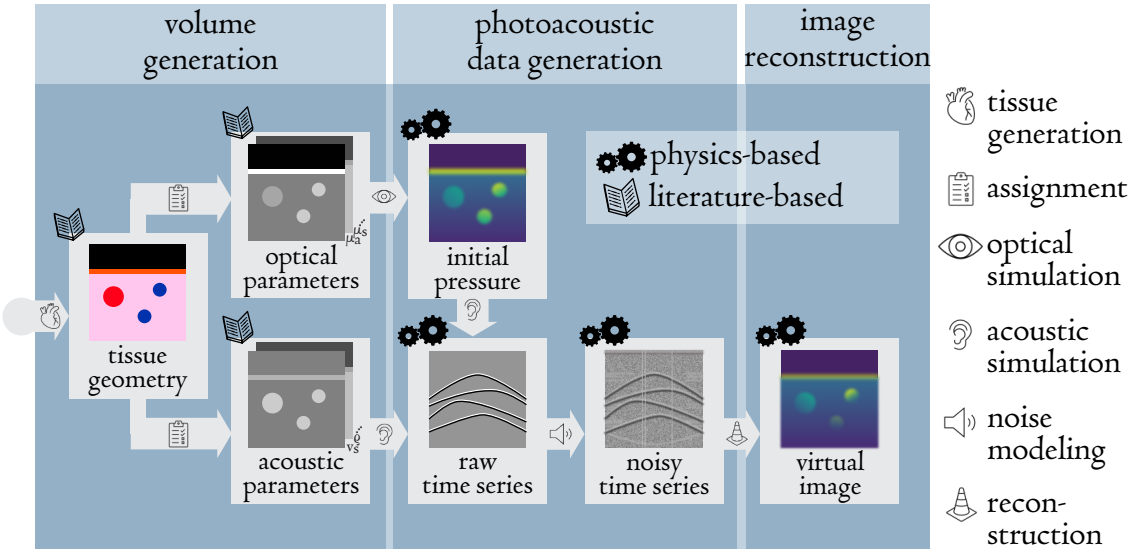

Figure 3.2.1.: Different components of the Photoacoustic Imaging (PAI) simulation pipeline as implemented in the toolkit for Simulation and Image Processing for Photonics and Acoustics (SIMPA). Based on the generated tissue geometries, optical tissue properties, such as the absorption coefficient μ_a and scattering coefficient μ_s, and acoustic tissue properties, such as the sound speed v_s and the density ρ, are assigned to the different tissue classes to perform the optical and acoustic simulations. Device-specific noise is added to the acoustic time series data that are reconstructed to a virtual Photoacoustic (PA) image representing the initial pressure distribution.

3.2.1. Image Simulation and Synthesis in Computer Vision and Medical Imaging

In recent years, remarkable progress has been made in generating virtual images in computer vision that offer image quality comparable to that of photographs [Rombach et al., 2022, Ramesh et al., 2023]. A current area of research relevant to this thesis's context is based on a disentanglement concept. More specifically, the image content and appearance, also referred to as style or texture, are considered separately. Different works showed that either one or both factors could be learned in a sequential [Zhu et al., 2017c, Esser et al., 2018, Park et al., 2019] or simultaneous manner [Li et al., 2021a, Zhang et al., 2021e, Pakhomov et al., 2021].

An outstanding contribution to this field of computer vision is the work by Kar et al., 2019, called Meta-Sim. In this approach, scene graphs were used to generate images whose content distribution matches that of the target images. More specifically, these scene graphs were generated by a probabilistic grammar commonly used in gaming and graphics. The nodes within these scene graphs represent different content objects, such as cars or roads in a traffic scene, and carry attributes describing, for example, the location and position of the objects. During the training process, the task was to optimize these attributes to minimize the content domain gap between the converted graph-to-image representations and the given target images. In other words, the underlying grammar-based structure of the scene graph was assumed to be correct and remained unchanged throughout the optimization process of the node attributes. A prerequisite for achieving optimization was the use of a differentiable graph-to-image converter. A key advantage of Meta-Sim is its flexibility to learn, per construction, the distributions of any number of attributes that match those of the real data, as demonstrated by its successful application to various data sets. Additionally, the optimization could be extended to simultaneously optimize a downstream task simultaneously.

Compared to computer vision, virtual images are at least as important in the field of medical imaging. For example, it allows image analysis and reconstruction algorithms to be understood, developed, and validated with available GT information. Additionally, large numbers of heterogeneous images that are often unavailable in the field can be easily generated for neural network training. Furthermore, virtual data can overcome data privacy hurdles. Research in this area is emerging, which is strongly related to the fact that the number of data-driven research projects and medical applications is increasing. In parallel, events such as the *Simulation And SyntHesis In Medical Imaging* (SASHIMI) workshop introduced in 2016 at the *Medical Image Computing*

and Computer Assisted Intervention (MICCAI) conference, one of the leading conferences in the field, are also increasing in order to disseminate current research.

In the specific context of tissue geometry generation, there are several works on the generation of digital phantoms of human anatomy via simulation, synthesis, and a combination thereof [Frangi et al., 2018, Segars et al., 2010]. For simulation, mathematical models defining tissue geometries following equations or simple geometric primitives are typically used [Segars et al., 2010]. For synthesis, semantic segmentations of data sets are usually provided as digital phantoms, such as the open-source visible human project [Ackerman, 1998] or the human forearm and hand data set [Kerkhof et al., 2018]. In hybrid approaches, the surfaces of segmented volumes are often modeled, for example, by B-splines or polygon meshes. This allows the accurate description of the individual structures and easy modification, as presented, for example, by the 4D extended cardiac-torso (XCAT) phantom [Segars et al., 2010]. Here, multimodal images could be simulated based on the virtual anatomy modified as required, which also accounted for cardiac and respiratory motion.

Similar to the disentanglement concept in computer vision, a notable advancement in the field of DL for medical image synthesis is the disentanglement of anatomical factors from remaining factors to image generation. There are different approaches to follow that principle [Yi et al., 2019]. For example, anatomical priors such as segmentation masks [Costa et al., 2017b, Unberath et al., 2018, Pham et al., 2020, Rusak et al., 2020] or semantic features [Xu et al., 2019] are leveraged as a basis for image synthesis. In addition, some studies have explicitly disentangled the different factors in a latent space [Chartsias et al., 2019, Li et al., 2019]. Moreover, researchers have explored modifying tissue geometries either directly in the image space [Shin et al., 2018] or in a latent space [Oliveira et al., 2018, Joyce et al., 2019].
Building on this foundation, research has turned to the generation of entirely new tissue geometries [Costa et al., 2017a, Guibas et al., 2017, Li et al., 2020a, Tudosiu et al., 2022]. These efforts are closely aligned with the objective of this thesis. A highly relevant study by Li et al., 2020a demonstrated that generative models can model tissue geometries and material maps that are subsequently used as input for physics-based simulations. The tissue geometry generation model was constrained by a statistic shape model with a reduced set of parameters to ensure the plausibility of the generated tissue geometries. Their training was implemented in a federated fashion, allowing data from multiple sites to be used and adapted to site-specific characteristics by training another neural network.

3.2.2. Image Simulation and Synthesis in Photoacoustic Imaging

Data-driven solutions to the problem of quantifying PA images usually require training data with GT tissue properties. Since there is no gold standard method that provides these quantities in vivo, virtual PA image data supplying a priori GT tissue properties are essential [Gröhl et al., 2021b].

In general, a lot of work exists on PA image simulation, which consists of several steps, and each of them needs careful consideration, as shown in Figure 3.2.1. The typical simulation workflow involves the following steps:
Tissue geometries are generated, such as 3D layers representing skin or tubular structures representing blood vessels. Then, optical and acoustic tissue properties are assigned to different tissue classes. Typically, these properties are either chosen randomly within certain boundaries or according to literature values [Gröhl et al., 2021b].
The optical forward model can be applied based on these optical parameter images. There are two primary strategies. On the one hand, photon propagation in turbid media can be approached by statistics-based MC-based methods [Jacques, 2014, Fang et al., 2009, Leino et al., 2019]. On the other hand, photon propagation can be accomplished by analytical methods that approximate the RTE, for example, by leveraging the diffusion equation or finite element methods [Schweiger et al., 2014, Dehghani et al., 2009].
Given the initial pressure distribution as the output of the optical forward model and the acoustic parameter images, acoustic modeling can be performed, which requires the solution of the partial differential PA wave equation. PDEs are typically solved by finite-difference, finite-element, or boundary-element methods. However, the time domain modeling with these conventional methods can become cumbersome and slow with broadband or high-frequency waves, as in the case of PAI. Therefore, the PA wave equation is typically solved by k-space pseudo-spectral methods, as implemented in k-Wave [Treeby et al., 2010]. k-Wave is the most commonly used open-source MATLAB toolbox and is specifically designed for time domain simulation and reconstruction of PA wave fields in tissue-realistic media accounting for spatial heterogeneities of acoustic properties, namely the sound speed, density, and acoustic absorption [Treeby et al., 2010].
Usually, a noise model is designed to match the Signal-to-Noise Ratio (SNR) ratio of a specific PA device and then added to the time series data [Dehner et al., 2022a].
Finally, the simulated PA images are obtained by reconstructing the raw data, which represents the initial pressure distribution and for which various methods exist [Xu et al., 2005, Park et al., 2008, Matrone et al., 2014, Grün et al., 2007, Hauptmann et al., 2018, Xu et al., 2002].

Several open-source tools are available to the PA community that address individual components of the simulation pipeline [Jacques, 2014, Fang et al., 2009, Leino et al., 2019, Treeby et al., 2010, Else et al., 2023, Gröhl et al., 2023a]. Some frameworks combine different simulation components into one toolkit [Sowmiya et al., 2017, Fadden et al., 2018].
The open-source toolkit for Simulation and Image Processing for Photonics and Acoustics (SIMPA) [Gröhl et al., 2021a] offers a simple and modular way to assemble different open-source computational models, data processing algorithms, and digital device twins properly, which is why it was applied in this thesis.

While the tools for optical and acoustic simulations are well established in the field, the generation of tissue geometries, which are the basis for PA simulations, is approached in a variety of ways in the literature. After analyzing 134 research papers on DL-based PAI spanning from January 2017 to June 2023, seven categories could be identified for modeling tissue geometries, as shown in Figure 3.2.2. A list of the categorized papers can be found in the Supplemental Material D.
Most of the papers used random geometric shapes as tissue geometries. More specifically, point, circular, elliptical, or rectangular shapes were randomly placed on a homogeneous background. Some papers used pattern phantoms, such as the Shepp-Logan and Derenzo phantom or logos, as the basis of the tissue geometries. While these concepts of tissue geometry generation are simple to implement, they are a poor approximation of human tissue.
Since vessels are generally easily visible in PAI, a significant amount of literature bases tissue geometries on model-based or segmentation-based vasculature, usually placed on a homogeneous background. For example, mathematical models, such as a Lindenmayer system or segmentations from open-source data of other imaging modalities, such as CT or MRI, were exploited to generate vascular trees.
Compared to these concepts, the fraction that models the entire tissue structure is smaller. However, one can again distinguish between model-based and segmentation-based approaches. A model-based example is the generation of forearm-specific tissue geometries based on literature knowledge, as used in this work, or on empirical observations of measured data [Susmelj et al., 2022]. Segmentation-based examples are numerical phantoms that are again obtained from other imaging modalities, such as the Optical and Acoustic Breast Phantom Database (OA-breast) [Yang et al., 2019a] and an open-source brain phantom [Li et al., 2022a], both based on MRI.

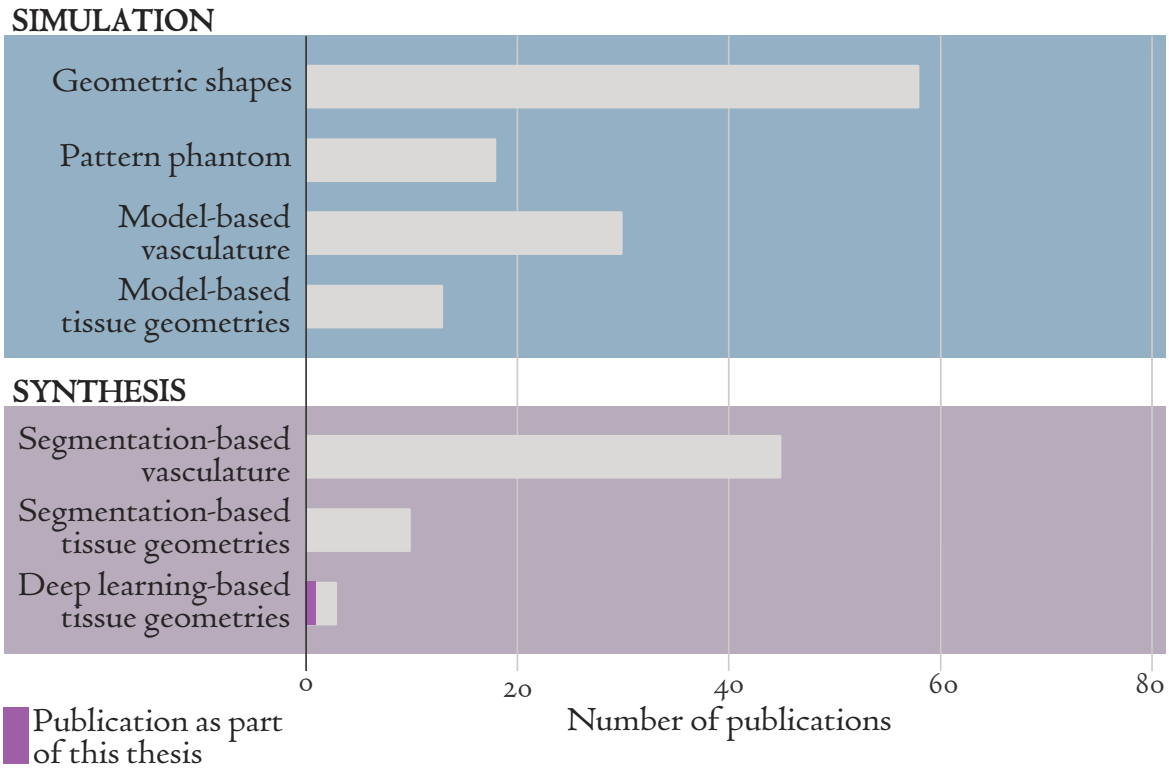

Figure 3.2.2.: Comparative analysis of concepts for modeling tissue geometries used in more than 100 DL-based publications between January 2017 and June 2023. The definitions of simulation and synthesis were chosen according to the work by Frangi et al., 2018. The search string on Google Scholar was defined as ("Deep Learning" OR "Neural Network") AND ("Photoacoustic" OR "Optoacoustic") as in previous work by Gröhl et al., 2021b. These initially over 300 papers were refined based on screening of abstracts and figures, if available, resulting in 217 publications. Only papers in the area of DL-based PA imaging/angiography/microscopy were included in this figure, i.e., pure spectroscopy papers were excluded. Scientific letters were also included unless they were below one page in length. The peer-reviewed version was preferred if there were multiple versions of a paper. Six publications could not be accessed, so either another version was used or the respective publication was excluded. The final papers were analyzed with respect to the use of simulated data as detailed in the Supplemental Material D. Network training relied on simulated data in $\sim 60\,\%$ ($N = 134$). This subset was further analyzed with respect to seven identified categories for tissue geometry generation.

Depending on the research questions addressed in these publications, these mostly unrealistic concepts may be appropriate for their specific application to in vivo data. For example, binary vasculatures derived from MRI may be sufficient to enable DL-based reconstruction of undersampled microscopic PA images. However, in some cases, the underlying concept for tissue geometry generation is co-decisive for in vivo application of the algorithms. For example, a

quantification algorithm trained on data with random geometric shapes could not yet be applied to tomographic PA images [Kirchner et al., 2018a]. Numerical phantoms based on different imaging modalities can also be challenging since measurement-related influences, such as the pressure of the hand-held probe on the tissue, may alter the structures in the tissue and should be considered. In summary, it is noticeable that PA research has not yet focused on a realistic conceptualization of the anatomy under investigation (whole tissue structure) and realistic training data for DL-based approaches in general.

This is also reflected in the fact that after the first publication of a preprint of this work in 2021, only two other papers have been published dealing with the automatic synthesis of PA images. Both of these works are based on a GAN and were published within the last year.
The work by Ma et al., 2022 allowed the augmentation of PA images to boost the performance of a super-resolution network. Here, a *BicycleGAN* [Zhu et al., 2017b] was used to augment PA images. By conditioning the BicycleGAN on a given resampling mask, an input image could be translated into a new image that resembled image characteristics from the input image and structural features of the resampling mask.
The work by Bench et al., 2023 dealt with generating realistic training data to enable DL-based quantification of PA images. The principle leverages an *ambient GAN*. First, a DL-based optical model was trained with simulated data, which was then included in the ambient GAN. The spatial distribution of absorption coefficients could thus be learned using the ambient GAN by converting the generated absorption estimates into PA images through the DL-based forward model and comparing them to real PA images through a discriminator.

While both works are closely related to this thesis, they address tissue geometry generation only implicitly. In Ma et al., 2022, realistic tissue geometries were only used as conditioning of a GAN-based strategy to augment images. In other words, the tissue geometries themselves were not modified and solely used several times. On the other hand, Bench et al., 2023 went one step further than this thesis by directly learning the spatial distribution of the absorption coefficient, which, in principle, also represents the underlying tissue geometries. However, this more unconstrained problem led to highly noisy and artifact-corrupted absorption maps and did not allow the generation of distinct tissue geometries.

4. Contributions

This chapter presents the contributions developed for DL-based modeling of realistic tissue geometries for PA image analysis. First, the acquired and manually annotated PA images used in this work are described (cf. Section 4.1). This is followed by a detailed description of the three approaches, each of which addresses one of the three research questions RQ1 - RQ3 (cf. Sections 4.2 - 4.4).

4.1. Photoacoustic Data

The three approaches were all based on acquired in vivo PA data or patterns derived from that. This section describes the PAI device used for the acquisition and details of the recorded healthy volunteer data (cf. Section 4.1.1) as well as specifics of image processing (cf. Section 4.1.2) and image annotation (cf. Section 4.1.3).

4.1.1. Image Acquisition

Disclosure to this work:
Lena Maier-Hein and Janek Gröhl initiated the idea of the healthy volunteer study. Lena Maier-Hein supervised the entire study and provided valuable feedback on various steps of the process. She and Janek Gröhl were the main contributors to the application for ethics approval. The acquisition of the data used in this thesis was mainly performed by Janek Gröhl, Kris K. Dreher, Niklas Holzwarth, Jan-Hinrich Nölke, Minu D. Tizabi, and myself. The manual annotations were performed by Janek Gröhl, Andrei Cosmin Siea, and myself following the annotation protocol that was part of the *Photoacoustics* journal publication by Schellenberg et al., 2022b.

Photoacoustic imaging device

The in vivo data used in this thesis was acquired with the Multi-Spectral Optoacoustic Tomography (MSOT) Acuity Echo device, iThera Medical, Munich, Germany (cf. Figure 4.1.1). The MSOT device is a commercially available solution for simultaneous PA and US imaging. The bimodal acquisition is achieved by integrating an optical light source into an US system. More precisely, the MSOT device allows laser pulses in the near-infrared range from 660 nm to 1300 nm to be sent into the tissue, which leads to a thermoelastic expansion after absorption of the laser energy by different chromophores. The resulting acoustic waves can be detected with the US transducer, which has a center frequency of ~ 4 MHz and a bandwidth of 55 %, assuming a Gaussian distribution. Using the same transducer, the MSOT allows for typical US imaging by transmitting US waves and receiving their reflections.

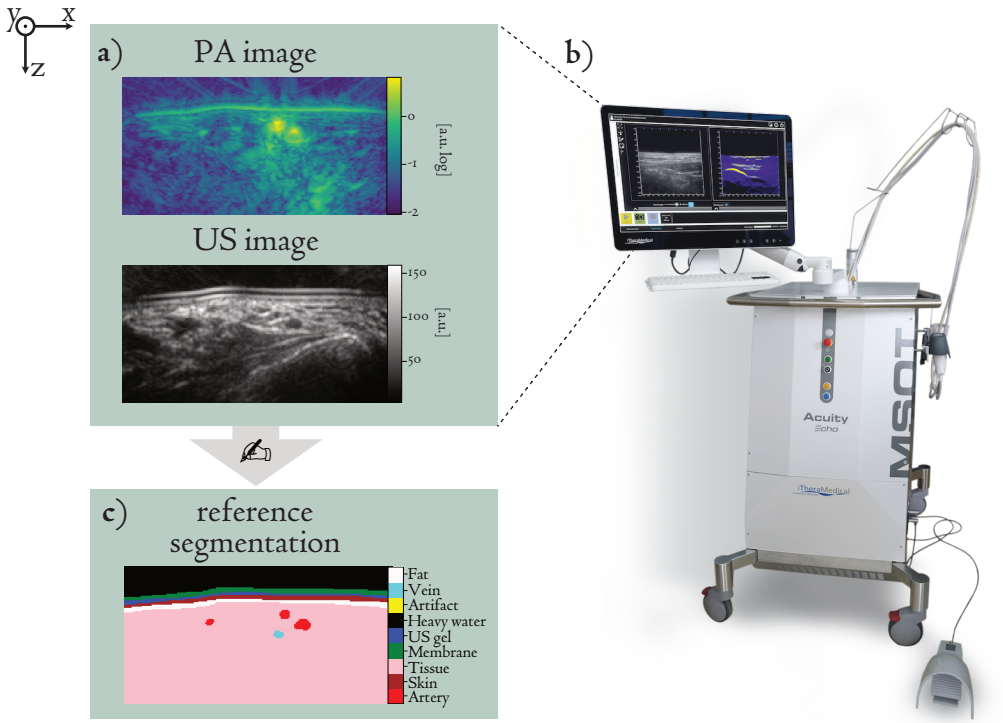

Figure 4.1.1.: Example (**a**) pair of Photoacoustic (PA) and Ultrasound (US) images acquired with the (**b**) Multi-Spectral Optoacoustic Tomography (MSOT) Acuity Echo device [iThera Medical GmbH, 2021] and their corresponding (**c**) manual annotation. In addition, the directions of the x-, y-, and z-axes are defined.

During imaging, PA and US images are recorded sequentially. First, a laser pulse of a specific wavelength is transmitted into the tissue, producing one PA image. Then, a laser pulse of a different wavelength produces the subsequent PA image, and so on. After typically five PA

images, an US image is recorded, and the workflow is repeated. The US acquisition frequency is 5 Hz, meaning that the PA repetition rate is up to 25 Hz, depending on the number of wavelengths.
The MSOT device saves the reconstructed PA and US images as well as the raw PA data. In addition, metadata, such as the laser pulse energy specified with a maximum pulse energy of 25 mJ at 750 nm by the vendor, is made available.

The reconstructed PA and US images display not only the imaged tissue but also parts of the MSOT device's handheld probe (cf. Figure 4.1.1). These components include the coupling fluid, the membrane of the probe, and the US gel used for coupling between the probe and the skin. The 1 mm thick mediprene membrane covers the probe's region in contact with the skin. The coupling fluid fills the gap between the detector elements and the membrane and is made primarily of heavy water. The manufacturer chose heavy water for its low absorption and scattering properties in the near-infrared window.

Healthy volunteer data

In an explorative pilot study involving healthy volunteers, in vivo data was acquired with the MSOT device. The study's primary goal is to build a database that enables the analysis of PA image features, which is of particular importance for the development of data-driven methods. The ethics of the study were approved by the committee of the medical faculty of Heidelberg University under reference number S-451/2020, and informed consent was obtained from all subjects before measurements. Additionally, the study is registered with the German Clinical Trials Register under reference number DRKS00023205.

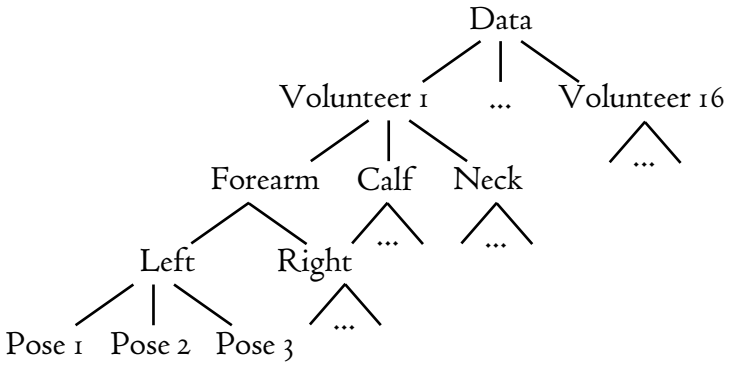

Figure 4.1.2.: Hierarchical structure of data. Each healthy human volunteer was imaged at the forearm, calf, and neck on both the left and right sides of the body and at three distinct locations each.

For this thesis, the data from 16 healthy human volunteers older than 18 years of any gender and human skin color was used. The forearm, calf, and neck area of each subject were imaged with both US and PA imaging. These sites were chosen because they are easily accessible and clearly show vessels that are generally important for many potential photoacoustic applications. At other sites, vessels are often not comparatively this superficial. In more detail, three distinct locations on the left and right sides of the forearm, calf, and neck were imaged freehand and as statically as possible for approximately 30 s with the MSOT device, which leads to a hierarchical structure of the data (cf. Figure 4.1.2). Similar to Chlis et al., 2020 and allowing the data to be used for various applications, the PA images were acquired using 26 wavelengths equidistantly selected between 700 nm and 950 nm yielding $N = 288$ pairs of multi-spectral PA and US images.

4.1.2. Image Processing

The image processing of the acquired data was different for the PA and US data since the vendor does not provide US raw data. Instead, the vendor-provided reconstruction based on a proprietary backprojection algorithm delivered 2D US images. The 2D PA images were reconstructed from the available raw data using a custom implementation of the delay-and-sum backprojection algorithm [Kirchner et al., 2018b] within the Medical Imaging Interaction Toolkit (MITK) [Nolden et al., 2013].

Four-step post-processing was performed on the PA images (cf. Figure 4.1.3):

1. The multi-spectral PA images were divided by the wavelength-dependent laser pulse energy of the MSOT device to account for laser energy variations.

2. Since the fields of view of the reconstructed PA and US images differed in depth by the use of two different algorithms, the PA images were co-registered to the US images. The offset was manually determined using the skin signal, which is clearly visible in both imaging domains. All PA and US images were cropped in depth according to the established offset to have a height of 2 cm.

3. Resampling the images resulted in an isotropic resolution of 0.16 mm.

4. The images were averaged to increase the SNR. For this, the sequence of PA and US image pairs was divided into four stacks, each corresponding to approximately eight

seconds. The image pairs of every stack were pixel-wise averaged, resulting in four aggregated image pairs. The image pair with the sharpest edges calculated as the averaged image gradient in the US image was chosen as the final pair of PA and US images.

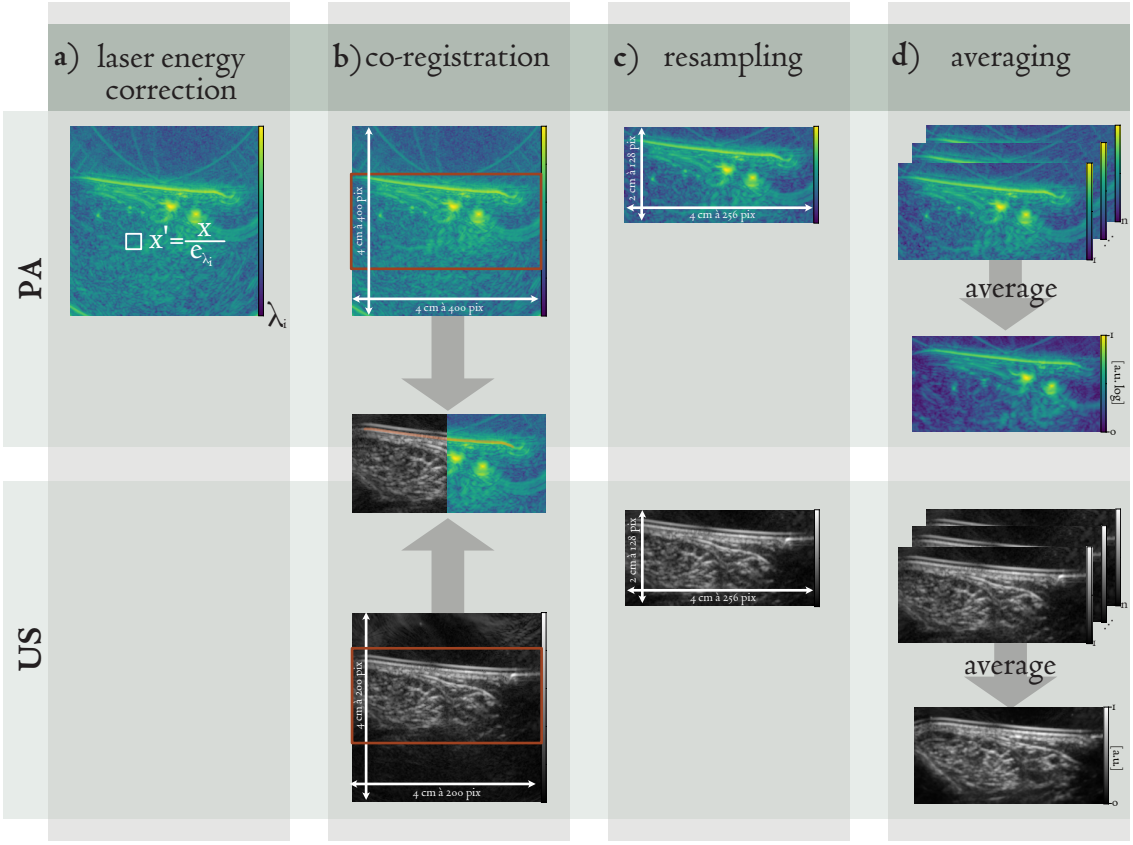

Figure 4.1.3.: Four-step data processing pipeline. (**a**) For every wavelength λ_i Photoacoustic (PA) images were divided by the laser energy e_{λ_i}. (**b**) The PA and Ultrasound (US) images were co-registered using the skin signal visible in both imaging domains to determine a global offset. Using this offset, the images were cropped to have a height of 2 cm. (**c**) The images were resampled to have an isotropic spacing of $\Delta x = \Delta z = 0.16$ mm. (**d**) The image pairs were divided into four stacks and averaged pixel-wise. The stack with the sharpest edges in the US image on average was chosen as the final pair. Note that PA images are shown in log-scale, and all images are min-max normalized.

4.1.3. Image Annotation

According to recommendations by Mongan et al., 2020, the PA and US image pairs were manually annotated in semantic segmentation masks by one out of three domain experts following a

standardized and detailed annotation protocol (cf. Supplemental Material A). In other words, every pixel of an image was classified as one out of nine annotation classes: (1) *artery*, (2) *skin*, (3) background *tissue*, (4) *US gel*, (5) transducer *membrane*, (6) *heavy water* of the transducer head, (7) *coupling artifact*, (8) *vein*, and (9) subcutaneous *fat*. An example pair of PA and US images and its corresponding annotation are shown in Figure 4.1.1. The background tissue class refers to the tissue below the fat layer, excluding the vessels. Among others, it comprises muscle and conjunctive tissue. The coupling artifact class was introduced because some measurements suffered from an insufficient coupling of the transducer to the skin, resulting in a signal loss at the images' edge.

Both the US and PA signals were considered for annotation. However, because the visual appearance for some tissue classes differs between the US and PA domains, one domain was defined as decisive for each tissue class. In general, tissue classes were distinguished by characteristics describing their spectral behavior, their US signal relative to the US signal of the background tissue or their spatial location.

4.2. Tissue Geometry Estimation with Neural Networks

Addressed Research Question RQ1:

Can discriminative neural networks be leveraged to extract tissue geometries from real PA images via automatic semantic segmentation?

This section presents the work addressing RQ1. For this purpose, two types of neural networks with input data of different granularities were investigated. The overall concept of this approach is explained in Section 4.2.1, and Section 4.2.2 provides the related material and methods. The experiments and corresponding experimental conditions are explained in Sections 4.2.3 and 4.2.4, respectively. Section 4.2.5 presents the corresponding results that are discussed in Section 4.2.6.

Disclosure to this work:
The idea for semantic segmentation of PA and US images originated from Janek Gröhl. He and Kris. K. Dreher implemented the very first experiments under the supervision of Lena Maier-Hein. I joined them and continued to work with Janek Gröhl on the experiments. Lena Maier-Hein supervised the entire project and offered valuable feedback and guidance throughout its various phases to ensure its successful completion. This work led to a poster at the Photons Plus Ultrasound meeting of the SPIE Photonics West conference in 2021 with Janek Gröhl and myself as joint first authors [Gröhl et al., 2021d]. I took over the project and conducted additional experiments. During this phase, in addition to Lena Maier-Hein, Kris K. Dreher, Alexander Seitel, Niklas Holzwarth, Annika Reinke, and especially Janek Gröhl were extremely supportive and invaluable for detailed discussions and excellent feedback. Patricia Vieten and Niklas Holzwarth verified the reproduction of the project's results. The re-annotations were performed by Janek Gröhl, Alexander Seitel, Niklas Holzwarth, Kris K. Dreher, and myself. I greatly appreciate all the support. The work was published in the journal *Photoacoustics* [Schellenberg et al., 2022b] and the content, Figures 4.2.2- 4.2.6, and Tables 4.2.1- 4.2.2 are taken (partly modified) from this publication with permission.

4.2.1. Concept Overview

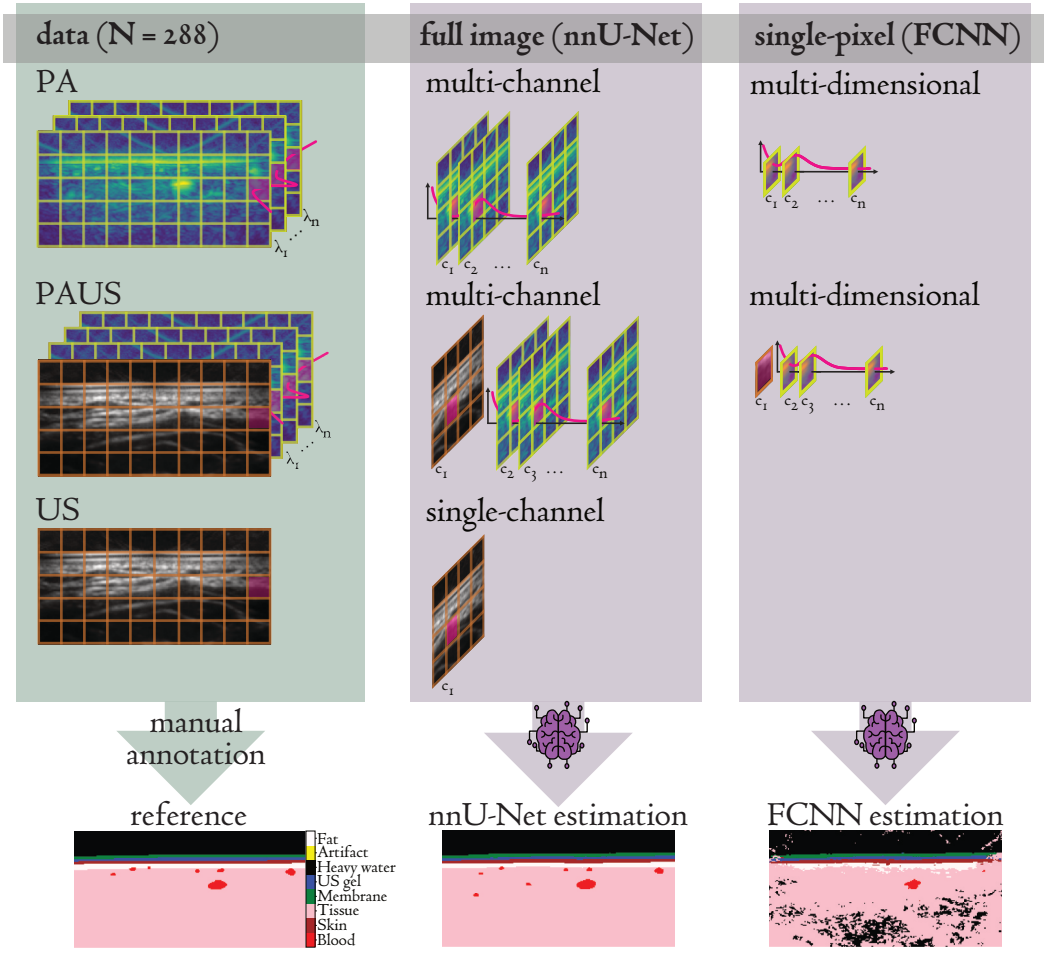

Figure 4.2.1.: The concept of estimating tissue geometries from PA images is based on two neural networks with input data of different granularities, the nnU-Net [Isensee et al., 2021] and a Fully-Connected Neural Network (FCNN). The benefit of additionally using US images, which is referred to as PAUS images, is investigated for both networks. Furthermore, the nnU-Net is trained on only US images.

The concept leveraged to assess the feasibility of tissue geometry estimation from PA images is based on two popular neural network architectures with data inputs of different granularities, both allowing semantic segmentation. In more detail, whole image information or single-pixel spectral information is used as the input of a U-Net [Ronneberger et al., 2015] or a FCNN, respectively. By design, the U-Net leverages local context by inherent convolutional kernels, in contrast to the FCNN, which only relies on single-pixel spectral information in exchange for more training data. This concept additionally investigates whether PAUS images provide an added value in terms of segmentation performance compared to networks trained on PA

images exclusively. A direct comparison with a network based on US images alone is studied for the U-Net since no tissue class-specific information is expected in the single-pixel signal intensities of the US images (cf. Figure 4.2.1).

4.2.2. Material and Methods

To accomplish the concept for semantic segmentation of PA, US, and PAUS images, the manually annotated data was used, and a U-Net and FCNN were implemented. Specifications of the reference data and the networks are provided in this section.

Data

The manual annotations described in Section 4.1.3 were leveraged for this approach. Note that the annotation classes artery and vein were combined into the class *blood*. The data was split into training/validation and test sets, considering the underlying hierarchical structure (cf. Section 4.1.1). The images of ten volunteers were randomly selected as training/validation data, and six volunteers' remaining images (N = 108) were chosen as a held-out test set. To enable five-fold cross-validation, the training/validation set was split into five subsets of two randomly selected subjects each. In other words, each of the five subsets was used exactly once as the validation set (N = 36), and the remaining data was used as the training set (N = 144). Note that this unconventional data split was deliberately chosen to allow statistical analysis on the test set with $N > 5$ subjects while still allowing for a sufficient amount of training data.

To better assess the performance of the networks and the annotation quality, a human annotation reliability study for the clinically highly relevant class blood [Attia et al., 2019] was carried out. Ten test images of one human volunteer were randomly selected, requiring at least one image of every body region and body side. Five domain experts annotated the ten images, and the resulting annotations were assessed with respect to the original annotations.

U-Net

The U-Net was chosen to consider the local context in the semantic segmentation. This was motivated on the one hand because, according to [Ronneberger et al., 2015] and compared to other network architectures, it is well suited for medical applications, requires less training data, and also provides less blurry results. On the other hand, the U-Net was most commonly and successfully used for PA image analysis according to a recent literature review [Gröhl et al., 2021b].

A genuinely outstanding framework for semantic segmentation in the biomedical field using the U-Net is the nnU-Net [Isensee et al., 2021], which outperformed a range of international biomedical segmentation challenges by a large scale. By comparing different modifications of the classical U-Net architecture, the authors hypothesize that not a complex network architecture but rather key design choices, such as the patch size, the augmentation strategy, and the method for ensembling of folds, are essential for achieving good performance. The main feature of the nnU-Net is having these design decisions modeled as a set of fixed parameters, interdependent rules, and empirical choices and being fully self-configured.

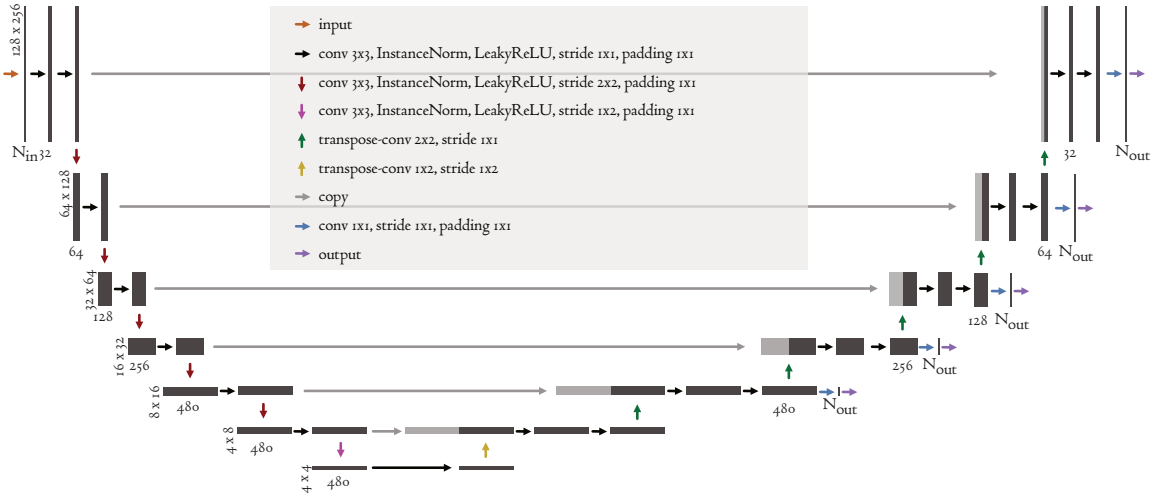

Figure 4.2.2.: Network architecture of the 2D nnU-Net [Isensee et al., 2021]. The input Photoacoustic (PA) and/or Ultrasound (US) images (N_{in} = 26/27/1) of size (256 x 128 px) are processed by different convolutional, normalization, activation, transpose convolutional, and copy operations to enable the one-hot encoded semantic segmentation outputs with N_{out} = 9 classes and the same size as the input images.

According to the best initial results performed on our validation data, the 2D nnU-Net configuration was chosen. This U-Net architecture closely follows the one of the original publication [Ronneberger et al., 2015]. However, minor changes, such as using strided convolutional downsampling layers instead of max-pooling layers, were implemented. Further details are reported in the work by Isensee et al., 2021. Figure 4.2.2 shows a schematic of the implemented architecture. The input size corresponded to the whole image size (256 x 128 px), and the multi-spectral PA image nature was designed as multi-channel input. In other words, the number of channels was set to one for US data, 26 for PA data, and 27 for PAUS data. As this project aimed to segment eight tissue classes, the output size was designed as the full image size with

nine output channels. This enabled a one-hot encoded representation of the eight tissue classes and one background class that was required due to the augmentations. Five-fold cross-validation was applied for training the nnU-Net. At inference, the corresponding five estimations were ensembled.

Fully-Connected Neural Network

The most straightforward network for single-pixel information is a FCNN. In the context of PAI, a FCNN has already been successfully proposed, for example, among others, for oximetry [Gröhl et al., 2021c, Gröhl et al., 2021b]. Here, the network architecture was based on previous work by Gröhl et al., 2021c. It consisted of an input layer of single-pixel size (1 × 1 px). In analogy to the U-Net implementation details, the input layer came with N_{in} dimensions that corresponded to N_{in} = 26 for PA data and N_{in} = 27 for PAUS data. The output layer was of single-pixel size with N_{out} = 8 dimensions according to the eight tissue classes. As shown in Figure 4.2.3, the network consisted of five layers. First, a fully-connected transition and a TanH activation were used to upscale the input layer to a hidden layer of the dimension of $4 \times N_{\text{in}}$. Then, four layers were calculated with fully-connected transitions, LeakyReLU activations, and dropout layers (20 %) followed. A final fully-connected transition downscaled the last hidden layer to the output size. As the nnU-Net, five-fold cross-validation was applied for training the FCNN, and the inferred estimations were ensembled.

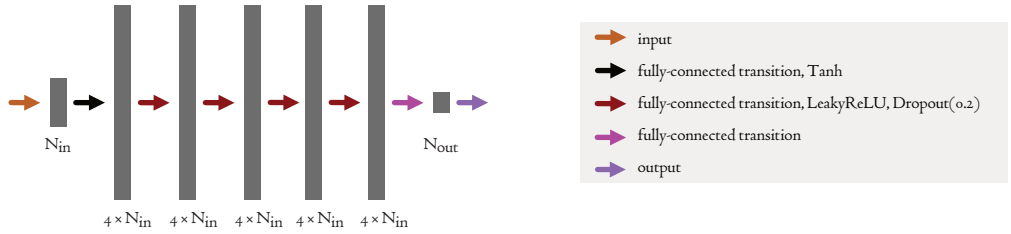

Figure 4.2.3.: Network architecture of the Fully-Connected Neural Network (FCNN). The single-pixel spectral information of Photoacoustic (PA) or a combination of PA and US (PAUS) images (N_{in} = 26/27) is fed into the network. The network consists of 5 layers, each of 4 x N_{in} dimensions. The six transitions are fully-connected with a Tangens Hyperbolicus (TanH) (first transition only) or Leaky Rectified Linear Unit (LeakyReLU) activation function. The five center transitions additionally have a dropout operation (20 %). The last transition is only fully-connected without activation and calculates the single-pixel output of N_{out} = 8 dimensions according to eight tissue classes.

4.2.3. Experiments

Two experiments were performed to assess the feasibility of automatic tissue geometry estimation from PA, US, or PAUS images with discriminative networks (RQ1). The first experiment is considered the *baseline*, and the second explores the *robustness* to measured body regions.

Baseline experiment

For the first *baseline* experiment, data from all body sites (forearm, calf, and neck) was leveraged to train each of the two network types, namely the nnU-Net and FCNN. As described in Section 4.2.1 and to estimate the added benefit of US data, both networks were trained on PA data and PAUS data. In contrast to the FCNN, the nnU-Net was additionally solely trained on US data. At inference, the networks were tested on the held-out test set, including all body sites.

Robustness experiment

For the second *robustness* experiment, the first experiment was repeated, but with training data from two body regions and test data from the remaining body region. Thus, the *robustness* of the models to morphologically different test data was investigated. The experiment was performed for all three data combinations from different body sites. In other words, the nnU-Net and FCNN were trained on (A) neck and calf, (B) forearm and neck, and (C) forearm and calf images of the training set. The estimations at test time were performed on (A) forearm, (B) calf, and (C) neck images of the test set.

4.2.4. Experimental Conditions

Hardware and training specifications of the nnU-Net framework and the FCNN used for both experiments are provided in this section. Additionally, the validation strategy to assess the performance of the different network and data configurations is presented.

Computing resources

All training was performed either on a Ubuntu 18.04 workstation with an Intel(R)Core(TM) i7-8700 processor (6 cores) and 64 Gigabyte (GB) Random Access Memory (RAM) or on a node with an Intel(R)Xeon(R) E5-2620V4 processor (8 cores) and 188 GB RAM of an in-house Graphics Processing Unit (GPU) cluster. In both cases, an Nvidia 2080 RTX Ti graphics card

with 10.7 GB RAM was used. The inference, pre- and post-processing was performed on the workstation.

nnU-Net configuration

The nnU-Net did not require hyperparameter optimization, as it is a self-configuring framework. The 2D nnU-Net trainer (version two)[1] with default settings was applied. Correspondingly, the loss was calculated as the sum of CE and Soft Dice losses. The Soft Dice loss was calculated per minibatch using a smoothing factor ϵ_{smooth} of $1 \cdot 10^{-5}$. Additional information about configurations can be found in the work by Isensee et al., 2021.

Fully-Connected Neural Network configuration

Before training the FCNN, the data was z-score normalized to the mean and standard deviation of the training data set. The FCNN was implemented in PyTorch [Paszke et al., 2019] and trained with the Soft Margin loss (cf. Section 2.2.3). In this context, the zero values of the one-hot encoded labels were set to -1. The hyperparameters were optimized with a grid search using the validation loss of the baseline experiment. The network was trained for 200 epochs with the Adam optimizer [Kingma et al., 2014] and a minibatch size of $1 \cdot 10^4$. For the training of one epoch, $1 \cdot 10^3$ minibatches were used. Before performing the optimizer step, the network parameters were clipped at ± 1. A learning rate of $5 \cdot 10^{-5}$ was used. The learning rate decayed with respect to the improvement of the validation loss. In particular, the ReduceLROnPlateau scheduler provided by PyTorch was applied with a patience of 40 epochs and a factor of 0.5.

Performance assessment

The feasibility of the tissue geometry estimation algorithms was assessed by calculating segmentation performance metrics on the estimated and reference test tissue geometries. Following the recommendations by Maier-Hein et al., 2022, the overlap-based DSC[2] and distance-based NSD metrics (cf. Section 2.2.3) were applied. In particular, to account for imbalances in the number of pixels of the different tissue classes, they were calculated for each test image and tissue class. The DSC and NSD metrics were not computed if the tissue class was missing in the reference. The NSD was also not computed if a tissue class was not estimated. Aggregation of the individual metric results was performed considering the underlying hierarchical structure of the data. First, the class-specific metric values were averaged per test image, resulting in class

[1] `https://github.com/MIC-DKFZ/nnUNet/blob/nnunetv1/nnunet/training/network_training/nnUNetTrainerV2.py`

[2] `https://docs.monai.io/en/latest/metrics.html`.

averages. Second, the per-test image results were aggregated to give one metric-specific overall score for each tissue class and the class average. The tissue classes *skin* and *blood* were defined as target classes, as they often become crucial for various photoacoustic applications, especially because of their high absorption [Gröhl et al., 2021b]. The results section, therefore, provides the outcomes of these target classes and the class averages. To provide a comparison between the developed algorithms, rankings were generated using the challengeR toolkit[3] [Wiesenfarth et al., 2021]. Further details about the NSD configuration, the ranking, and the analysis of the inter-rater reliability study follow.

Normalized Surface Distance configuration For the NSD, the tolerance values for all tissue classes, excluding the class blood, were set to the most critical value $\tau = 1$ px. In contrast, the value for the class *blood* was chosen based on the inter-rater reliability analysis, as recommended in the work by Nikolov et al., 2021. For this, the average nearest neighbor distances (surface distance) between the surface of the original blood annotations and the re-annotated ones were calculated for each image and annotator. To aggregate the results considering the small amount of data and the underlying hierarchical structure, a linear mixed model, as described, for example, in the work by Roß et al., 2023, was applied. The body site was set as a fixed effect, and the image identifier and annotator were set as random intercepts. The resulting NSD tolerance value was $\tau = 5$ px chosen as the intercept of the fitted model.

Ranking To provide a comparison between the algorithms, the challengeR toolkit [Wiesenfarth et al., 2021] was applied. This toolkit helps analyze and visualize benchmarking experiments. In particular, rankings are generated based on user-specific choices of aggregation methods and different calculation schemes. Special emphasis is placed on the stability of the rankings, e.g., concerning the use of different aggregation methods or the number of data.
For this work, the DSC was chosen as the primary metric to compare the algorithms with the challengeR toolkit. More specifically, a ranking was performed for the target classes and the class averages (leading to three analyses) per volunteer and body site level. Hence, the DSCs were aggregated accordingly for each analysis, yielding $N = 18$ test cases each. Following the work by Winzeck et al., 2018, a rank-then-aggregate approach was applied to the test cases. The first step was to rank the 18 values per algorithm. As a second step, the corresponding ranking results were averaged to determine the final ranking.

[3] `https://github.com/wiesenfa/challengeR`

Inter-rater reliability analysis To relate the performance results to the inter-rater reliability, the DSC for the *blood* class was calculated for the re-annotations with respect to the original annotations for the ten test images. The results were aggregated using a linear mixed model analogous to calculating the surface distance. The mean and standard deviation of the fitted model's intercept were used to measure the overall human performance.

4.2.5. Results

Qualitative and quantitative results for both the *baseline* and *robustness* experiments are provided in this section. More specifically, example estimations and the analyzed metric results for the target classes and the class average are shown.

Baseline experiment

Figure 4.2.4 shows example results of the tissue geometry estimation for both the nnU-Net and FCNN. The network estimations for all data configurations resemble the reference tissue geometries. However, compared to the nnU-Net, the FCNN results show more noise in the estimated labels. Here, images were selected according to the median blood DSC with a minimum of 60 blood pixels. The improved performance of the nnU-Net is also reflected in the calculated overall performance scores for the class average and the target classes aggregated over all test images (cf. Table 4.2.1). For example, the class average DSC values for the nnU-Net and FCNN trained on PAUS images are 0.85 and 0.66, respectively. For the NSD, the respective values are 0.89 and 0.61.
Moreover, the performance scores demonstrate that for the blood class, the multi-spectral information of the PA data is essential for semantic segmentation. The blood DSC of the nnU-Net trained on PA images is 0.71 compared to 0.32, which corresponds to the nnU-Net trained on US images. The results for all remaining classes can be found in Table B.1 in the Supplemental Material.
Figure 4.2.5 gives more insight into the DSC results and emphasizes these findings. It shows the distribution of the per-test volunteer DSC values and the ranking results for the class average and the target classes. Note that the findings of the superior performance of the nnU-Net compared to the FCNN and the importance of the multi-spectral information were not altered when using the NSD values, a different level of aggregation, or another (non-test-based) aggregation scheme for the ranking. The distribution of the per-test volunteer DSC values indicates that there is no clear trend of differing DSC performances for different volunteers, body sides, or body sites.

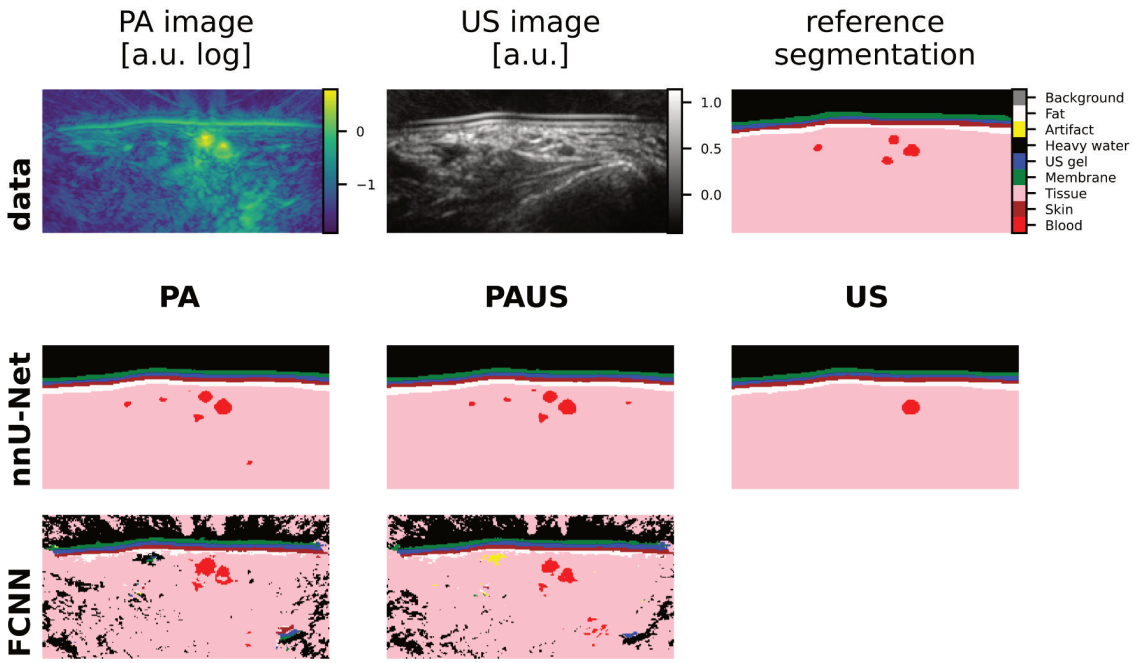

Figure 4.2.4.: Both networks, the nnU-Net and the Fully-Connected Neural Network (FCNN), estimate segmentations that agree with the reference segmentations. The first row shows (*left*) the log-scaled Photoacoustic (PA) image at 800 nm, (*center*) the Ultrasound (US) image, and (*right*) the reference segmentation of the representative example. The estimations of the (*second row*) nnU-Net trained on (*left*) PA, (*center*) a combination of PA and Ultrasound (US) (PAUS), or (*right*) US images are less noisy compared to the ones of the (*third row*) FCNN which was trained on (*left*) PA or (*center*) PAUS images. The example image was chosen according to the median blood Dice Similarity Coefficient (DSC) (calculated on images with at least 60 px classified as blood) for the nnU-Net trained on PAUS images.

The performance of the human annotators for the blood class (mean of 0.66 and standard deviation of 0.09) is also shown in Figure 4.2.5. The detailed results of the linear mixed model can be found in Table B.2 in the Supplemental Material. The DSC values achieved by the nnU-Net trained on PA or PAUS data are, on average, higher than the human annotator performance.

Additional qualitative and quantitative results of the baseline experiment be found in the Supplemental Material B.

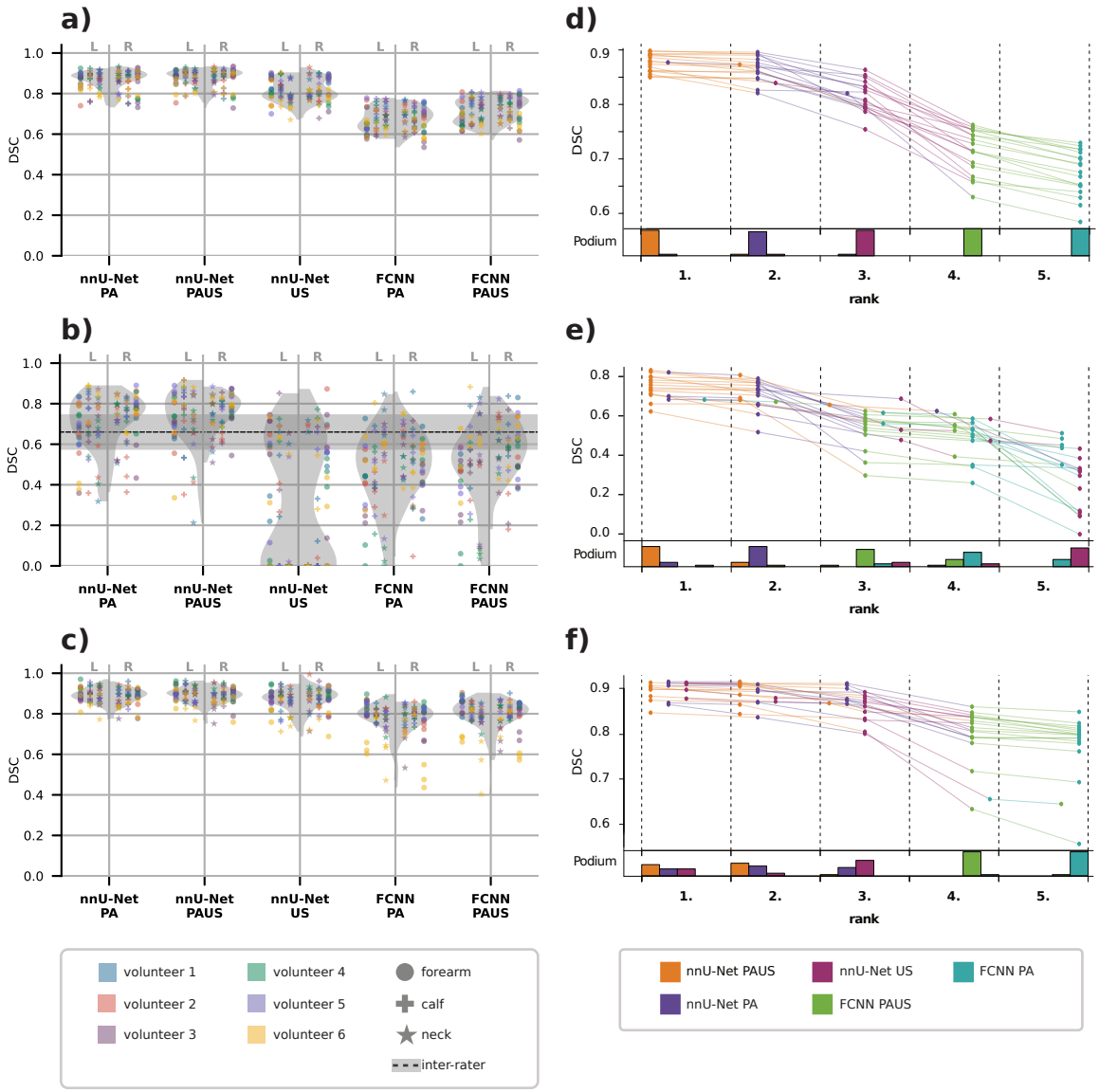

Figure 4.2.5.: The (*left*) raw data plots and (*right*) ranking plots for all combinations of networks and training data (Photoacoustic (PA), Ultrasound (US), and a combination of PA and US (PAUS)) show superior performance of the nnU-Net trained on multispectral images compared to the Fully-Connected Neural Network (FCNN). (*Left*) For each volunteer (color-coded), body site structure (shape-coded), and volunteers' body side (left L or right R side of the respective vertical line), the Dice Similarity Coefficient (DSC) score was calculated (**a**) averaged over all tissue classes, (**b**) for the blood class, and (**c**) for the skin class. For each body side, the relative score frequencies are shown as grey distributions. The mean and standard deviation of the performance of the human annotators are plotted for the class blood as the dotted line and the shaded area, respectively. (*Right*) The (**d**) averaged, (**e**) blood, and (**f**) skin DSC values were aggregated across the three poses and two body sides, color-coded according to the network, and ordered from the highest rank (1) to the lowest (5). One line corresponds to the DSC values achieved with the five networks on one identical test case. The relative frequency of a network achieving a rank is shown in bar charts in the respective bottom areas.

Table 4.2.1.: The overall performance scores of the baseline experiment were calculated with the Dice Similarity Coefficient (DSC) and Normalized Surface Distance (NSD). The scores of the nnU-Net and the Fully-Connected Neural Network (FCNN) trained on Photoacoustic (PA), Ultrasound (US), and a combination of PA and US (PAUS) data were computed for the class average and the target structures, blood and skin. Note that higher values of DSC and NSD (maximum of 1) indicate better performance.

	Structure	nnU-Net PA	nnU-Net PAUS	nnU-Net US	FCNN PA	FCNN PAUS
DSC	Average	0.83	0.85	0.80	0.62	0.66
	Blood	0.71	0.74	0.32	0.48	0.53
	Skin	0.89	0.89	0.87	0.77	0.79
NSD	Average	0.88	0.89	0.84	0.59	0.61
	Blood	0.84	0.85	0.47	0.75	0.75
	Skin	0.98	0.98	0.97	0.87	0.89

Robustness experiment

The results of the *robustness* experiment are in line with the *baseline* experiment. According to the overall performance scores shown in Table 4.2.2, that were calculated for the class average and the target classes with both the DSC and NSD, the nnU-Net showed improved performance compared to the FCNN. For instance, the class average DSC values for nnU-Net and FCNN trained on calf and neck PAUS data and tested on forearm data were 0.82 and 0.65, respectively. The analogous NSD results were 0.87 and 0.59.

Furthermore, these results emphasize that the multi-spectral information underlying the PA images improves the segmentation performance for the blood class. For example, as shown in Table B.3 in the Supplemental Material, the nnU-Nets trained on PA calf and neck data and tested on forearm data achieved an overall blood DSC value of 0.66. The corresponding nnU-Net trained on US data achieved a value of 0.21.

Overall, as shown for the DSCs in Figure 4.2.6, the performance of the robustness experiment was decreased compared to the baseline experiment. However, the differences of DSC results between the robustness and baseline experiments were smaller for the FCNN compared to the nnU-Net, especially clear for the class average.

Supplemental qualitative and quantitative results of the robustness experiment can be found in the Supplemental Material B.

Table 4.2.2.: The overall performance scores of the robustness and baseline experiment calculated with the Dice Similarity Coefficient (DSC) and Normalized Surface Distance (NSD) for the nnU-Net and the Fully-Connected Neural Network (FCNN) trained on a combination of PA and US (PAUS) data. The performance of the models changes when applied to test data with morphologically different structures compared to the training data. In contrast to the baseline experiment, where the models were trained and tested on data from all body sites (forearm, calf, and neck), the models of the robustness experiment were trained on data from two sites and tested on data from the remaining body site. The metrics were computed for the class average and the target structures, blood, and skin. Note that higher values of DSC and NSD (maximum of 1) indicate better performance.

Combination			**A**	**B**	**C**
Training data		All (baseline)	Neck &calf	Forearm & neck	Forearm & calf
Test data		All (baseline)	Forearm	Calf	Neck
	Structure	**nnU-Net / FCNN PAUS**	**nnU-Net / FCNN PAUS**	**nnU-Net / FCNN PAUS**	**nnU-Net / FCNN PAUS**
DSC	Average	0.85 / 0.66	0.82 / 0.65	0.86 / 0.66	0.83 / 0.64
	Blood	0.74 / 0.53	0.70 / 0.52	0.74 / 0.54	0.72 / 0.49
	Skin	0.89 / 0.79	0.86 / 0.75	0.90 / 0.80	0.89 / 0.78
NSD	Average	0.89 / 0.61	0.87 / 0.59	0.89 / 0.61	0.88 / 0.60
	Blood	0.85 / 0.75	0.84 / 0.68	0.83 / 0.78	0.85 / 0.75
	Skin	0.98 / 0.89	0.97 / 0.87	0.98 / 0.86	0.98 / 0.89

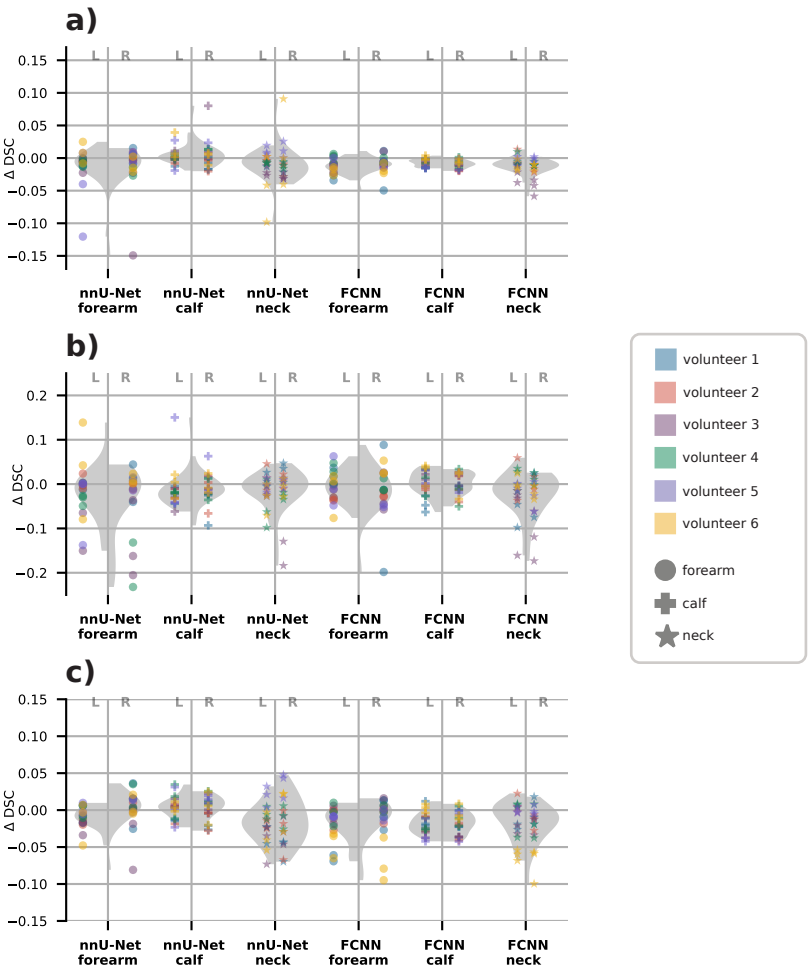

Figure 4.2.6.: Compared to the nnU-Net, the Fully-Connected Neural Network (FCNN) tends to be more robust to body sites not included in the training. The differences of the Dice Similarity Coefficient (DSC) values (**a**) averaged over all structures, (**b**) for the blood class, and (**c**) for the skin class between the robustness experiments and the baseline experiment for the nnU-Net and FCNN trained on Photoacoustic (PA) and Ultrasound (US) (PAUS) images were calculated, respectively. Positive values describe an improved performance of the robustness experiments compared to the baseline results, and negative values mean the opposite. The naming convention of the *x*-axis describes the network and the body site of the test set (cf. Table 4.2.2). As for Figure 4.2.5, the DSC values are plotted for each volunteer (color-coded), body site (shape-coded), and body side (left L and right R of the vertical line). The relative score frequencies are shown as grey distributions, which tend to be narrower for the FCNN compared to those of the nnU-Net.

4.2.6. Discussion

The feasibility of automatic semantic segmentation of PA images with discriminative neural networks was investigated with RQ1. Although the number of available PA and US images was limited, both the *baseline* and *robustness* experiments showed that the **nnU-Net and the FCNN** trained on multispectral images **estimate plausible tissue geometries**, achieving high overlap- and contour-based metric values for the majority of classes.

The validation of the **baseline experiment**, in which the networks were trained and tested on all (forearm, calf, and neck) body sites, revealed that the overall performance was improved for the nnU-Net. The qualitative estimations were less noisy, and the metric scores were higher. This is likely due to the incorporation of spatial context in the nnU-Net that seems to be beneficial for many tasks, as shown in various other semantic segmentation challenges [Isensee et al., 2021]. This finding holds true for the robustness experiment. However, in the robustness experiment, in which the networks were trained on data from two body sites and tested on the remaining one, the FCNN tends to show an advantage compared to the nnU-Net. The difference in DSC scores between the robustness and the baseline experiment was smaller for the FCNN compared to the nnU-Net. This could indicate that the FCNN might be more robust with regard to morphologies not included in the training data. By design, FCNNs rely solely on single-pixel spectral information for estimation, which could explain this outcome. However, the number of learnable parameters was smaller for the FCNN compared to the nnU-Net, which could also influence that result.
Broadly speaking, depending on the task at hand, one could choose one of the two networks or combine the strengths of both methods. For the latter purpose, one could combine their estimates, for example, via ensembling strategies.

Overall, the performance of both network types trained on a combination of PA and US images was slightly worse for the **robustness experiment** compared to the baseline experiment in most of the cases. The reason for this might be a domain gap between the various body sites that were included in the experiment. While the tissue compositions of the imaging regions included in the robustness experiment were relatively similar, an increased performance drop is expected when applying the models on morphologically more different body structures or pathological regions, such as surgical abdominal or cancerous images.

When interpreting the results, it is, however, important to note the **number of test cases** used in this study. Only 108 test images from six volunteers were used in the baseline experiment, and the number dropped by a factor of three for the robustness experiment. Therefore, the results

reported should be interpreted with care, especially with respect to the relative performance of the different architecture/input combinations in the baseline experiment and the overall decrease in performance for the robustness experiment. Particularly for the robustness experiment, the small number of test data could explain the partly higher scores compared to those of the baseline experiment. Including more data in the future would allow a rigorous validation of the different architecture/input combinations, and not only for this study. Yet, the acquisition and manual annotation of PA data is time-consuming, and the availability of large open-source datasets is generally considered one of many existing barriers in the relatively young field of PAI [Assi et al., 2023].

Moreover, the data used in this study was not independent, given a **hierarchical structure**. 16 volunteers were imaged at three body sites on both sides of the body with three images each. In other words, the three images per body side represent a lower-level unit subject to a higher-level unit, the body sides. This may result in a higher correlation within the images of the same body side compared to the correlation with images of the other body side. As shown in Figure 4.1.2, the body sides are another lower-level unit of the higher-level unit of body regions, which in turn is a lower-level unit of the volunteer unit. There was no clear trend that the performance of the algorithms differed on test images of the left and right sides of the body or of different body sites. However, a clustering of DSC results of one test volunteer within the same site could be identified in some cases. Nevertheless, with leave-one-out cross-validation, no significant variations in the estimations of individual volunteers or annotators were noticed. Still, because the test data could not be considered independent and standard statistical analyses, such as variance computations, do not account for interdependencies and lead to biased results, only the mean was reported in this study without a standard deviation.

Across the different tissue classes, the worst performances were achieved for the *blood* and *coupling artifact* classes. Most likely, this is due to the fact that these classes were the most **difficult ones to annotate**. A systematic analysis of the results identified two types of failure cases (cf. Figure B.1): (1) over-segmentation of small superficial vessels and (2) missing annotation of vessels located deeper in the tissue.
An inter-rater reliability study with five additional annotators for the *blood* class revealed that manual annotations of MSOT images are error-prone as there was variation in manual annotations performed by different annotators. For example, the size, location, and number of blood vessels were ambiguous. However, as mentioned before, obvious differences in annotations performed by different annotators could not be detected with the leave-one-out cross-validation. Interestingly, the DSC results of the nnU-Net were slightly improved compared to the human

performance, which might indicate its generalizability across annotation variability.
There are several possibilities that might improve the reliability of the manual annotations. First, the US and PA images could be reconstructed with the same algorithm, which is currently hindered by the MSOT not providing US raw data. Second, the reconstruction algorithm itself could be improved, for example, by accounting for differences in the speed of sound across different morphological structures and volunteers [Dehner et al., 2022b] or the limited bandwidth of the US detection elements and their impulse response [Chowdhury et al., 2021, Chowdhury et al., 2020]. In addition, multi-modal image registration of PAI and other imaging modalities, such as MRI [Ren et al., 2021a], or leveraging the 3D context [Holzwarth et al., 2021b] could improve the annotation quality.

This project was limited to **26 wavelengths and eight tissue classes**. However, additional initial experiments showed that fever wavelengths representing tissue class characteristics could be sufficient. A nnU-Net trained on a combination of PA and US data performed only marginally worse when using five wavelengths evenly sampled from the 26 wavelengths. However, another initial experiment that trained the networks on the original manual annotations that differentiate blood vessels into arteries and veins did not reveal a similar performance to the one presented, highlighting the annotation uncertainty mentioned before. As the manual annotations were performed on both PA and US images, the networks trained on data from both imaging modalities might have had an advantage. Additionally, the number of learnable parameters was highest for this setup, as the number of input channels was highest [Koonce, 2021]. In the future, methods could be developed that analyze the most relevant wavelengths and take annotation uncertainties into account.

In conclusion, discriminative neural networks enable automatic semantic segmentation of multispectral PA images. They replicate the annotation uncertainty of human annotators and could be used to replace manual annotations of PA images in the future.

4.3. Tissue Geometry Generation with Generative Adversarial Networks

Addressed Research Question RQ2:

Can GANs be leveraged for the generation of plausible tissue geometries?

To investigate RQ2, a concept for GAN-based augmentation of reference tissue geometries and its use for subsequent simulation of PA images was developed, which is described in Section 4.3.1. The relevant material and methods are provided in Section 4.3.2. The performed experiments, the corresponding experimental conditions, as well as the results and the discussion, are detailed in the respective Sections 4.3.3-4.3.6.

Disclosure to this work:
The idea for GAN-based tissue generation was developed by Lena Maier-Hein, Janek Gröhl, and myself. Lena Maier-Hein continued to be an exemplary supervisor, providing indispensable suggestions and feedback at various moments. Furthermore, it was a pleasure to discuss the project with Janek Gröhl, Kris K. Dreher, Jan-Hinrich Nölke, Niklas Holzwarth, and Alexander Seitel. I am very grateful to all of them for their tremendous support and valuable feedback, which contributed significantly to the project's outcome. A special thanks goes to Kris K. Dreher, who spent a considerable amount of time helping me with the simulations and also verified the reproduction of the project's results. This work was presented as a talk at the Photons Plus Ultrasound meeting of the SPIE Photonics West conference in 2021 [Schellenberg et al., 2021] and published in the journal *Photoacoustics* [Schellenberg et al., 2022b]. The content, Figures 4.3.1- 4.3.8, and Table 4.3.1 are taken (partly modified) from this publication with permission.

4.3.1. Concept Overview

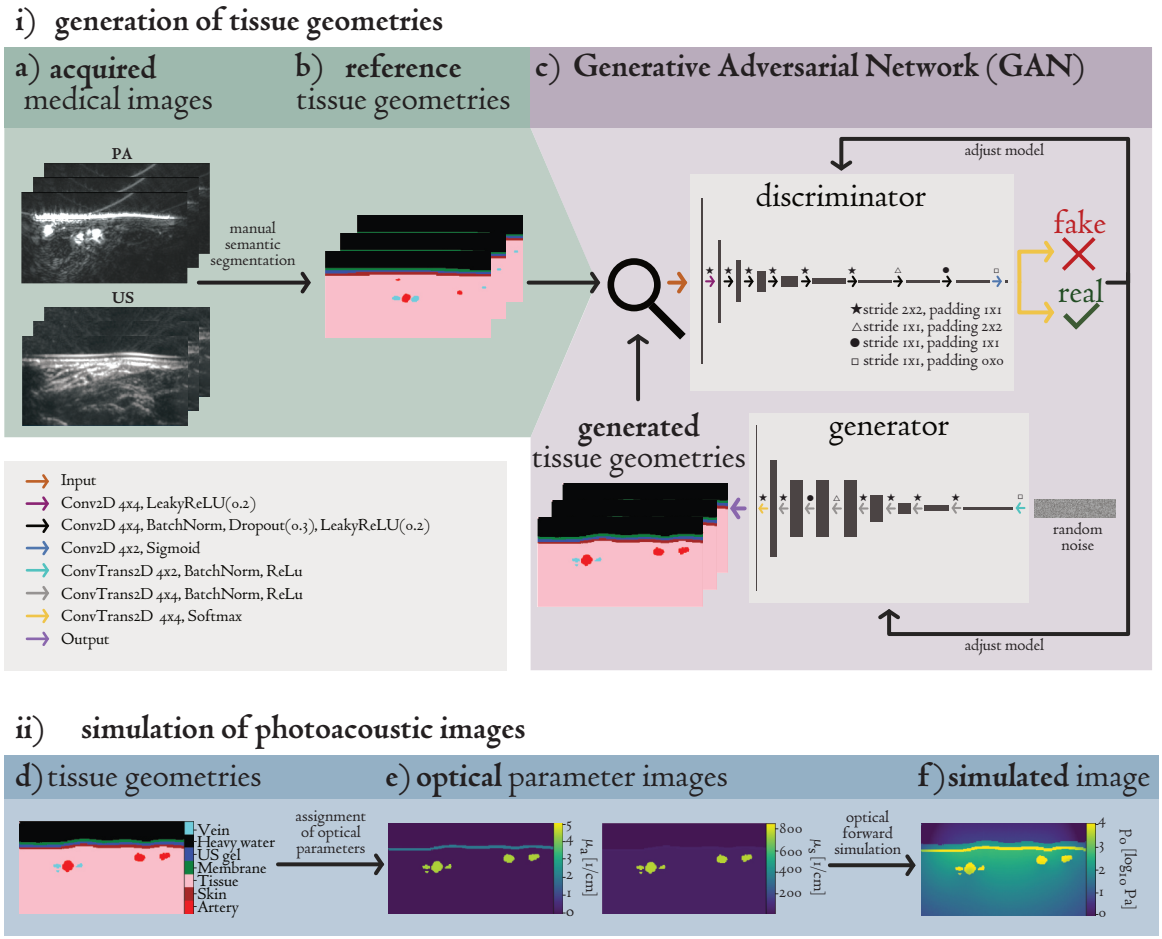

Figure 4.3.1.: Concept for (**i**) generating tissue geometries from a small set of manually annotated reference tissue geometries with a Generative Adversarial Network (GAN) and (**ii**) simulation of Photoacoustic (PA) images based on tissue geometries. (**a**) PA and Ultrasound (US) images are acquired and (**b**) manually annotated by domain experts. (**c**) A GAN is trained on these reference tissue geometries to generate any number of new tissue geometries that follow the training data distribution. (**d**) Based on reference or generated tissue geometries, (**e**) optical tissue properties can be assigned to different tissue classes to simulate a (**f**) PA image.

The concept for the generation of tissue geometries is based on GANs, which have been widely applied in the field of computer vision with great success [Isola et al., 2017, Wang et al., 2018, Zhu et al., 2017a] (cf. Figure 4.3.1). The core idea is to leverage annotated reference tissue geometries from acquired images to generate any number of new geometries resembling the training data

distribution. In other words, the GAN is trained to augment these reference geometries. After assigning optical (and perspective acoustic) parameters to the reference and generated tissue geometries in a probabilistic manner, PA images can be simulated. The simulation of the images serves as an important step for validation with a downstream task (cf. Section 4.3.4).

As illustrated in Figure 4.3.1, the following six steps are required to achieve GAN-based generation of tissue geometries:

a) **Image acquisition:** 2D (or perspective 3D) images of the target anatomy are acquired. For example, PA and US imaging can be performed jointly as in this thesis. However, the method is generally not limited to these modalities, and instead, any modality, such as CT and MRI, could be used.
b) **Image annotation:** Reference tissue geometries are provided, for example, by manual semantic segmentation of the acquired images.
c) **Training of GAN:** A GAN is trained on a training split of the reference tissue geometries.
d) **Inference of GAN:** At inference time, the GAN generates any number of plausible tissue geometries that resemble the training data distribution.
e) **Assignment of tissue parameters:** Based on literature knowledge, optical (and perspectively acoustic) parameters can be assigned to the reference and generated tissue geometries in a probabilistic manner.
f) **Simulation of PA images:** PA images can be simulated based on the optical (and acoustic) parameter images.

4.3.2. Material and Methods

In this section, details on how the concept for GAN-based tissue geometry generation was achieved are presented. In particular, the human-annotated tissue geometries from PA and US images, the deep-convolutional GAN, and the simulation pipeline are described. As described at the end of this section, an additional data set specifically for forearm data with tissue geometries based on literature knowledge was simulated to validate the plausibility of the generated geometries (cf. Section 4.3.3).

Data

The manual annotations of the PA and US images acquired at three body sites, namely forearm, calf, and neck, as described in Section 4.1.3 served as training data for the GAN. Here, the

annotation classes *coupling artifact* and *subcutaneous fat* were merged into the *tissue* class. Considering the hierarchical nature of the data, the annotated images of 13 randomly picked volunteers (N = 78) were selected as training data for each of the three body sites (forearm, calf, and neck). Thus, the data of three subjects ($\sim$ 20 %) were held out as test data.

Generative Adversarial Network

A deep convolutional GAN [Radford et al., 2015] (cf. Figure 4.3.1 c) was trained. The network architecture was primarily chosen because it is considered a baseline for more complex models and showed stable training and high resolutions across a range of data sets [Radford et al., 2015]. As shown in Figure 4.3.1, the generator used a 100-dimensional vector of Gaussian random noise as the input. The output was an image of full image size (256 x 128 px) with seven channels corresponding to the number of tissue classes. The discriminator classified an image of this kind into a single logit. At inference, tissue geometry masks were generated by applying the arg max operator on the generated images of the GAN.

Simulation

PA images were simulated based on the tissue geometries using SIMPA[4]. The first step was to place the 2D tissue geometries in the 3D volume. Then, optical parameters were assigned to the different classes of the tissue geometries, similar to the work by Ma et al., 2020. This spatial distribution of optical tissue parameters allowed the MC-based [Fang et al., 2009] optical forward simulation. Multispectral images with wavelengths from 700 nm to 850 nm in equidistant steps of 10 nm, an isotropic resolution of $\Delta x = \Delta y = \Delta z = 0.16$ mm were simulated with $5 \cdot 10^7$ photons. The output of the simulation pipeline was a 3D initial pressure distribution that was cropped to match the original 2D tissue geometries. Details about the tissue geometry positioning and the assignment of optical parameters follow.

Positioning of tissue geometries in simulation volume The 3D simulation volume was of size 75.0 mm x 20.0 mm x 68.2 mm along the *x*-, *y*-, and *z*-axis. Note that the thickness of the US gel t_{USgel} was added to the z-axis during simulation, such that the final dimension along the z-axis corresponded to $z = 68.2 \text{ mm} + t_{\text{USgel}}$. SIMPA's internally implemented device digital twin of the MSOT Acuity Echo was used to simulate multi-spectral PA images. The MSOT was set in the center top part of the volume, placing the probe origin at $(x,\ y,\ z) = (37.5,\ 10.0,\ 43.2)$ mm. The generated tissue geometries were placed in the simulation volume accordingly. In other

[4] `https://github.com/IMSY-DKFZ/simpa`

words, the highest point of the US gel layer was positioned at the bottom end of the probe origin, at z = 43.2 mm. Since the tissue geometries were 2D and of size 40 mm x 20 mm, they were stacked along the *y*-axis and extrapolated along the *x*-*z*-axis to fill the 3D volume. After simulations, the 2D center *x*-*z* slice was selected and cropped. The corresponding field of view with respect to the probe origin was $x = [-20,\ 20]$ mm and $z = [-4.35,\ 15.66]$ mm, which determined the final 2D slice of size 40 mm x 20 mm.

Optical parameter assignment The optical parameters were assigned to the tissue geometries by accessing the internal tissue library of SIMPA containing literature optical parameters for different tissue classes. For some classes, the optical parameters, i.e., wavelength-dependent absorption, scattering, and anisotropy, are indirectly defined by the sO_2 and Blood Volume Fraction (BVF). This allows a probabilistic assignment of the parameters by sampling from a pre-defined distribution for sO_2 and BVF. The default settings were used for the tissue classes *membrane*, referred to as *mediprene* in SIMPA, *skin*, modeled as *epidermis*, *US gel*, and *heavy water*. The classes *artery* and *vein* were modelled as SIMPA's *blood* class with $sO_2 \sim \mathcal{U}(0.9, 1.0)$ and $sO_2 \sim \mathcal{U}(0.6, 0.8)$, respectively. The background *tissue* class corresponded to SIMPA's class *soft tissue* with $sO_2 = 0.1$ and BVF $\sim \mathcal{U}(0.005, 0.010)$.

Literature-based forearm tissue geometries

An additional *literature-based* forearm data set was generated, including the same tissue classes as the manual annotations. The human forearm model was based on previous internal works [Gröhl et al., 2021c, Dreher et al., 2020]. The tissue geometries were modeled in 3D with the same simulation size and resolution as the simulations based on tissue geometries generated with the GAN. A key aspect here was to first assemble a 2D cross-section of a human forearm with geometries based on literature in a probabilistic manner (cf. Figure 4.3.2). For example, the number of vessels, their radius, eccentricity, and position were drawn from given distributions. In analogy to the simulation procedure, the 2D cross-section was stacked along the *y*-axis to enable a 3D simulation. To mimic possible deformations of the superficial layers, the 3D model was deformed along the z-axis. Further details of the literature-based model can be found in the Supplemental Material C.

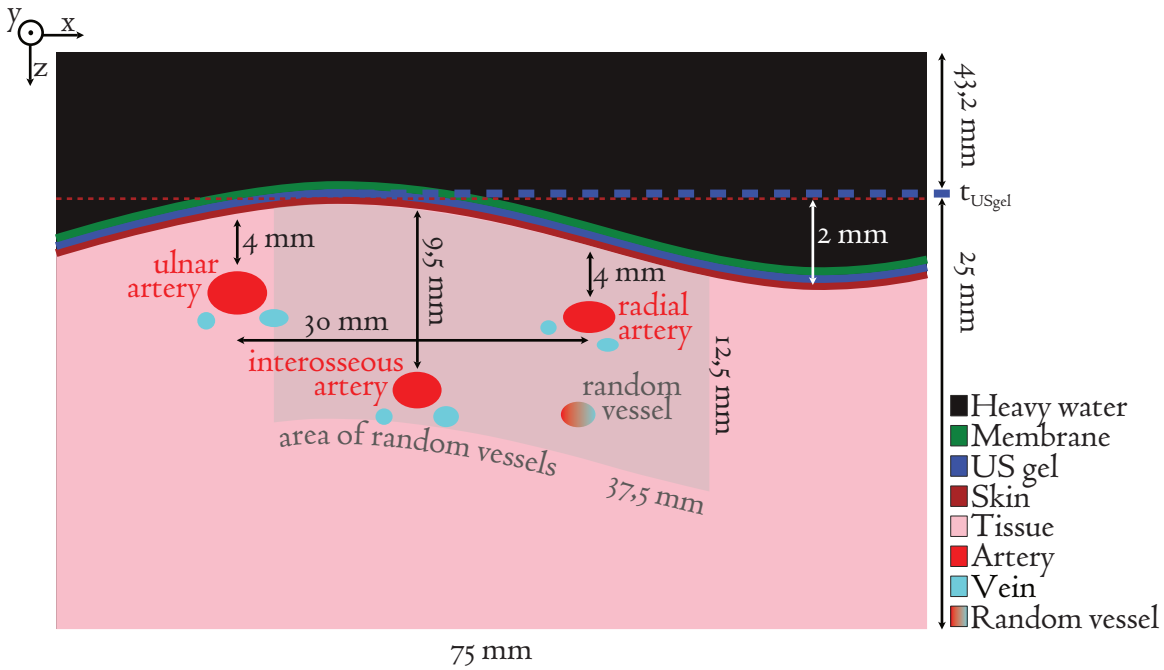

Figure 4.3.2.: An example cross-section (x-z center slice) of the literature-based forearm model. The size of the cross-section is 75 mm along the x-axis and 25.0 mm + $t_{\text{US gel}}$ + 43, 2 mm along the z-axis. The tissue classes *heavy water*, *membrane*, *US gel*, *skin*, background *tissue*, *artery*, and *vein* were modeled according to the literature. Note that *random vessels* were either arteries or veins. The arrows next to the vessels and the skin define the mean of the distances and extension, respectively. The grey area is the region where random vessels were modeled. Note that the upper part of the cross-section is not to scale.

4.3.3. Experiments

To investigate the feasibility and the added value of tissue geometries generated with GANs (RQ2), three experiments were performed that assessed different approaches to tissue geometry generation. In particular, for each body site, one experiment was conducted.

Forearm experiment

For the *forearm* experiment, three approaches to tissue geometry generation were investigated. The simplest approach comprises the already existing reference *annotation-based* tissue geometries (N = 96), which are based on manual annotation. In the second approach, a GAN was trained on the training split of forearm reference annotations (N = 78), which allowed the inference of (N = 1000) *GAN-based* tissue geometries. Specifically for the forearm experiment, a third approach was explored that applied the additional *literature-based* tissue geometries

(N = 500) generated with a model derived from the literature knowledge. These three sets of tissue geometries served to simulate three corresponding forearm PA simulation datasets.

Calf experiment

In analogy to the *forearm* experiment, the reference *annotation-based* calf tissue geometries were investigated (N = 96). Additionally, a GAN was trained on the training set of these calf reference annotations (N = 78), and (N = 500) *GAN-based* calf tissue geometries were generated. Due to the lack of available literature about superficial vessels in the calf, model-based approaches could not be implemented for this body site. The annotation- and GAN-based sets of tissue geometries served to simulate two corresponding calf PA simulation datasets.

Neck experiment

The *neck* experiment is analogous to the *calf* experiment except that neck data was used. Since there is also a lack of available literature about superficial vessels in the neck, model-based approaches could not be implemented for this body site.

4.3.4. Experimental Conditions

This section first presents the hardware and training specifications of the GAN valid for all three experiments. The strategy for validating the different approaches to tissue geometry generation follows. Specifically, the simulated datasets served as training data for a quantification downstream task that allowed a comparative assessment of the plausibility of the underlying tissue geometries. Key to this was that all downstream task models were tested on the same body site-specific realistic target data set.

Computing resources

All training was performed either on a Ubuntu 18.04 workstation with an Intel(R)Core(TM) i7-8700 processor (6 cores) and 64 GB RAM or on a node with an Intel(R)Xeon(R) E5-2620V4 processor (8 cores) and 188 GB RAM of an in-house GPU cluster. In both cases, an Nvidia 2080 RTX Ti graphics card with 10.7 GB RAM was used. The inference, pre-, and post-processing were performed on the workstation.

Deep convolutional Generative Adversarial Network configuration

The deep convolutional GAN was implemented with PyTorch Lightning[5]. To increase the number of training data, a horizontally flipped copy of every training image was added to the training set as augmentation. Since the number of training data was still limited, the data were additionally augmented with rotation and translation based on the work by Karras et al., 2020b before input into the discriminator.
The hyperparameters were set with a grid search on the training data. For each body site, a GAN was trained for 700 epochs with a minibatch size of 3 using the binary CE loss. One-sided label smoothing was performed such that the label for being part of the real data was decreased by a number value from $\mathcal{U}(-0.3,\ 0.0)$. In addition, labels were flipped with an initial probability of $P_{\text{flip}} = 0.2$. The probability decreased with increasing epochs with a slope of $-2.9 \cdot 10^{-4}$. Both the generator and discriminator were trained with a learning rate of $2 \cdot 10^{-4}$. The channel size of the first hidden layer of the generator was 100 for the forearm and neck data and 106 for the calf data. For the discriminator, the channel size of the first hidden layer was 56 for all body regions. The affine transformations applied to the data before entering the discriminator were rigid (rotation and translation in *x*- and *y*-direction) and performed with a probability of $P_{\text{rigid}} = 0.6$ each. The values for the rotation and translations were sampled from $\mathcal{U}(-45\,°,\ 45\,°)$ and $\mathcal{U}(-5\text{ px},\ 5\text{ px})$, respectively.

Photoacoustic quantification downstream task

The benefit of the GAN approach to tissue geometry generation was assessed with a quantification downstream task. For this, the simulated *annotation-*, *GAN-* and, for the forearm experiment, *literature-based* data sets were split into training (70 %), validation (10 %), and test (20 %) data sets and combined into different data configurations as shown in Table 4.3.1. Note that the splits for the annotation-based data were identical to the ones used for the GAN training. Furthermore, since the GAN was trained on the annotation-based training and validation data set, the GAN-based data set can be considered as an augmentation.

For each body region and data set configuration, a U-Net of the same architecture (cf. architecture details in Figure 4.3.3) was trained to estimate the underlying spatial distribution of absorption coefficients from the simulated PA images (initial pressure p_0). Note that the U-Net was chosen based on its successful applications in the field of PAI for quantification tasks [Gröhl et al., 2021b]. Thus, the network input was a PA image of size (256 x 128 px) with

[5] `https://github.com/Lightning-AI/lightning`

16 channels corresponding to the 16 simulated wavelengths. The output was an absorption coefficients map of the same size and channel number. As the GAN, the implementation was performed with PyTorch Lightning.

Table 4.3.1.: Configurations of the data sets for the U-Net-based quantification downstream task. For each of the three experiments (forearm, calf, and neck), three identical compositions with corresponding training, validation, and test splits were configured for the data based on annotations (anno) and the Generative Adversarial Network (GAN). The target test data set is highlighted in bold. Three additional configurations were defined for the forearm experiment. These include data based on literature knowledge (lit) and extended GAN-based data (ext).

			Training	**Validation**	**Test**
			70%	10%	20%
	Abbreviation	**Simulated data set**			
forearm, calf, and neck	**anno**	annotation-based	66	12	**18**
	GAN	GAN-based	350	50	100
	anno-GAN	mix of **anno** and **GAN** (all anno data, fill up with GAN data)	350	50	100
forearm	**lit**	literature-based	350	50	100
	anno-GAN-lit	mix of **anno**, **GAN**, and **lit** (sum of all)	766	112	118
	anno-GAN-ext	mix of **anno** and **extended GAN**-based data (all anno data, fill up with GAN data)	766	112	118

The PA images and absorption coefficient maps were log-scaled for training. As augmentation, Gaussian noise ($\mu = 0$, $\sigma = 0.5$) was added to the PA images, and the training images were horizontally flipped with a probability of 50 %. The U-Nets were trained with the Adam optimizer [Kingma et al., 2014] and the MSE loss. After hyperparameter optimization using a grid search with respect to the validation data set, the learning rate was set to $1 \cdot 10^{-4}$, and a minibatch size of 3 was used. The learning rate was decreased using the ReduceLROnPlatau scheduler of PyTorch with default settings and a minimum learning rate of $1 \cdot 10^{-6}$. The number of epochs until convergence of the validation loss was 30 000 for the annotation-based,

20 000 for the GAN-based, 12 000 for the literature-based, 16 000 for the anno-GAN-based, 14 000 for the anno-GAN-lit-based, and 16 000 for the anno-GAN-ext-based models.

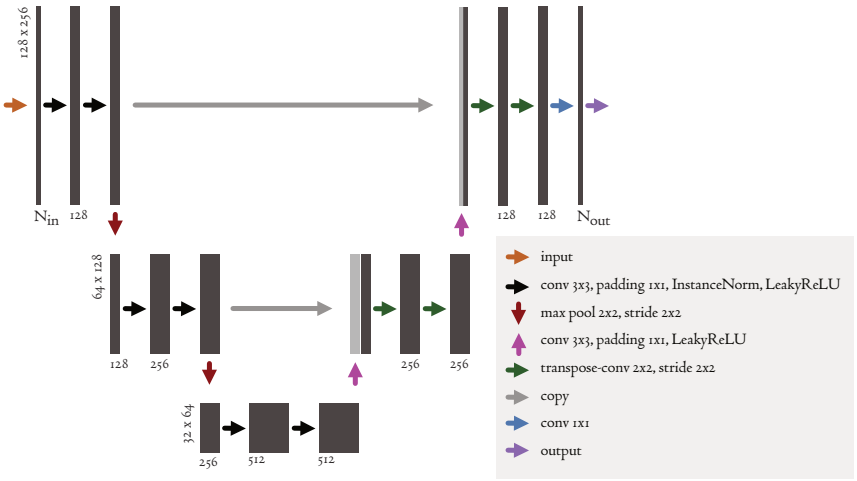

Figure 4.3.3.: The U-Net architecture for the quantification downstream task that was implemented recursively[6]. The input of the network is an initial pressure distribution of size 128 x 256 px with $N_{\text{in}} = 16$ wavelengths encoded as channels. The output is an absorption coefficient map of the same dimensions but with $N_{\text{out}} = 7$ channels according to the tissue classes. The number of channels of the hidden layers is shown at the bottom of every layer, and the image sizes are shown to the left of every downsampling step.

Performance assessment

The performance of the downstream task was investigated on both the respective held-out test data sets (cf. Table 4.3.1) and the realistic annotation-based target test data set (highlighted in bold in Table 4.3.1). Quantitative validation between the estimated and GT absorption coefficient, $\hat{\mu}_a$ and μ_a, was conducted with three metrics, namely the absolute and relative error, $\text{AE}_{x,i_c,\lambda}(\hat{\mu}_a, \mu_a)$ and $\text{RE}_{x,i_c,\lambda}(\hat{\mu}_a, \mu_a)$ and the $\text{SSIM}_{x,\lambda}(\hat{\mu}, \mu)$. Here, x denotes the test image identifier, i the pixel index of tissue class c defining (1) artery, (2) skin, (3) background tissue, (4) US gel, (5) transducer membrane, (6) heavy water, and (7) vein, and λ specifies the wavelength.

[5] `https://github.com/MIC-DKFZ/basic_unet_example/blob/master/networks/RecursiveUNet.py`

Similar to the experiments on tissue geometry estimation (cf. Section 4.2.4), the challengeR toolkit [Wiesenfarth et al., 2021] was applied to compare the performance of the quantification models (for the forearm experiment $N = 6$, for calf and neck experiments $N = 3$) on the target test data set ($N = 18$). For each experiment, the toolkit was run as a multi-task challenge separately for each metric. The tasks were defined according to the quantification performance for different wavelengths ($N = 16$). In analogy to Section 4.2.4, the toolkit was used in the rank-then-aggregate mode to rank the models and establish their stability, thereby assessing the plausibility of the underlying tissue geometries. For this, the pixel-wise results per tissue class of the $\mathrm{AE}_{x,i_c,\lambda}$ and $\mathrm{RE}_{x,i_c,\lambda}$ were aggregated. In line with the work in Section 4.2.4, the classes *artery*, *skin*, and *vein* ($c = 1, 2, 7$) were defined as target classes. Additionally, the resulting per-class values ($\overline{\mathrm{AE}}_{x,c,\lambda}$ and $\overline{\mathrm{RE}}_{x,c,\lambda}$) were averaged to obtain class averages labeled with $c = 0$. Thus, the toolkit was applied to class average results and separately for each tissue class. In contrast, the SSIM results did not require aggregation and could be analyzed directly.

4.3.5. Results

This section presents the results of the GAN-based approach to tissue geometry generation. For each of the three experiments, the performances of the quantification downstream task are comparatively assessed.

Forearm experiment

Figure 4.3.4 shows one randomly chosen example for each of the *literature-* , *annotation-* , and *GAN-based* data sets used for the forearm experiment. The tissue geometries, the assigned optical parameter images (here absorption coefficient), and the resulting simulated PA images are shown.

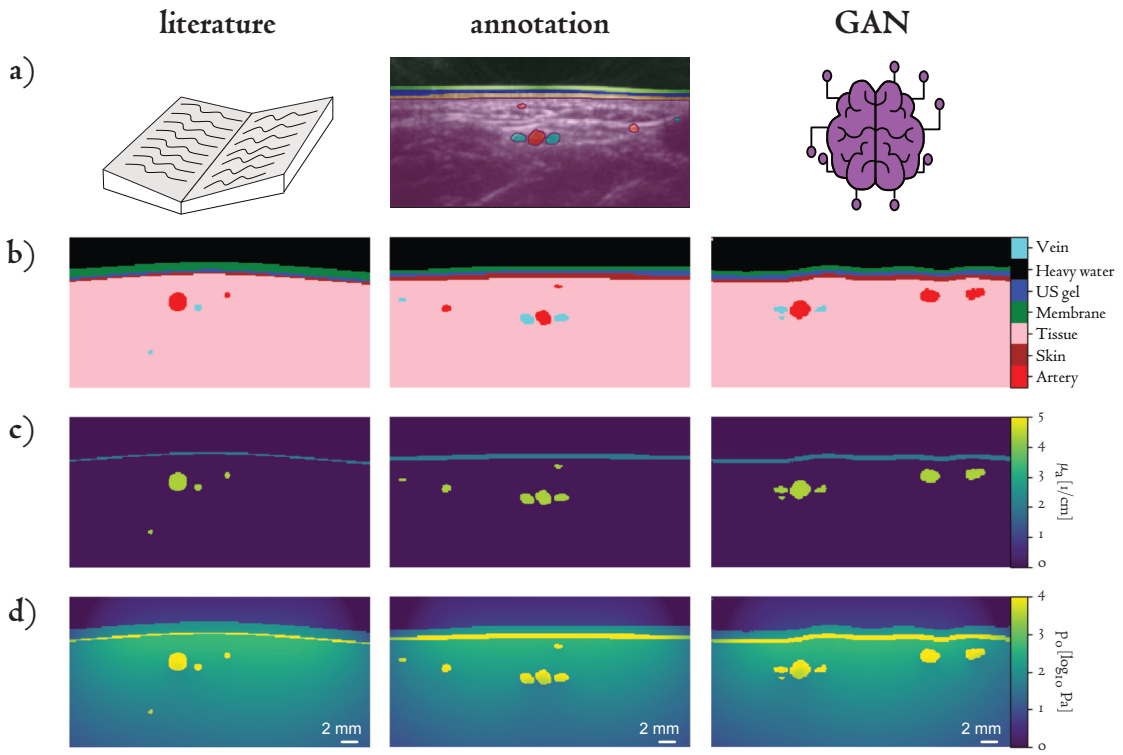

Figure 4.3.4.: Randomly chosen examples of the (*left*) literature-, (*center*) annotation-, and (*right*) Generative Adversarial Network (GAN)-based data sets of a human forearm. For each of the data sets, the (**a**) approach to tissue geometry generation, (**b**) the corresponding tissue geometries, (**c**) the absorption coefficients μ_a, and (**d**) the resulting simulated PA image (initial pressure p_0) at 800 nm are shown.

Figure 4.3.5 displays both qualitative and quantitative outcomes of the downstream task models. These models were trained on literature-, annotation-, and GAN-based images and tested using the same annotation-based forearm data set. The image was selected using the median of the class average absolute errors at 700 nm ($\overline{\text{AE}}_{x,c=0,\lambda=700\text{nm}}$) for the model trained on the literature-based data set. According to the pixel-wise absolute and relative errors at 700 nm and 800 nm, the estimated absorption coefficient maps for the annotation- and GAN-based models more closely resemble the GTs compared to the literature-based model. However, the errors for target structures are generally larger than for the other tissue classes.

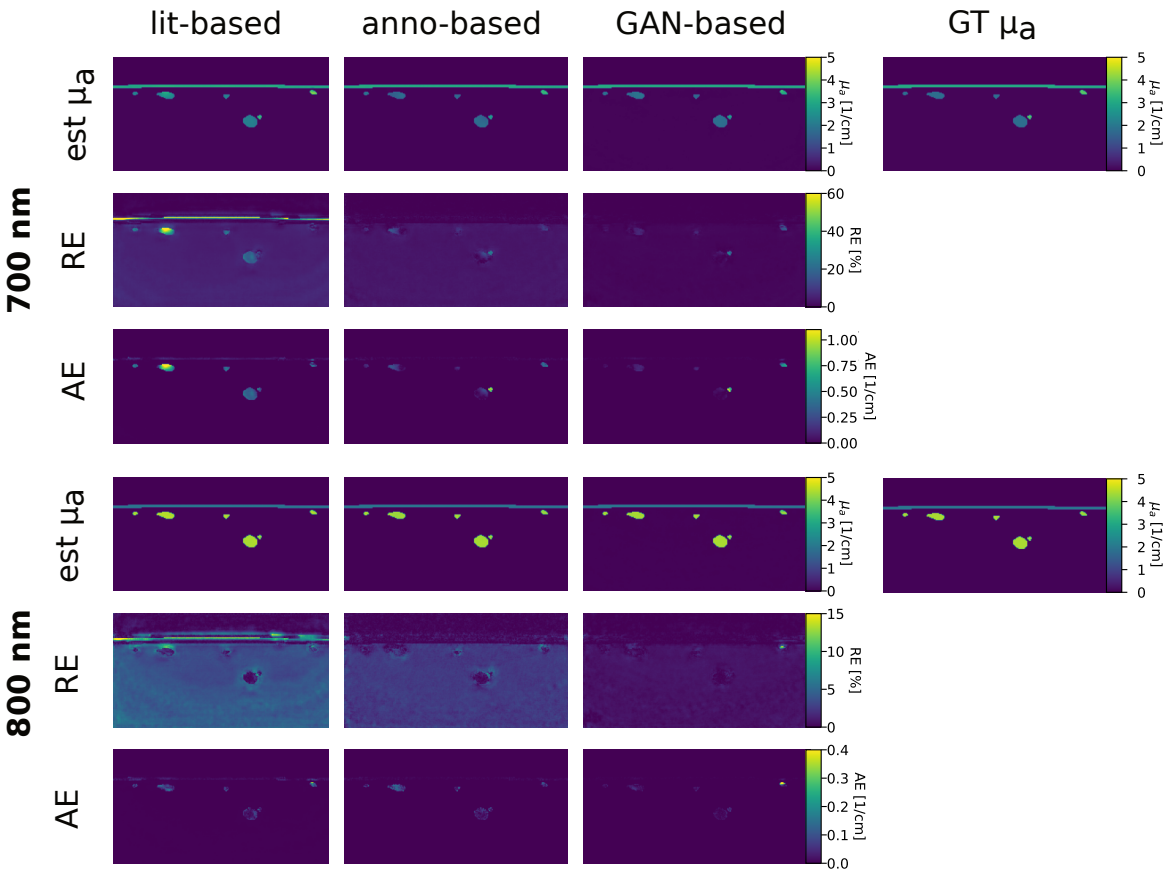

Figure 4.3.5.: Qualitative comparison of the quantification results on a representative annotation-based forearm test case for the models trained on (*left*) literature (lit)-, (*center*) annotation (anno)-, and (*right*) Generative Adversarial Network (GAN)-based data. The estimated absorption coefficient (est μ_a), the relative error (RE), the absolute error (AE), and the corresponding ground truth (GT μ_a) at (*top*) 700 nm and (*bottom*) 800 nm reveal that the annotation- and GAN-based models more closely resemble the μ_a GTs than the literature-based model. The example image was chosen according to the median of the per-image mean absolute errors at 700 nm ($\overline{\text{AE}}_{x,c=0,\lambda=700\text{nm}}$) for the model trained on literature-based data.

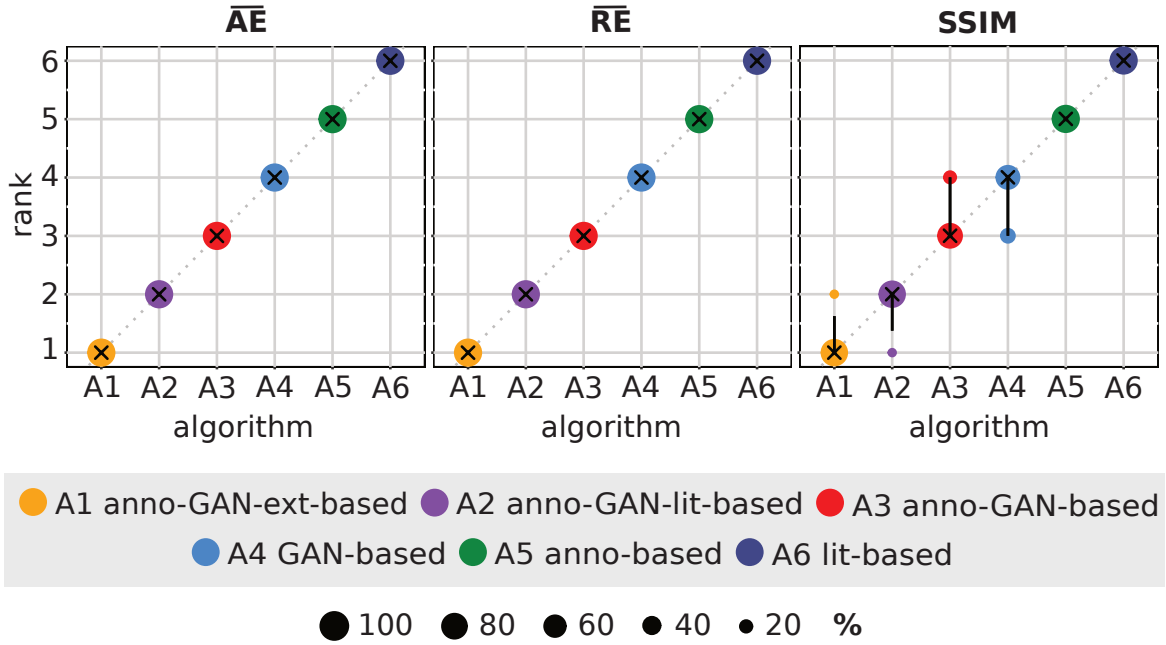

Figure 4.3.6.: Comparative performance assessment of the forearm experiment with six different quantification models and data configurations as described in Table 4.3.1 demonstrate the benefit of Generative Adversarial Network (GAN)-based data. Using the challengeR toolkit, uncertainty-aware rankings (lower is better) were computed for the class average absolute errors ($\overline{\text{AE}}_{x,c=0,\lambda}$), relative errors ($\overline{\text{RE}}_{x,c=0,\lambda}$), and structural similarity indices ($\text{SSIM}_{x,\lambda}$). The rank-then-aggregate scheme was applied to the per-image (x) metrics. A circle's area is proportional to the relative frequency with which the algorithm reached that rank in all tasks. The tasks were to solve the optical inverse problem for 16 wavelengths (λ). The median rank for each model is shown as a black cross, and the black lines mark 95 % confidence intervals ranging from the 2.5th to the 97.5th percentile.

The ranking results emphasize the superior performance of the data-driven method (cf. Figure 4.3.6). Here, the GAN-based model outperforms the (smaller) annotation-based model irrespective of the metric applied. Augmenting annotation-based data with GAN-based data (anno-GAN) further boosts the quantification performance. With the same number of training data, a model trained on a combination of annotation- and GAN-based data (anno-GAN-ext) performs better than the model trained on a mixture of annotation-, GAN- and literature-based (anno-GAN-lit) data. This finding is regardless of the wavelengths and tissue class under investigation (cf. Figure C.1 in the Supplemental Material).

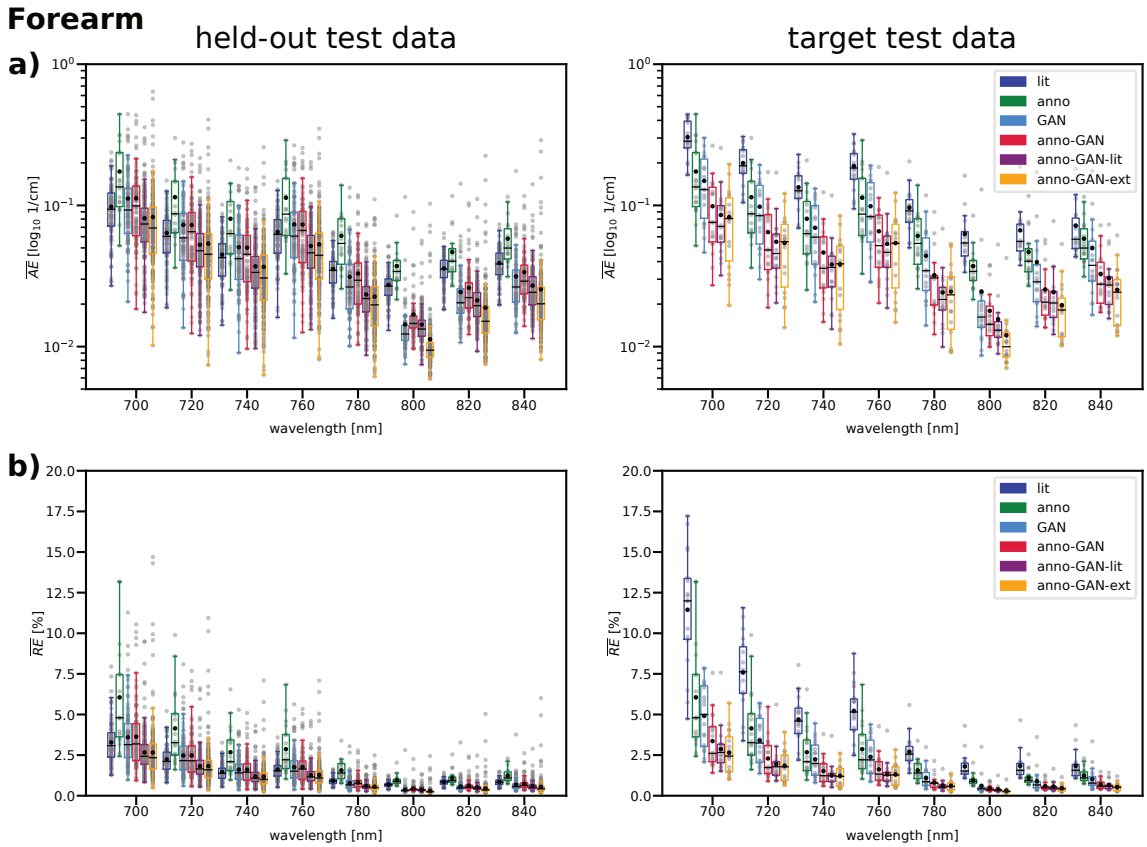

Figure 4.3.7.: Quantitative results of the forearm experiment with six quantification models trained on the different data configurations as shown in Table 4.3.1. Except for the literature (lit)-based model, the (**a**) absolute and (**b**) relative errors of the models (*left*) tested on the in-distribution held-out test set are in the same order of magnitude as (*right*) when applied to the target data. The per-image and per-wavelength absolute and relative errors ($\overline{\text{AE}}_{x,c=1,2,7,\lambda}$ and $\overline{\text{RE}}_{x,c=1,2,7,\lambda}$) aggregated over the target classes artery, skin, and vein (gray dots) are shown. The median, the interquartile range, and the mean values per respective wavelength are indicated as a black bar, colored box, and black dot, respectively.

The absolute and relative errors of the downstream task models trained on the six data configurations (cf. Table 4.3.1) are shown for each test case and averaged over the target annotation classes, $\overline{\text{AE}}_{x,c=1,2,7,\lambda}$ and $\overline{\text{RE}}_{x,c=1,2,7,\lambda}$, in Figure 4.3.7. Here, the models were analyzed on the respective in-distribution and the most realistic annotation-based target test data sets. For the literature-based model, the performance decreases drastically when applied to the target data compared to the held-out test results. Note that for all models, the performance varies with wavelengths on both the held-out and target test data sets.

Additional qualitative and quantitative results of the forearm experiment can be found in the Supplemental Material C.

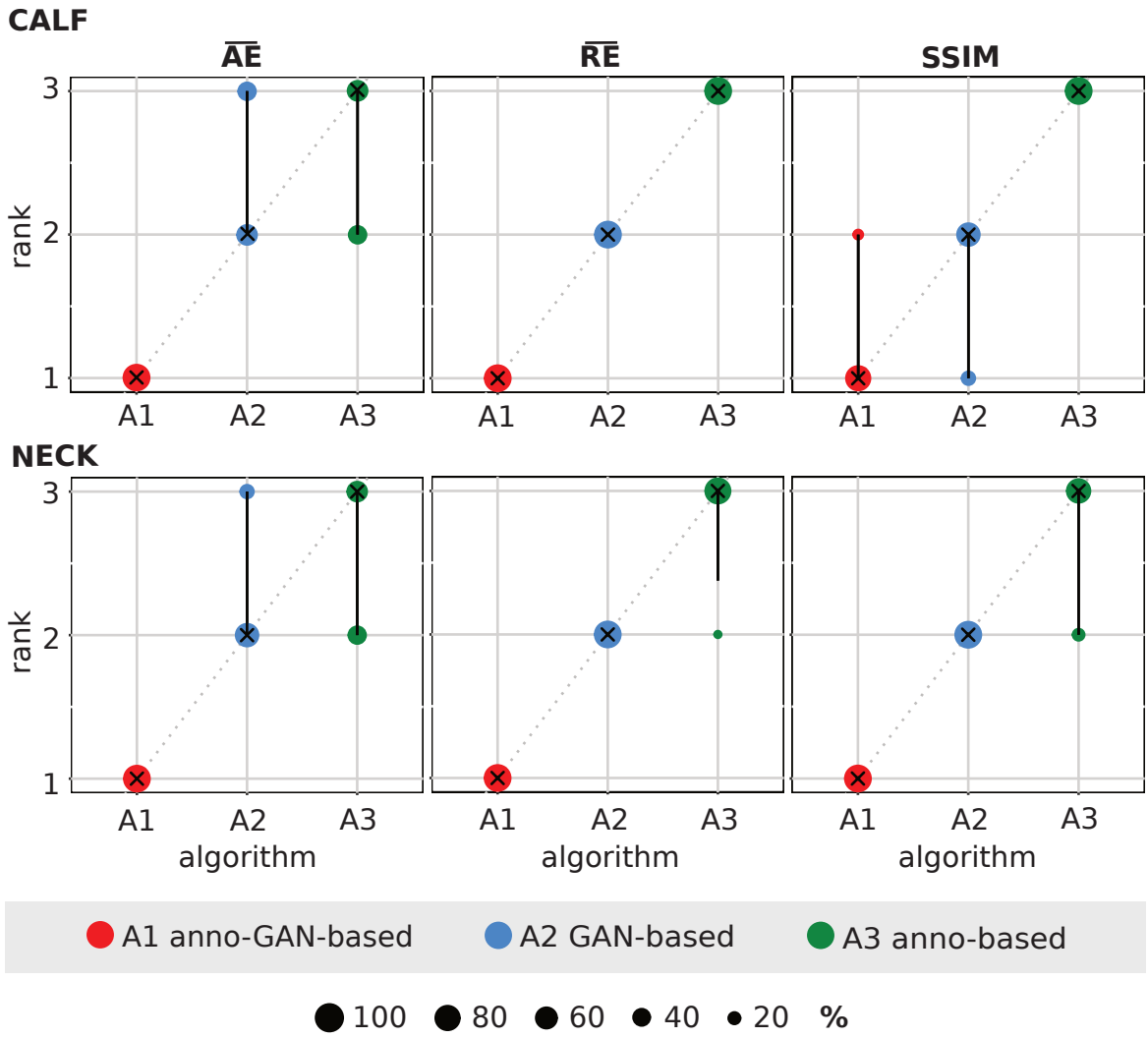

Figure 4.3.8.: Comparative performance assessment of three quantification models of the (*top*) calf or (*bottom*) neck experiments with data configurations as described in Table 4.3.1. The benefit of augmenting annotation (anno)-based data with Generative Adversarial Network (GAN)-based data is significant, as shown by the uncertainty-aware rankings (lower is better) computed for the per-image (x) class average absolute errors ($\overline{\text{AE}}_{x,c=0,\lambda}$), relative errors ($\overline{\text{RE}}_{x,c=0,\lambda}$), and structural similarity indices ($\text{SSIM}_{x,\lambda}$). A rank-then-aggregate scheme was used. A circle's area is proportional to the relative frequency with which the algorithm reached that rank in all tasks. The tasks were to solve the optical inverse problem for 16 wavelengths (λ). The median rank for each model is shown as a black cross, and the black lines mark 95 % confidence intervals ranging from the 2.5th to the 97.5th percentile.

Calf experiment

The findings of the *forearm* experiment agree with the *calf* experiment. The ranking results show substantial improvement in combining data-driven (GAN-based) tissue geometries with manual annotations compared to only relying on the (smaller) annotation-based data (cf. Figure 4.3.8). As for the forearm data, this result is neither affected by the metric used nor by the tissue class analyzed. Additional results are presented in the Supplemental Material C.

Neck experiment

The superior performance of the model that used both annotation- and GAN-based data for training also holds for the *neck* experiment (cf. Figure 4.3.8). This finding is, again, independent of the metric applied and independent of the tissue class analyzed. Further results are presented in the Supplemental Material C.

4.3.6. Discussion

RQ2 whether **GANs allow realistic modeling** of tissue geometries can be affirmed based on the improved performance of the quantification downstream task using the GAN-based augmentation approach. Combining GAN-based tissue geometries with the manually annotated ones improved the quantification performance, even compared to a model that was solely trained on manual annotations, which were assumed to resemble the test data geometries the most due to the same underlying data distribution. Most likely, this is due to the smaller data set size, highlighting the need for sophisticated augmentation strategies. In comparison to models trained on literature-based knowledge, the GAN-based methods showed improved performance, demonstrating successful learning of the data distribution of the reference tissue geometries.

The direct comparison of the forearm models trained on data sets of different sizes (N = 766 and N = 350) showed that the training of the downstream task was still in a regime where **more data improved the estimations**. However, optical forward simulations with the chosen spatial resolution and the number of photons to provide sufficient SNR in initial pressure distributions were time-consuming, even though the actual photon propagation was performed with GPU-based MC methods. Therefore, the number of generated images was chosen as a trade-off between a large number of images and the computing time for the simulations. The model trained on the largest data set combining GAN- and annotation-based data achieved the highest rank in the comparative assessment study.

This emphasizes the value of this **GAN-based augmentation strategy**. In fact, data augmentation for quantitative PAI applications is non-trivial. Standard mechanisms, such as affine transformations, are generally not applicable because they corrupt the accuracy of light propagation in the tissue. This problem was successfully resolved by modeling the tissue geometries independent from the optical properties, which is considered a form of disentanglement. It is worth noting that hand-crafted augmentation of the tissue geometries could be an alternative option instead of relying on the GAN-based approach. However, manual augmentation, in contrast to the GAN-based approach, requires anatomical prior knowledge and fine-tuning, which prevents its transfer and generalizability to different applications without substantial effort.

The general feasibility of the **U-Net for estimating the absorption coefficient** from the initial pressure can be confirmed by the results on in silico held-out test data [Gröhl et al., 2018, Cai et al., 2018, Chen et al., 2020b, Gröhl et al., 2023b]. A previously developed network was used for quantification with no advanced optimization performed. Therefore, there are several options that might improve the quantification performance. For example, ensembling strategies [Allen-Zhu et al., 2020] could be applied, or sophisticated network architectures could be utilized. In addition, elaborated data augmentation techniques could be developed. For example, one could simulate multiple PA images for the same tissue geometries by assigning optical parameters to the tissue structures multiple times and resampling them each time. Since the simulations were performed in 3D and only the center 2D slide was used within this study, one could further augment the training data with the off-center slides of the simulated volumes.
The quantitative downstream task results showed a dependency on the wavelengths. A possible explanation for this behavior can be the inherent wavelength-dependent fluence distributions and corresponding signal intensity differences mainly caused by the target structures. The existing imbalance between the number of pixels assigned to the different target classes and the other classes could amplify this behavior.

One of the main limitations of this work is that the gain of augmenting tissue geometries was **validated with only one in silico downstream task**. In order to better identify the general applicability as well as strengths and weaknesses of the method, additional downstream tasks and further application areas with a larger diversity of anatomical regions and classes, including pathologies such as cancer, should be investigated in the future. In this context, it would also be important to perform the validation of the quantification downstream task on in vivo data. However, validation on real PA data is by no means straightforward and poses two challenges. On the one hand, validation is hampered by the fact that further steps in the simulation pipeline,

such as the selection of optical and acoustic parameters or the modeling of noise, are not yet considered sufficiently realistic (the gap between simulation and reality is still too large). On the other hand, assuming one has realistic virtual data, one would need experimental setups for which the underlying properties are precisely known. Both challenges are part of active research and have not yet been solved.

Overall, this GAN-based method represents a simple concept to generate new tissue geometries in an automated way. While it provides an essential step towards the realistic synthesis of PA images, it can be easily transferred to any imaging modality, such as CT or MRI, thus providing a general concept that could enhance medical image synthesis.

4.4. Tissue Geometry Generation with Scene Graphs

Addressed Research Question RQ3:

Can scene graphs be leveraged for the generation of plausible tissue geometries?

This section presents the work dedicated to RQ3. The overall concept of the developed approach is presented in Section 4.4.1, which is followed by Section 4.4.2 on material and methods. Two experiments were performed, which are described along with their experimental conditions in Sections 4.4.3 - 4.4.4. Corresponding results can be found in Section 4.4.5, which are discussed in Section 4.4.6.

Disclosure to this work:
The concept of generating tissue geometries with scene graphs was developed by Lena Maier-Hein and myself. Lena Maier-Hein supervised the entire project and provided invaluable guidance and essential feedback. In addition, Kris K. Dreher, Jan-Hinrich Nölke, Alexander Seitel, Niklas Holzwarth, Tom Rix, and Christoph Bender were always available for in-depth and beneficial discussions, and Damjan Kalšan verified the reproduction of the project's results. I greatly appreciate all the support and valuable feedback.

Before presenting the concept of the scene graph-based approach, it is important to understand its rationale. This method was conceived from an analysis contrasting the traditional model for tissue geometry generation, based on literature knowledge, with that using the GAN (cf. Section 4.3.5). As shown in Figure 4.4.1, both methods have unique strengths and limitations. A literature-based method involves a detailed semantic representation of the tissue, which provides an understanding of the contextual relationships between the various tissue structures. In general, such a semantic representation of the underlying anatomy is valuable for many reasons [Li et al., 2022b]. For example, insights into the relationship between anatomical structures can be gained that may be beneficial for the diagnosis and treatment planning of pathologies. While the GAN implicitly learns this semantics without providing explicit user access, it provides more realism and is versatile, as it is applicable to reference data from any anatomical region, even from healthy or diseased individuals. The goal of the scene graph-based technique was to leverage the strengths of both paradigms: the ability to semantically represent a scene while generating novel, realistic data based on a reference data set.

	literature knowledge	GAN	scene graphs
semantic representation	✓	x	✓
realism	x	✓	✓
broad applicability	x	✓	✓

Figure 4.4.1.: The potential of the scene graph-based approach to tissue geometry modeling combining the strengths of the literature- and Generative Adversarial Network (GAN)-based methods.

Research outside the field of biomedical image analysis has demonstrated that scene graphs originally proposed by Johnson, 2019 are a powerful representation that encapsulates the semantic structure of an image (or a language) by capturing its objects, attributes, and relationships in the form of a graph. While scene graph-based representations encode context that can be used for basic recognition tasks, they also offer high potential for mastering and improving visual tasks due to their structured abstraction and improved semantic representation capacity compared to conventional image features [Zhu et al., 2022a]. Their success has been shown in various studies related to scene understanding in the field of computer vision, including image captioning, visual question answering, content-based image retrieval, and image generation [Zhu et al., 2022a].

The approach presented here builds on previous image synthesis methods that explicitly benefit from conditioning on scene graphs [Johnson et al., 2018, Kar et al., 2019]. More specifically, the concept closely follows the work by Kar et al., 2019. The key idea is to condition image generation on scene graphs that encode prior knowledge about the content of an image. This means that the nodes of the scene graphs represent different objects of an image. Their node attributes encode features of the objects. The relationships between objects are represented by the scene graphs' edges. The image synthesis is achieved with a GNN that learns the distribution of selected node attributes of the graphs such that the corresponding generated images resemble target images.
Unlike the literature-based model, this approach is data-driven and learns to mimic the reference data distribution, assuming the prior knowledge is correct (realism and broad applicability). In comparison to the GAN method (cf. Section 4.3), this approach requires prior knowledge during training and inference of the synthesis method. However, it improves upon the GAN-based approach in one key factor, namely that the distributions of selected attributes are explicitly learned by leveraging the scene graph representation.

4.4.1. Concept Overview

In the specific context of this thesis, the scene graph-based approach is intended for the synthesis of tissue geometries. The potential of this approach lies in the fact that this technique serves a dual purpose: as a generative model and as a mechanism for analyzing and modifying geometric quantities that are represented as node attributes. The underlying distributions of these geometric quantities learned from the provided data set might hold implications for understanding diseases characterized by anatomical alterations. Additionally, by adjusting the inferred distributions, one can tailor tissue geometry generation to specific needs, such as synthesizing data sets with intentional prevalence shifts (cf. Figure 4.4.2).

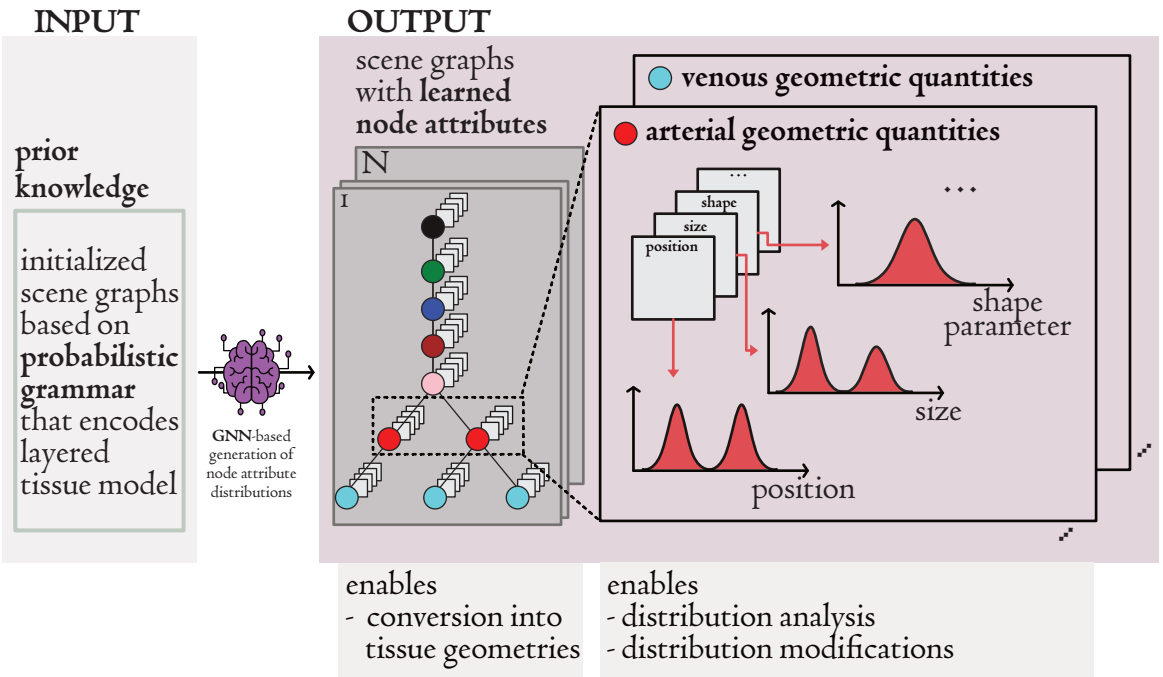

Figure 4.4.2.: The potential of a scene graph-based approach for tissue geometry generation and simultaneous analysis of geometric quantities. By using prior knowledge about the general structure of the tissue, scene graphs can be generated. A Graph Neural Network (GNN) generates the node attributes of N graphs. These N graphs can be converted into tissue geometries. Additionally, the distributions of the attributes, for example, those in relation to the position, size, and shape of one tissue class, can be analyzed and modified.

In this specific concept, the prior knowledge about the content of an image refers to the general tissue composition of an image. It encloses knowledge about the occurrence probability of different tissue structures and their spatial correlation in an image. For example, this prior

knowledge includes the varying number of veins and the US-gel layer always being above the skin. Scene graphs, as shown in Figure 4.4.3, can efficiently capture this prior knowledge by encoding the layered tissue structures of an image as nodes of a graph tree. This means that the nodes of a scene graph represent different geometric tissue structures of an image, and the edges encode the nodes' correlations. The node attributes represent geometric quantities, such as the size and position.
To generate these graphs, this concept builds, as the work by Kar et al., 2019, on a context-free probabilistic grammar encoding the prior knowledge. Such a grammar consists of a set of production rules that define with which probability a specific node follows another one.

Following this probabilistic grammar, scene graphs with initialized node attributes can be generated. The objective of the concept is to learn the distributions of the structure-specific node attributes of these graphs. To explicitly learn these geometric quantities and simultaneously generate plausible tissue geometries, a GNN learns to map the *input scene graphs* into *output scene graphs* of the same node structure but with transformed node attributes. Note that the number of node attributes of the input graphs is not directly related to the number of learned output node attributes (cf. Figure 4.4.3).

In order to allow optimization of the GNN parameters, two complementary training paradigms are explored, one requiring reference annotation masks (*annotation-based* optimization) and one relying solely on PA images (*image-based* optimization).
In a broader context, this training scheme belongs to the class of GMMNs, and the MMD is used as the loss function to compare the distributions of the generated and training minibatches of data (cf. Section 2.2.3). During inference, the trained GNN allows the user to generate new tissue geometries that follow the training data distribution and, at the same time, to gain insights into the optimized geometric quantities (cf. Figure 4.4.2).

Annotation-based optimization For the annotation-based training, the output of the GNN with optimized output node attributes is converted into tissue geometries whose data distribution can be compared with the ones of given target reference tissue geometries to compute the loss (cf. annotation-based optimization in Figures 4.4.3).
Note that the distribution of the different geometric parameters could, in principle, also be extracted from the target tissue geometry masks. However, extracting these parameters from the masks can be ambiguous, potentially requiring application-specific post-processing steps. In contrast, the approach chosen here only requires the defined forward path to be computed, meaning the conversion of the transformed scene graphs into tissue geometry masks.

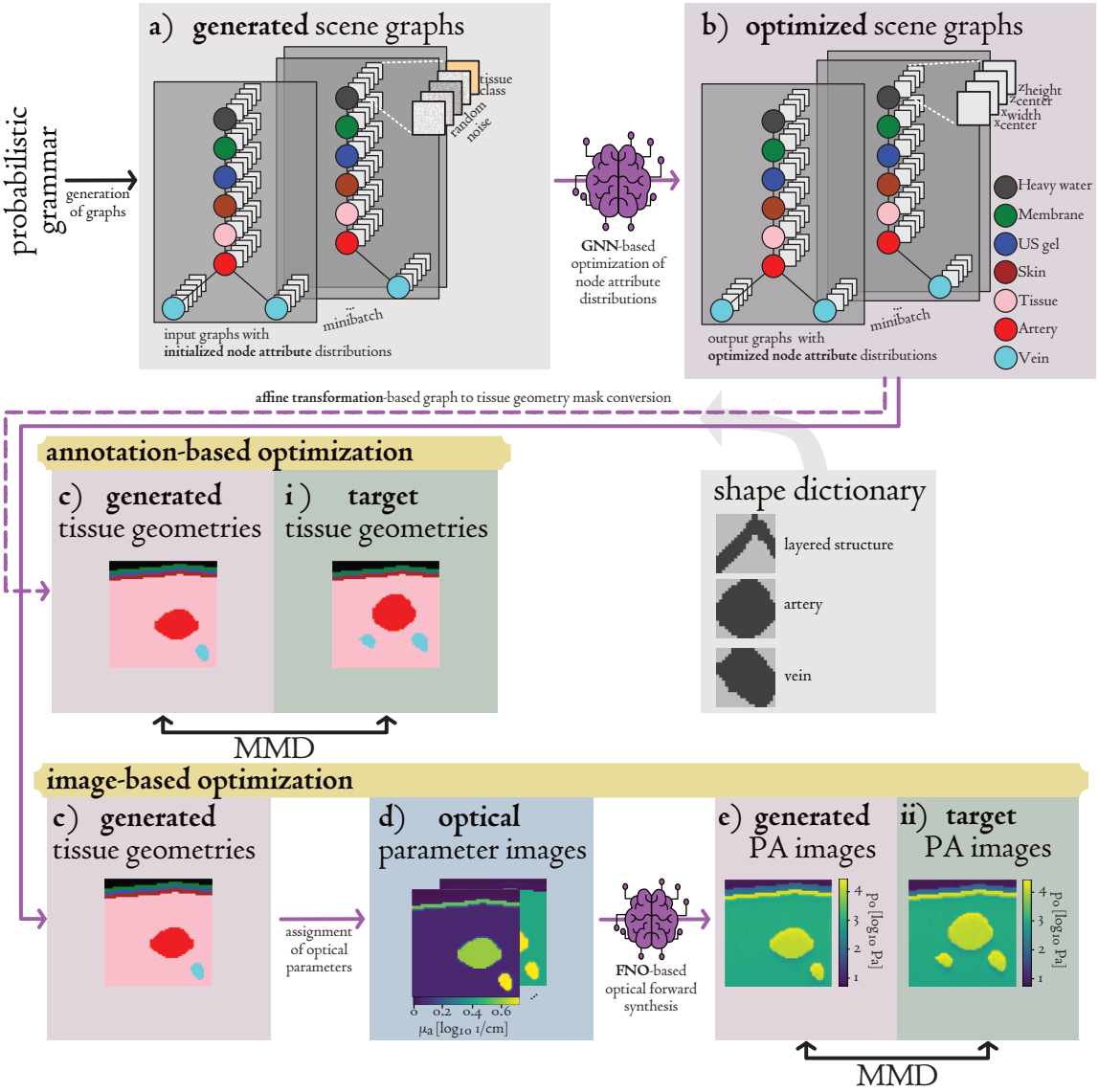

Figure 4.4.3.: Scene graph-based concept for tissue geometry generation. A probabilistic grammar encoding prior knowledge about the tissue composition allows for the generation of (**a**) *input scene graphs* with initialized node attributes. A Graph Neural Network (GNN) learns to map the input attributes into meaningful attributes, and the (**b**) corresponding *output scene graphs* are converted into (**c**) tissue geometry masks using a shape dictionary. For the *annotation-based* optimization, the generated masks are compared to (**i**) target masks to compute the Maximum Mean Discrepancy (MMD). For the *image-based* optimization, optical parameters, such as the absorption coefficient μ_a, are assigned to the tissue geometries, and the (**d**) optical parameter images serve as the input of a Fourier Neural Operator (FNO)-based PA image synthesis. The (**e**) generated Photoacoustic (PA) images can be compared with (**ii**) target images to compute the MMD.

Image-based optimization Since annotations are not always available, usually time-consuming, and often inhibit uncertainties, optimizing geometric quantities without the need for reference annotations but using PA images themselves is preferable. The concept accounts for this limitation (cf. image-based optimization in Figure 4.4.3) by computing the MMD between minibatches of virtual and target reference PA images. For this purpose, a pre-trained FNO-based model for photon propagation is included in the optimization process to synthesize virtual images.

The pipeline of the graph-based concept consists of the following steps (cf. Figure 4.4.3). While the image-based optimization includes all five steps, only steps (a) - (c) need to be performed for the annotation-based paradigm. Note that these steps are required to be differentiable to allow back-propagation of the computed gradients during optimization:

a) **Scene graph generation:** The first step is the generation of scene graphs that encode the prior information about tissue composition. Here, analogous to the work by Kar et al., 2019, a probabilistic grammar that encodes the prior knowledge is used to generate the hierarchical graph trees [Zhu et al., 2007]. However, in principle, there are several ways to generate scene graphs [Zhu et al., 2022a], and the approach is not limited to hierarchical graph structures.
b) **Optimization of node attributes:** A GNN is used to learn the distribution of the meaningful output node attributes, such as the size and position. Thus, the structure of the input graph is not changed during training, but only the node attributes. Note that, in principle, any number of node attributes can be set.
c) **Graph to tissue geometry mask conversion:** The transformed graph tree is converted into tissue geometry masks in a differentiable fashion.
d) **Assignment of tissue parameters:** As described in Section 4.3.1, the optical (and perspective acoustic) parameters are assigned to the generated tissue geometry masks based on literature knowledge.
e) **DL-based PA synthesis:** Taking the optical parameter images as input of a DL-based optical forward model, PA images can be generated.

4.4.2. Material and Methods

The proposed concept for tissue geometry generation can be implemented with various degrees of complexity. In this first proof-of-concept study, a comparatively simple setting to demonstrate general feasibility was explored. The related material and methods are presented in this section. First, the in silico data sets are described. Then, details of the GNN and the conversion of graphs into images follow. Last, the Monte Carlo-based simulation and the DL-based optical forward model are described.

Data

For this study, a tissue model was used that resembles the manual annotation data (cf. Section 4.1) in terms of occurring tissue classes and layered tissue structure. It consisted of heavy water, membrane, US gel, skin, and one artery with up to two accompanying veins. Following the work by Kar et al., 2019, a probabilistic context-free grammar was defined based on the top-down structure of the tissue geometries. As shown in Algorithm 1, it consisted of production rules with specific probabilities to generate new graph trees.

Two data sets were generated based on this grammar, referred to as *input* and *target* graphs.

Input graphs 10 000 input graphs were generated with the grammar to avoid generating them during the training. Each node of an input graph was assigned five node attributes. In each epoch during training, the first four node attributes were randomly sampled from a Gaussian distribution $\sim \mathcal{N}(0, 1)$. Using these first four node attributes of the input graphs, the graph convolutions could learn node-specific information only through the graph structure itself. To learn node-specific properties more efficiently, the fifth attribute was introduced. It encoded the tissue class by an integer between zero and seven according to the number of tissue classes. Initial results with the added fifth attribute showed a significant increase in the convergence time compared to training with only the first four attributes. A training, validation, and test split was created with respective 6000 ($\sim$ 60 %), 2000 ($\sim$ 20 %), and 2000 ($\sim$ 20 %) randomly sampled data.

S	→	Heavy_Water Connection1 [1.0]
Heavy_Water	→	heavy_water [1.0]
Connection1	→	Membrane Connection2 [1.0]
Membrane	→	membrane [1.0]
Connection2	→	US_Gel Connection3 [1.0]
US_Gel	→	us_gel [1.0]
Connection3	→	Skin Connection4 [1.0]
Skin	→	skin [1.0]
Connection4	→	Background_Tissue Connection5 [1.0]
Background_Tissue	→	background_tissue [1.0]
Connection5	→	Artery Connection6 [1.0]
Artery	→	artery [1.0]
Connection6	→	none [0.1]
Connection6	→	Veins [0.9]
Veins	→	Vein1 Connection7 [1.0]
Vein1	→	vein [1.0]
Connection7	→	vein [0.4]
Connection7	→	none [0.6]

Algorithm 1: Context-free probabilistic grammar. A new graph can be generated following the production rules sequentially, starting at symbol S. The rules' probabilies are shown in brackets. Words starting with capital letters denote non-terminal symbols, and small letter words denote terminal symbols.

Target graphs The target graphs were generated with the same probabilistic grammar and served to generate the target tissue geometries and PA images with known distributions of underlying geometric quantities. In contrast to the input graphs, each node was assigned four node attributes encoding the size and position. More specifically, these four node attributes encoded the width, height, center position in the x-direction, and center position in the z-direction of the corresponding bounding box of a structure. The target data set's node attributes followed distributions that were predefined specifically for this in silico study (cf. Table 4.4.2). In addition, some node attributes were specified as mutable, meaning that the corresponding attribute values were considered in the graph to tissue geometry mask conversion (cf. step c). Here, the mutable nodes were of tissue classes membrane, artery, and vein, with only the attribute position in the z-direction being mutable for the membrane. Note that the shape

was not optimized and instead retrieved from a shape dictionary (cf. Figure 4.4.3). The shape dictionary included one example shape for an artery, one for a vein, and one for the membrane. These example shapes were extracted from manual annotations (cf. Section 4.1.3) and rescaled to a size of 20 x 20 px. In total, 20 000 target graphs were generated, 10 000 of which were used to train the GNN. The remaining 10 000 graphs were solely used for training the DL-based optical forward model, as detailed in one of the following paragraphs. In analogy to the input graphs, training, validation, and test splits with respective 6000 ($\sim$ 60 %), 2000 ($\sim$ 20 %), and 2000 ($\sim$ 20 %) randomly sampled data were created.

Table 4.4.1.: Mean and standard deviations of Gaussian distributions of mutable geometric quantities of the target data set. For the center position in the x-direction of veins, two Gaussian distributions that are symmetrically centered around the origin ($\pm$ 17 px) were superimposed.

		Mean	**Standard deviation**
Tissue class	**Attribute**	[px]	[px]
Artery	center position in x	0.0	3.5
	center position in z	27.0	2.0
	width	20.0	3.5
	height	20.0	3.5
Vein	center position in x	$\pm$ 17.0	1.0
	center position in z	17.0	1.0
	width	9.0	2.0
	height	9.0	2.0
Membrane	center position in z	5.0	1.0

if $\pm$, then bimodal

Graph Neural Network

A GNN of five hidden layers was designed. Every layer consisted of a transformer convolutional layer [Shi et al., 2020], a layer normalization, and a LeakyReLU activation. Three of the transformer convolutions used 15 multi-head-attentions, and two of them used five. A final transformer convolution followed the last hidden layer to predict the output. The input of the network was a graph with $N_{\text{in}} = 5$ node attributes (input graphs). The output was a graph of the same node structure but with $N_{\text{out}} = 4$ node attributes representing the tissue's geometric

quantities (output node attributes). The network architecture and dimension size N can be extracted from Figure 4.4.4. Note that the choice of network architecture is based on the work by Kar et al., 2019, and implementation details were determined by initial experiments, for example, by varying the number of transformer convolutional layers.

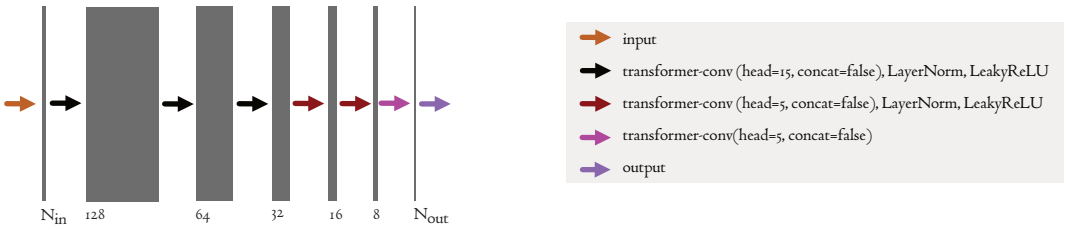

Figure 4.4.4.: Network architecture of the Graph Neural Network (GNN). The scene graphs with $N_{\text{in}} = 5$ node attributes were input to the GNN, which transforms the node attributes by five hidden layers, including transformer convolutions, layer normalization, and activation functions to an output scene graph with $N_{\text{out}} = 4$ node attributes.

Graph to tissue geometry mask conversion

To allow optimization of the GNN using reference tissue geometries or PA images, a differentiable conversion from graphs into tissue geometries was required. For this proof-of-concept study with a relatively simple setting, a differentiable transformation was implemented using affine transformations and a shape dictionary. The exact procedure consisted of the following four steps:

1. An array filled with zeros of image size 64 x 64 px with channels according to the number of tissue classes was initialized (similar to one-hot encoding). This *class-encoded* array is referred to as *tissue geometry masks*.

2. The tissue geometries of the mutable nodes were created using the shape dictionary and placed in the corresponding channel of the class-encoded array using affine transformations (scaling and translation). The shape dictionary containing binary example shapes of size 20 x 20 px for each mutable class (cf. Figure 4.4.3) was scaled to the specific size and placed at the appropriate position in the corresponding channel. Note that these transformations required bilinear interpolation for differentiability, leading to the channels not being piecewise constant as the edge regions of the geometries were smoothly interpolated between zero and one.

3. The remaining immutable nodes were placed in their associated channel using the prior knowledge encoded in the graph. For example, heavy water was always positioned above the membrane, and the US gel layer was always positioned below the membrane. For simplicity, the same shape was chosen for the US gel and skin as for the membrane. More details of the algorithm are shown in Algorithm 2.

4. The tissue geometry masks were clamped between values of zero and one. This step ensured that the range of values between the generated and the target data was the same, even in the case of overlapping structures.

X := array filled with zeros of image size 64 x 64 px and with i = 7 channels
`/* mutable nodes */`
for i *in mutable node types* [*membrane M, artery A, vein V*] **do**
  for $j = 0;\ j \leq$ # *nodes of that type;* $j = j + 1$ **do**
    load binary shape map for that node type form dictionary
    scale and position shape according to attribute values into X_i
  end
end
`/* non-mutable nodes */`
for i *in non-mutable node types* [*skin S, US gel U, heavy water H, background tissue B*] **do**
  if i *is skin S* **then**
    load binary shape map for node type membrane from dictionary
    scale and position shape according to fixed attribute values into X_S
  end
  if i *is US gel U* **then**
    find the area between membrane and skin
    fill X_U at these positions with $1 - (X_M \cup X_S)$
  end
  if i *is heavy water H* **then**
    find the area on top of the membrane
    fill X_H at these positions with $1 - X_M$
  end
  if i *is background tissue B* **then**
    find the area under the skin
    fill X_B at these positions with $1 - (X_M \cup X_A \cup X_V)$
  end
end
clamp X between 0 and 1

Algorithm 2: Graph to tissue geometry mask conversion.

Loss calculation with Maximum Mean Discrepancy

To compare the generated and target tissue data, the MMD was calculated per minibatch and with a polynomial kernel (cf. Section 2.2.3). As described in Section 2.2.3, the MMD highly depends on the chosen kernel parameters, and usually, multiple kernel parameters are applied. Therefore, the final MMD loss value was computed as the sum of the MMD results calculated with different kernel parameters, a and b (cf. Section 2.2.3).

Specifically for the annotation-based optimization scheme, the MMD was calculated separately for each mutable class c between the generated output and target tissue geometry masks, $\hat{X}_c$ and X_c, because initial results showed that this allowed faster convergence. Thus, the final loss score was the sum of the individual MMD results:

$$L = \sum_{c} \sum_{a,b} \mathrm{MMD}^2_{a,b} \left(\frac{\hat{X}_c}{s} + 1, \frac{X_c}{s} + 1 \right). \tag{4.1}$$

Note that s denotes a scaling factor. It was introduced to allow the use of the same kernel parameters for both optimization schemes (annotation- and image-based) and chosen proportional to the maximum value of the target data set. The constant value of one was added to the scaled images to prevent multiplications with zero for the computation of the correlation matrices included in the MMD.

In accordance, the MMD calculation for the image-based optimization scheme with generated output and target PA images, $\hat{P}$ and P, was the following:

$$L = \sum_{a,b} \mathrm{MMD}^2_{a,b} \left(\frac{\hat{P}}{s} + 1, \frac{P}{s} + 1 \right). \tag{4.2}$$

Simulation

In analogy to the GAN project (cf. Section 4.3.2), optical parameters were assigned to the tissue geometries depending on the tissue classes. The internal tissue library of SIMPA was used. For the artery, vein, and background tissue classes, the sO_2 and BVF were fixed. More specifically, the sO_2 for arteries was set to 95 %, the sO_2 for veins was set to 70 %, and the sO_2 and BVF were set to 3 % and 65 % for background tissue, respectively. The class-specific optical parameters were multiplied by the tissue geometry masks. In other words, for each of the optical parameters (absorption, scattering, and anisotropy), one class-encoded parameter array of the same size

as the class-encoded tissue geometry array was calculated. Each parameter array was summed along the class-specific channel dimension to obtain the final optical parameter images.

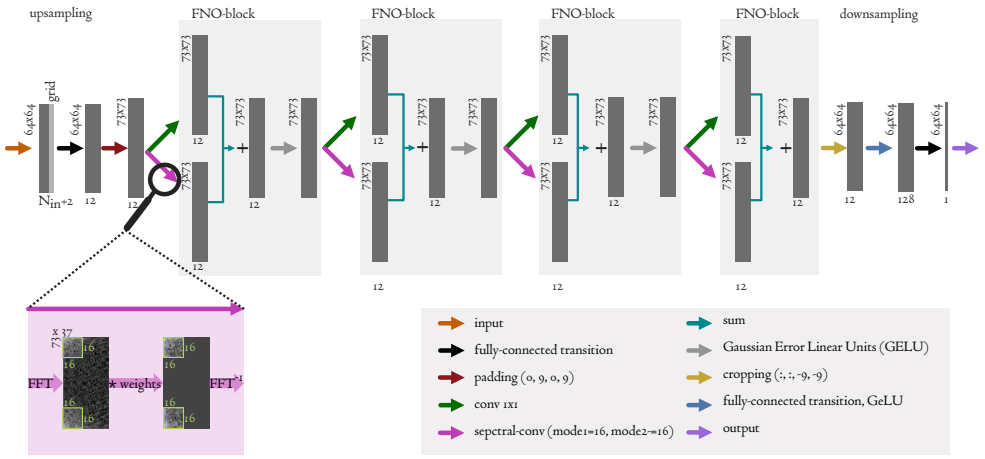

Figure 4.4.5.: Architecture of the Fourier Neural Operator (FNO)-based optical forward model. First, the images are upsampled by concatenation with the grid and a fully-connected transition. Four FNO blocks follow, each consisting of a typical convolution and spectral convolution whose outputs are added. A Gaussian Error Linear Unit (GELU) activation ensues for all except the last block. The output of the last layer is cropped and downsampled by two fully-connected transitions and a GELU activation. The spectral convolutions include a Fast Fourier Transformation (FFT), a selection of the specified Fourier modes, the multiplication with learned weights, and the inverse FFT.

Deep-learning based photoacoustic image synthesis

To provide a differentiable optical PA simulation, a FNO-based neural network was trained with Monte-Carlo-based simulations performed with one-half of the target data set (N = 10 000). First, the 10 000 graphs with their node attributes following target distributions were converted to tissue geometry masks (step c). These masks served as the input of PA simulations with SIMPA to provide data for network training. As this simulation pipeline currently requires piece-wise constant 2D tissue masks as input, the tissue geometry arrays were aggregated using the arg max operator. The optical parameters were assigned to the resulting 2D tissue geometries within SIMPA. In principle, the simulation procedure was analogous to the one described in Section 4.3.2. For example, the digital device twin of the MSOT was also positioned in the center top part of the simulation volume. However, the following settings were different. To increase the simulation speed, the 3D simulation volume was 75 mm x 0.2 mm x 60 mm

along the x-, y-, and z-axis. Additionally, only one wavelength of 750 nm was simulated with $1 \cdot 10^7$ photons. The isotropic resolution was set to $\Delta x = \Delta y = \Delta z = 0.1$ mm. The field of view was set such that the final cropped 2D slice of the simulation corresponded to the input tissue geometries. After simulation, a FNO-based neural network [Li et al., 2020c] was trained on the training split of the simulated data, similar to the work by Rix et al., 2023. The network input was an image with three channels corresponding to the log-scaled optical parameter images for absorption (μ_a^{FNO}), scattering (μ_s^{FNO}), and anisotropy (g^{FNO}). The output was the corresponding log-scaled synthesized PA image (p_0^{FNO}). More specifically, the logarithmic transformation was performed after adding a constant value of one to ensure positive values:

$$\begin{aligned} \mu_a^{\text{FNO}} &= \log_{10}(\mu_a + 1) \\ \mu_s^{\text{FNO}} &= \log_{10}(\mu_s + 1) \\ g^{\text{FNO}} &= \log_{10}(g + 1) \\ p_0^{\text{FNO}} &= \log_{10}(p_0 + 1) \end{aligned} \tag{4.3}$$

The default 2D FNO model was used. An overview of the architecture, including the dimensions of Fourier modes, is given in Figure 4.4.5.

Annotation- and image-based training paradigms

To implement the annotation- and image-based training paradigms, the processing steps described in the previous part of this section were applied. For the annotation-based training, the data described, the GNN, the graph to tissue geometry mask conversion, and the MMD loss calculation were performed. Algorithm 3 details the annotation-based training scheme. For the image-based training, the MC-based simulations and the FNO-based PA image synthesis were required in the process. The step-by-step training procedure of the image-based optimization is shown in Algorithm 4.

/* G^i are input graphs */

/* G^t are target graphs */

Previous steps:

convert target graphs G^t to tissue geometry masks X; X will be used as references for the annotation-based experiment.

Training:

for *n in epochs* **do**
  for *j in minibatches* **do**
    load input graphs G^i_j
    for *graph in* G^i_j **do**
      initialize the first four node attributes initialize Gaussian random noise
      assign the fifth node attribute according to tissue class
    end
    apply GNN: $\hat{G}_j$ = GNN(G^i_j)
    convert output graphs $\hat{G}_j$ to tissue geometry masks $\hat{X}_j$
    initialize loss L
    for *c in mutable classes* **do**
      for *a, b in kernel parameters* **do**
        compute and add loss: L += $\mathrm{MMD}^2_{a,b}$ ($\hat{X}_{c,j}$, $X_{c,j}$)
      end
    end
    optimize parameters of GNN using L
  end
end

Algorithm 3: Training workflow of annotation-based experiment.

/* G^i are input graphs */

/* G^t are target graphs */

Previous steps:

convert target graphs G^t to target tissue geometry masks X

assign optical parameters to X, which results in optical parameter images Y

Monte Carlo (MC)-based simulation:

perform MC-based optical simulation on 10 000 of the target optical parameter images Y_{MC}; this results in reference PA images P_{MC} for FNO-based training

Training of FNO-based simulation:

train the FNO-based model with Y_{MC} as input and P_{MC} as reference

Inference of FNO-based simulation:

apply FNO-based model on the remaining 10 000 target optical parameter images Y to get P; P serves as the reference for the image-based experiment

Training:

for *n in epochs* **do**

  for *j in minibatches* **do**

    load input graphs G_j^i

    for *graph in* G_j^i **do**

      initialize the first four node attributes with Gaussian random noise

      initialize the fifth node attribute according to tissue class

    end

    apply GNN: $\hat{G}_j = \text{GNN}(G_j^i)$

    convert output graphs $\hat{G}_j$ to tissue geometry masks $\hat{X}_j$

    assign optical parameters to $\hat{X}_j$ to get optical parameter images $\hat{Y}_j$

    apply FNO: $\hat{P}_j = \text{FNO}(\hat{Y}_j)$

    initialize loss L

    for *a, b in kernel parameters* **do**

      compute and add loss: $L \mathrel{+}= \text{MMD}^2_{a,b}(\hat{P}_j, P_j)$

    end

    optimize parameters of GNN using L

  end

end

Algorithm 4: Training workflow of image-based experiment.

4.4.3. Experiments

To validate the feasibility of scene graph-based generation of tissue geometries (RQ3) for PAI, two experiments with different degrees of complexity were designed (cf. Figure 4.4.3). As introduced in the previous section, one annotation-based experiment and one image-based experiment were conducted.

Annotation-based experiment

For the first *annotation-based* experiment, reference target tissue geometries were used for training. In other words, the pipeline was run only up to the generation of the tissue geometry masks (step c in Figure 4.4.3), and the generated and target class-encoded tissue geometry masks were used to compute a comparative loss (cf. Algorithm 3).

Image-based experiment

In the second *image-based* experiment, simulated target PA images were used as references during training. This means the entire pipeline (step e in Figure 4.4.3) was run, and the generated PA images were compared with the target PA images to compute the loss and to optimize the weights of the GNN (cf. Algorithm 4).

4.4.4. Experimental Conditions

This section provides specifications of the computing resources, the hyperparameters of the GNN and FNO-based neural network, the MMD calculation, and the performance assessment for both the annotation-based and image-based experiments.

Computing resources

All experiments were performed on a Ubuntu 20.04 workstation with an AMD Ryzen 9 3900x processor (12 cores), 64 GB RAM, and an NVIDIA GeForce RTX 3090 graphics card with 24 GB RAM.

Graph Neural Network and Maximum Mean Discrepancy configuration

The GNN was implemented with PyTorch geometric [Fey et al., 2019] and updated according to the MMD loss. Since the informative value of the MMD highly depends on the selected kernel parameters, a list of kernel parameters was empirically selected and verified with the help of a permutation test [Gretton et al., 2012]. For both experiments, the coefficients and exponents for the MMD were $a = 1$ and $b = \{1, 2, 3, 4\}$, respectively. Note that these settings were determined empirically and only confirmed with the permutation test. Other settings or a subset of the parameters would certainly also lead to successful optimization. Generally, choosing the kernel parameters for the MMD is an ongoing field of research. The hyperparameters of the GNN were chosen with a grid search using the validation results. For both experiments, a minibatch size of 200 was used to train the pipeline with the Rectified Adam optimizer [Liu et al., 2019] provided by PyTorch. The learning rate was decreased with progressive training using the ReduceLROnPlateau scheduler provided by PyTorch. A patience of 25 epochs, a factor of 0.5, and a minimum learning rate of $1 \cdot 10^{-7}$ was chosen. To decrease the convergence time, the outputs for vessels were adapted for a specified number of epochs. In more detail, the center widths and heights were extended by $+\,5$ px, and $+\,10$ px were added to the center position in the z-direction.

The experiment-specific settings, including the scaling factors of the MMD, are detailed in the following:

Annotation-based experiment The annotation-based experiment was trained for 300 epochs with an initial learning rate of $1 \cdot 10^{-3}$, and the output adaptations were performed for 40 epochs. For this experiment, the class-wise MMD (cf. Equation 4.1) with a scaling factor of $s = 10$ was used. It was computed separately per mutable class. This means one value was calculated for the membrane class, one for the artery class, and one for the vein class.

Image-based experiment For the image-based experiment, the initial learning rate was $5 \cdot 10^{-3}$, and the output adaptations were performed for 90 out of 200 epochs. The pipeline was trained with the MMD using a scaling factor of $s = 45$ (cf. Equation 4.2).

Fourier Neural Operator-based network

The FNO-based network was implemented with PyTorch Lightning[7]. As for the graph training, the hyperparameters of the FNO-based network were chosen based on a grid search with respect to the validation results. A learning rate of $1 \cdot 10^{-3}$ and a minibatch size of 100 were used to train the network for $40\,000$ epochs with the Adam optimizer [Kingma et al., 2014] and the MSE as loss function. The ReduceLROnPlateau scheduler provided by PyTorch was also applied here to decrease the learning rate with progressive training. A patience of 3000 steps, a factor of 0.5, and a minimum learning rate of $5 \cdot 10^{-5}$ was chosen.

Performance assessment

To assess the scene graph-based approach and the generated tissue geometries, the distributions of the geometric quantities of the learned and target data were visualized. Additionally, descriptive statistics were calculated on the distributions. The number of pixels assigned per tissue class and mask were also investigated for the two data sets.

4.4.5. Results

The ensuing section shows the results of the scene graph-based generation of tissue geometries. First, the outcomes of the annotation-based experiment are presented. Then, the results of the image-based experiment follow.

Annotation-based experiment

Qualitatively, the generated and target tissue geometries look similar (cf. Figures 4.4.6 and 4.4.7). The mutable classes membrane, artery, and vein have similar sizes and positions. However, an extensive analysis showed that two disjoint veins never appeared in the generated tissue geometries. Instead, in the case of two veins, they overlap in the generated tissue geometries.

[7] `https://github.com/Lightning-AI/lightning`

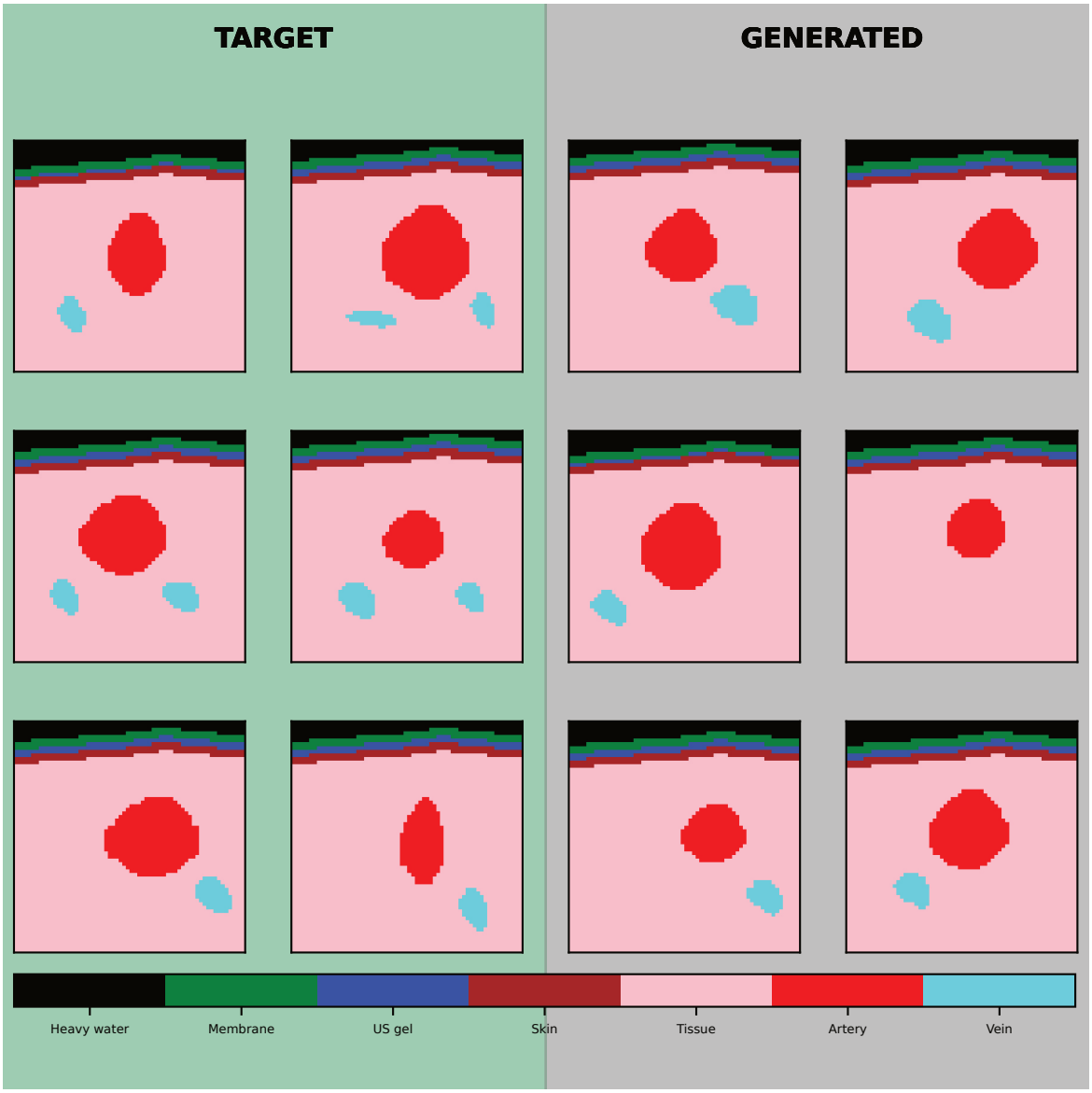

Figure 4.4.6.: Randomly chosen (unpaired) examples of (*left*) target and (*right*) generated tissue geometries of the annotation-based experiment look similar. While the size and position of the mutable classes membrane, artery, and vein match well, the generated data never contains two disjoint veins. Note that the class-encoded tissue geometry masks were aggregated with the arg max operator.

This finding is confirmed by an analysis of the number of pixels per mutable class in the generated tissue geometries (cf. Figure 4.4.8). The corresponding histograms for the classes membrane and artery of the generated data more closely resemble the ones of the target data compared to the vein histograms. The generated tissue geometries generally have fewer pixels assigned to the class vein compared to the target tissue geometries.

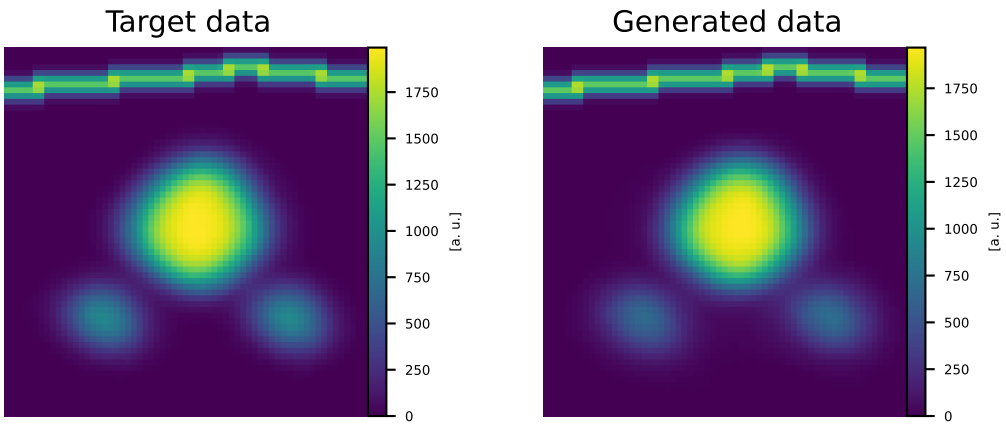

Figure 4.4.7.: Aggregated images of (*left*) target and (*right*) generated tissue geometry masks of the annotation-based experiment match. The size and position of the mutable classes are in good agreement. However, as the intensity of veins shows, veins occur less often in the generated aggregated images than in the target ones. For aggregation, the tissue geometry masks of the mutable classes membrane, artery, and vein were summed first, and the sum of these aggregated masks of all test data is shown.

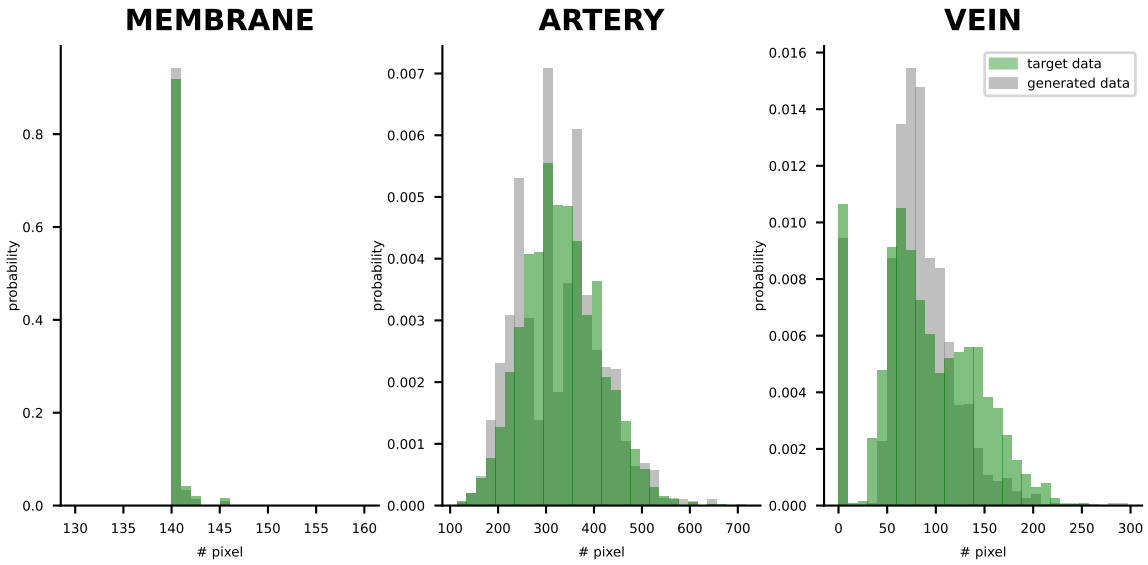

Figure 4.4.8.: Analysis of the number of pixels per mutable class in images generated with the annotation-based experiment. Overall, the number of pixels matches between the target and generated tissue geometries for (*left*) the membrane and (*center*) the artery. However, the number of (*right*) vein pixels per image is smaller in the generated tissue geometries compared to the target ones. The tissue geometry masks were aggregated with the arg max operator for this analysis.

As shown in Figure 4.4.9, the generated distributions of the geometric quantities of mutable attributes closely resemble the target distributions. The calculated mean values of the distributions coincide for all quantities within one target standard deviation. Even the bimodal distribution of the center position in the x-direction of veins was learned by the GNN. Note that descriptive statistics were calculated for negative and positive values separately in this case. The largest difference in the means exists for the width of veins.

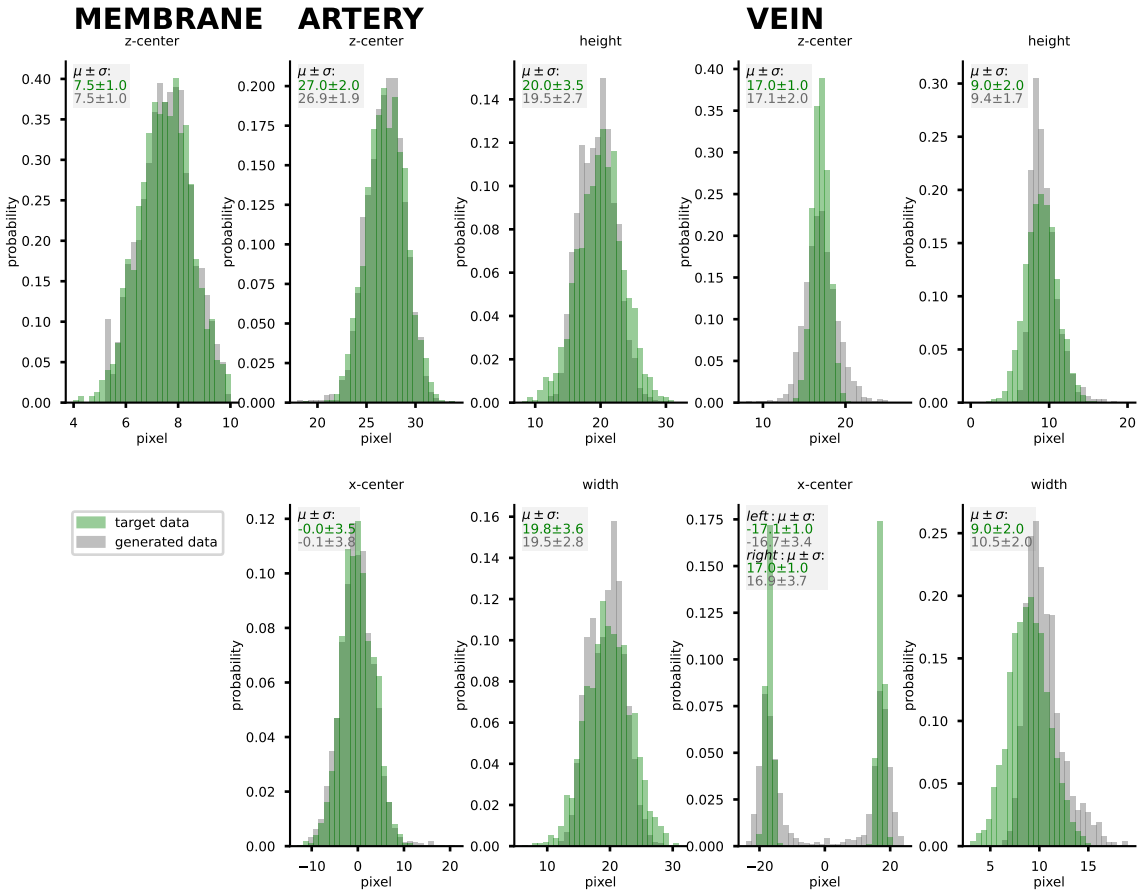

Figure 4.4.9.: Distributions of the target and generated geometric quantities closely resemble each other for the annotation-based experiment. The target and generated distributions of the optimized geometric quantities are shown for each mutable class (membrane, artery, and vein). Descriptive statistics (mean μ and standard deviation σ) of the center position in the z-direction (z-center), center position in the x-direction (x-center), width (width), and height (height) coincide within one target σ for all quantities. Note that the statistics of the veins' center positions in the x-direction were calculated separately for negative and positive values.

Image-based experiment

The findings from the annotation-based experiment hold true for the image-based experiment. As shown in Figure 4.4.10, the generated and target tissue geometries are qualitatively comparable. However, as for the annotation-based experiment, two veins always overlap in the generated tissue geometries, meaning that two disjoint veins never appear in generated images.

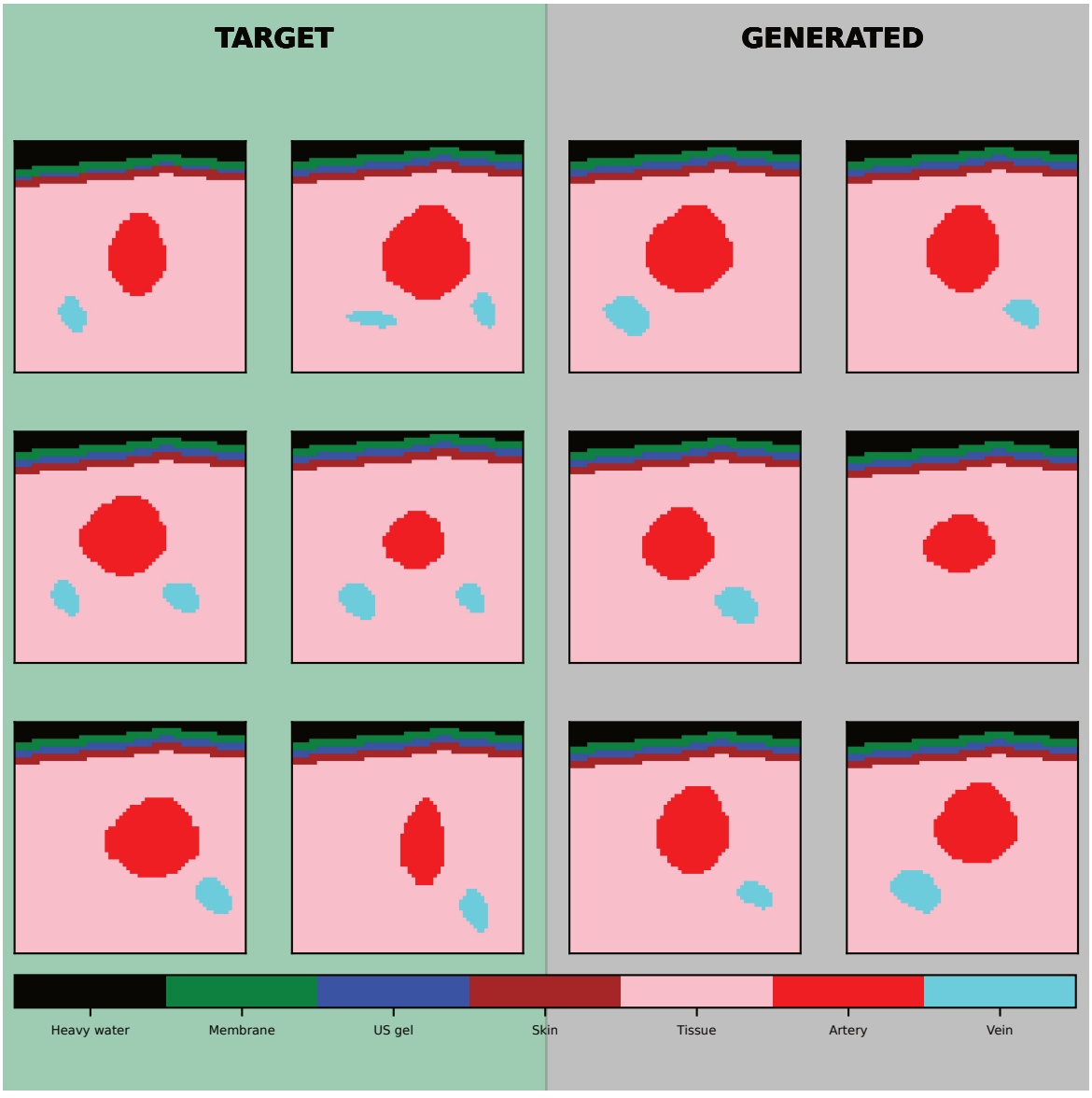

Figure 4.4.10.: Randomly chosen (unpaired) examples of (*left*) target and (*right*) generated tissue geometries of the image-based experiment are comparable. The size and position of the mutable classes membrane, artery, and vein match. However, the generated data never contains two disjoint veins. Note that the class-encoded tissue geometry masks were aggregated with the arg max operator.

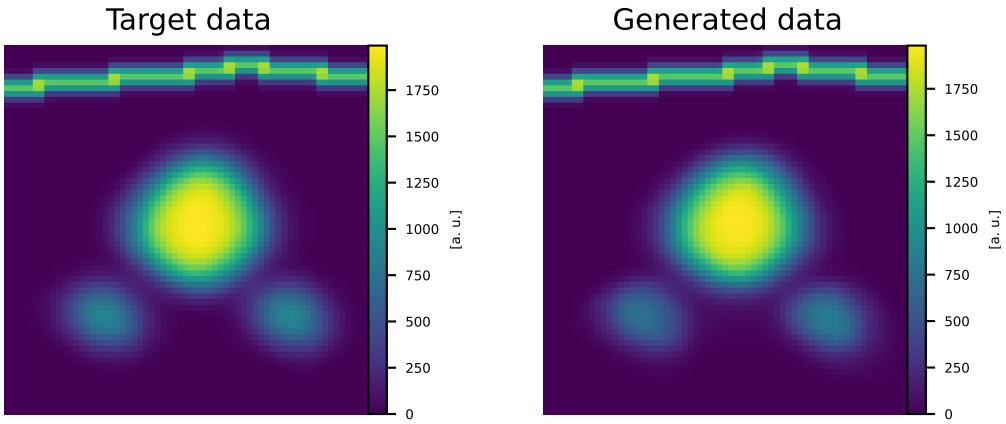

Figure 4.4.11.: Aggregated images of (*left*) target and (*right*) generated tissue geometry masks of the image-based experiment match. The size and position of the mutable classes are in good agreement. However, the intensity of veins in the generated aggregated images indicates that generated veins occur less often compared to target ones. For aggregation, the tissue geometry masks of the mutable classes membrane, artery, and vein were summed first, and the sum of these aggregated masks of all test data is shown.

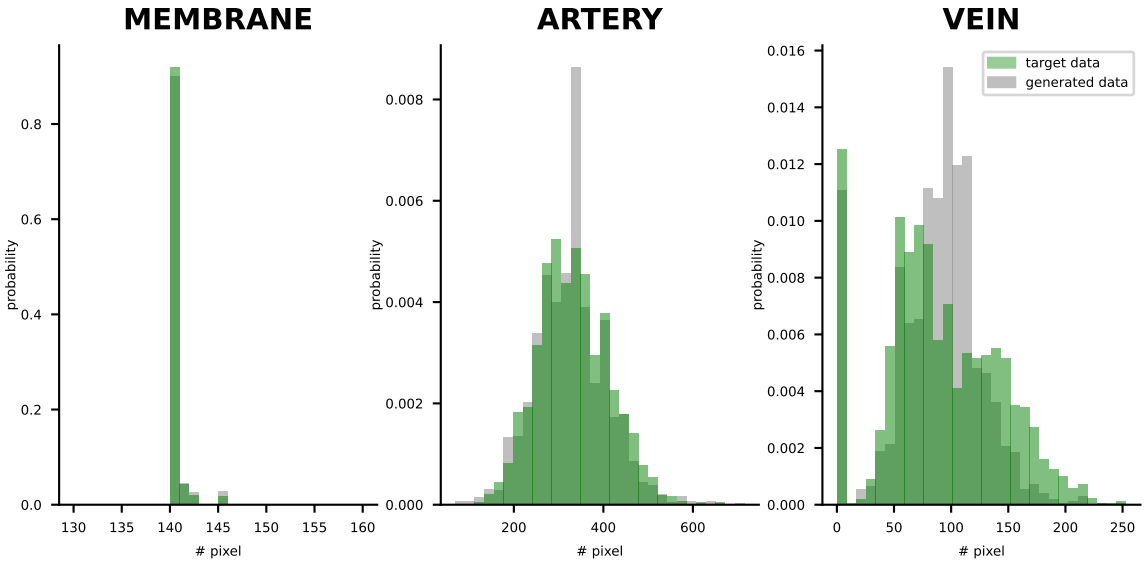

Figure 4.4.12.: Analysis of the numbers of pixels per mutable class in images generated with the image-based experiment. The number of pixels matches between the target and generated tissue geometries for (*left*) the membrane and (*center*) the artery. However, the number of (*right*) vein pixels per image is smaller in the generated tissue geometries compared to the target ones. The tissue geometry masks were aggregated with the arg max operator for this analysis.

The aggregated images (cf. Figure 4.4.11) and the analysis of the number of pixels per mutable class in the generated images (cf. Figures 4.4.12) support this insight.

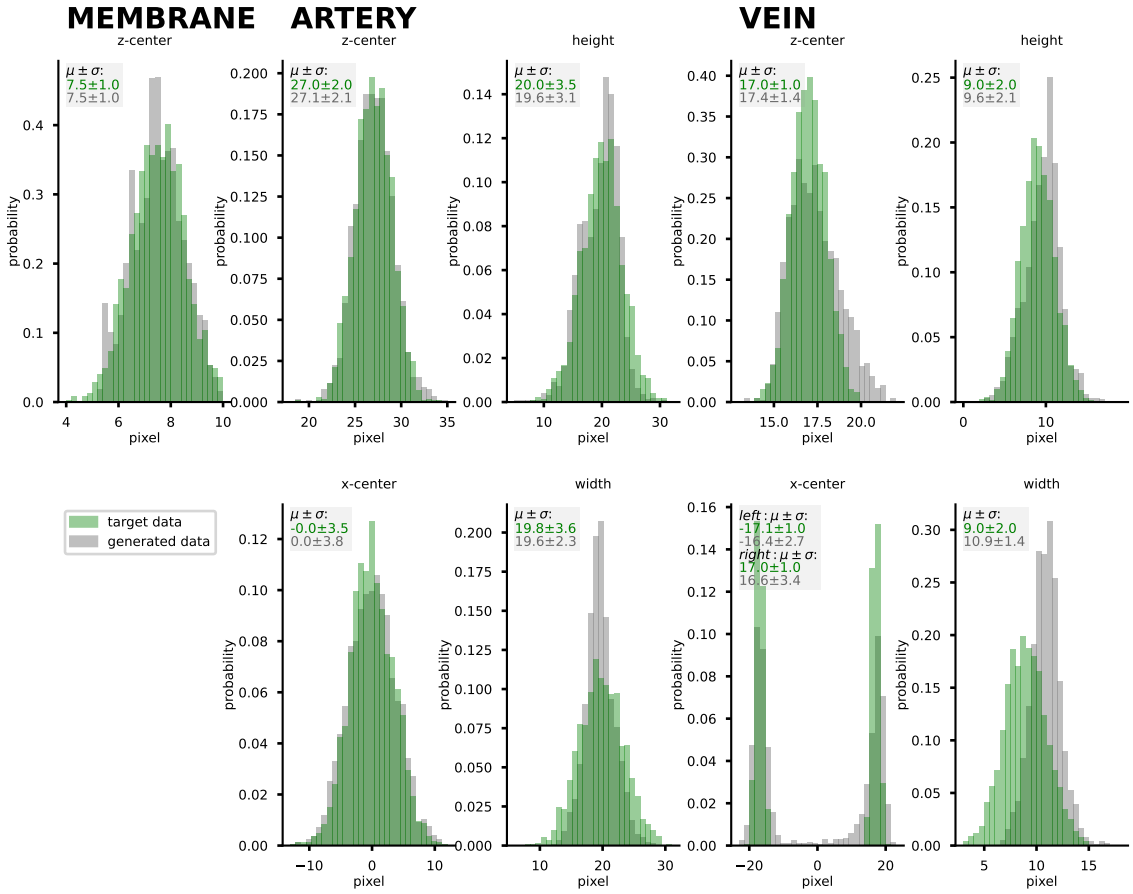

Figure 4.4.13.: Distributions of geometric quantities of the target data and the ones generated with the image-based experiment coincide. For each of the mutable classes (membrane, artery, and vein), the target and generated distributions of the optimized geometric quantities are shown. Descriptive statistics (mean μ and standard deviation σ) of the center position in the z-direction (z-center), center position in the x-direction (x-center), width (width), and height (height) coincide within one target σ. The GNN learned the bimodal distribution for the veins' center position in the x-direction. The largest difference in the means exists for the width of veins. Note that the means and standard deviations for the veins' center positions in the x-direction were calculated separately for negative and positive values.

In analogy to the annotation-based experiment, the generated distributions of the geometric quantities of mutable attributes coincide with the target distributions (cf. Figure 4.4.13). Descriptive statistics reveal that the means of the distributions overlap within one target standard deviation. The GNN learned the bimodal distribution of the center positions in the x-direction of veins. For this quantity, the means and standard deviations were calculated for negative and positive values separately. The largest difference in the means exists for the width of veins.

4.4.6. Discussion

The in silico results of the annotation- and image-based experiments indicate that tissue geometries can be generated using scene graphs, which addresses RQ3. While the success of the approach has been shown on natural images, this work demonstrates its **feasibility on PA images**, which have image properties very different from natural images.

For the **annotation-based experiment**, the pipeline was run until the tissue geometry masks were generated, which were then compared with the target masks. It turned out that overall, the generated and the target tissue geometries were similar in size and position. The distributions of the generated geometric quantities resembled the ones of the target geometric quantities. Only in rare cases did the position or size deviate from the targets, as shown by the agreement of the means of the distributions of the geometric quantities within one target standard deviation. The edge cases are represented by a difference in the standard deviations of the generated distributions compared to the target ones, which were the largest for bimodal vein quantities. This is most likely related to the bimodal distribution of the vein positions in the x-direction that made the optimization of vein quantities more difficult.

The **image-based experiment** demonstrated the feasibility of scene graph-based generation of tissue geometries purely based on target PA images, i.e., without the need for reference annotations. Compared to the annotation-based experiment, the image-based experiment was considerably more complex as it involved the DL-based forward model. Still, comparable results were achieved, and the annotation-based experiment's findings that the generated distributions of geometric quantities resemble the target ones hold true.
Although the FNO-based optical forward model was trained on data that followed the target distribution, it estimated reasonable initial pressure distributions during training that resulted in correctly optimized geometric quantities.
One important factor for the convergence of this experiment was to compute the loss between log-scaled images, which kept the ranges of values of the images in the same order of magnitude. For future applications where signal attenuation is greater with image depth, fluence corrections could become important.

In both experiments, the **bimodal distribution** of the center positions in the x-direction for veins was explicitly learned. However, two disjoint veins were never present within a single converted tissue geometry mask. In other words, in the case of a scene graph with two veins, the GNN estimated similar center positions in the x-direction for both veins. This led to overlapping veins, as the analysis of the number of vein pixels per image demonstrated. One reason for the

occurrence of overlaps could be that in the current implementation, including the clamping of the generated tissue geometry masks, cases with overlaps can result in the same tissue geometry masks as ones with only one vein. However, limitations related to overlaps were also mentioned in the publication by Kar et al., 2019, which this work is based on. This means that overlaps seem to be a general challenge with this approach. In future work, adjustments to the graph structure, node attributes, or adding a loss term that accounts for node interdependencies could be investigated to overcome this issue.

In **comparison to the GAN-approach**, this graph-based approach revealed slower training by a factor of approximately 16, even though the tissue model was simpler and the images were smaller by a factor of 8. This could be due to the fact that GNNs are generally slow by nature [Kose et al., 2022], and the inherent graph to tissue geometry mask conversion was implemented in a sequential manner. Parallelization or matrix-wise operations could help to reduce the training time.
Compared to GANs, the major advantage of this approach is that it offers the explicit learning of the distributions of the geometric quantities, which in the long run provides higher transparency and interpretability of important geometric parameters (especially compared to literature values).

One limitation of the approach is the **high computational cost**. On the one hand, the MMD is a statistics-based metric that benefits from larger minibatch sizes. On the other hand, the different networks with a larger minibatch size require more RAM on the GPU. Even though the GNN and FNO-based models are rather small (~ 5 GB), initial experiments that computed the loss of the image-based experiment in the latent space of a pre-trained inception network limited the minibatch size to 150 with the given image sizes of 64 x 64 px, one wavelength, and a GPU with 24 GB RAM.

While the results showed that the generated and target distributions could be well matched by using the MMD, the **informative value of the MMD** is highly dependent on the choice of the kernel [Gretton et al., 2012, Li et al., 2020c]. In this thesis, the kernel and associated parameters were chosen empirically and verified by a permutation test. However, more sophisticated methods for kernel selection, such as learning-based methods proposed by Biggs et al., 2023, could be investigated in the future.
When considering larger PA images with multiple wavelengths, the amount of data required for reliable kernel embeddings increases. Therefore, computing the MMD in a latent space might become beneficial [Li et al., 2020c], for example, by using an autoencoder. The work by Kar et al., 2019 followed this principle and computed the MMD in a latent space of a pre-trained

classification network. In addition, other loss functions, such as the Kullback-Leibler divergence that enables maximum likelihood training, could be explored, or one could add a downstream task in the optimization procedure, as performed in the work by Kar et al., 2019.

This scene graph-based approach shows promising results and is considered to be **easily extendable** by adding nodes and node attributes. Applying this technique to in vivo images, however, requires the solution of several challenges:
First, a more complex setting needs to be investigated. This setting would include an extended and more interconnected probabilistic grammar and, correspondingly, more complex graph trees. The work by Kar et al., 2019 could, however, show that this approach is well suited to deal with complex scenes and environments, emphasizing the versatility of the approach. Biological tissue is inherently highly heterogeneous and complex. Thus, finding a good approximation for the probabilistic grammar will likely depend on the specific application and potential downstream task. Another option could be to leverage data from other imaging modalities of the regions of interest to train methods on these images to generate scene graphs directly, circumventing the need for a probabilistic grammar. This study employed a shape dictionary to represent the shapes of the vessels and the layered tissue structures. For a more complex setting, this simple yet elegant solution could easily be extended to contain additional shapes based on literature knowledge or manual annotations. A graph attribute could be introduced that selects the most appropriate shape from the dictionary to increase the realism of the generated tissue geometries. Learning a parametric representation of the shapes is also an option that could be pursued in the future.
Second, the DL-based forward model was trained on data corresponding to the given target distributions. This introduces some prior which is not available in practice. However, the FNO-based forward model showed good generalizability when facing out-of-distribution data during the training process. Therefore, the FNO-based model could perform well even when trained on rough approximations of the target distributions or with a broad range of geometric parameter combinations.
Third, the remaining components of the PA simulation pipeline, such as the choice of optical and acoustic parameters, a differentiable acoustic forward simulation, noise modeling, and reconstruction, would likewise need to be incorporated to allow the approach to be applied to in vivo data. However, the realistic modeling of all components of the image formation pipeline in a differentiable manner is an active area of research [Bench et al., 2023]. Still, one could extend the approach by sampling the optical parameters from a literature-based distribution or modeling them as additional node attributes. The last of these options would add great value to the PAI field since there is so far no established method that learns the distribution of both

anatomical and optical (and acoustic) parameters in a disentangled manner.
The last challenge is related to the required number of training data. In this initial study, several thousand images were used with minibatch sizes in the order of 100. However, the minimal amount of data that is required was not investigated specifically. In general, open-source data is one of the limiting factors for data-driven PA image analysis in practice [Assi et al., 2023]. Therefore, it is of great importance to conduct future studies that provide the acquired data, preferably in a standardized form.

In short, this proof-of-concept study successfully showed the general feasibility of leveraging scene graphs for automatic tissue geometry generation. Although there are several hurdles to overcome before this approach can be applied to in vivo data, its versatility makes it promising for continued research.

5. Discussion

This thesis investigated whether DL-based methods can be leveraged to generate plausible tissue geometries (cf. RQ1 - RQ3 in Section 1.3) that can eventually improve the realism of PA simulations. While the details of the three methods are discussed in their respective sections, this chapter focuses on the general aspects.

Innovative concept This thesis pioneers the use of data-driven techniques to generate realistic tissue geometries for PA image analysis. The key innovation was to leverage neural networks as well as acquired PA images or patterns derived from them. This concept contrasts with previous studies in this area, which often rely on simulated data but place little emphasis on the underlying anatomical realism. Tissue geometry generation in the PAI domain is typically based on conventional methods, such as numerical pattern phantoms or model-based strategies, that lack realism and broad applicability. However, ensuring such realism is pivotal for certain research questions, such as the DL-based solution to the quantification problem.
Three distinct methods of varying complexity, building upon each other, were developed. The first approach for the automatic extraction of tissue geometries was considered a preliminary step of the thesis and was based on DL-based semantic segmentation, techniques that are already widespread in the PAI field. However, this work stands out as the first to enable automatic multi-label semantic segmentation of PA images. The two more advanced methodologies introduced an entirely novel *learning-to-simulate* approach to the field. Based on generative models that are successfully and frequently applied in other fields, these methods allowed the generation of new tissue geometries that match the distribution of the training data set. The results of these advanced approaches are indicative of the potential that DL-based techniques hold in realistically synthesizing PA images. Due to the inherent differentiability of these methods, this thesis is the first milestone towards achieving the overarching vision this thesis is part of, an encoding-decoding scheme for PA image generation and quantification.

Comparison of three approaches The results of all three approaches related to RQ1 - RQ3 confirmed the feasibility of generating realistic tissue geometries in an automatic and data-driven manner.

The first approach of semantic segmentation (addressing RQ1) is not a classic generation approach but was considered a preliminary step of the work. It confirmed the fundamental question of whether the anatomical information can be estimated from a set of acquired images using DL. In the future, this approach may replace the manual annotations that served as the training data for the two generative modeling approaches (addressing RQ2 and RQ3). Broadly speaking, the semantic segmentation approach could be interpreted as a naive modeling approach to tissue geometries. Each of the three techniques has its own strengths and weaknesses, especially with respect to inference requirements (cf. Figure 1.3.1). Unlike the generative modeling approaches, semantic segmentation requires real PA images to provide the paired image-tissue geometry information during inference. Thus, this "modeling approach" is limited by the number of available real images.

In contrast, the GAN-based approach requires no data or prior knowledge for inference. It generates any number of entirely new tissue geometries that match the data distribution of the training reference annotations from Gaussian-distributed noise alone. However, its primary limitation is that GANs are considered black boxes, lacking an inherent mechanism to understand the underlying geometric quantities thoroughly.

Contrarily, the scene graph-based approach leveraging prior knowledge about tissue composition allows for more insight concerning the distributions of geometric quantities that were explicitly and correctly learned. Thus, this model serves as a bridge between the two preceding methodologies. While it necessitates prior knowledge about the tissue scene's hierarchical composition, it autonomously generates the explicitly encoded geometric quantities from Gaussian-distributed noise during inference in analogy to the GAN. Moreover, this technique could be performed without the need for reference annotations, relying solely on PA reference images, giving it an advantage over the GAN-based approach. However, it is noteworthy that extending the GAN strategy with additional simulation steps is feasible, with ongoing research in this vein known under the term *ambient GANs*. Broadly speaking, bypassing the need for reference annotations for the generation of tissue geometries holds significant promise by sidestepping the uncertainties and time-intensive nature of manual annotations.

Manual annotation quality One limitation of this work is that there is no guarantee that the reference annotations resemble realistic anatomy. Within an inter-rater reliability analysis for the tissue class blood, it became clear that manual annotation of PA and US images can be

challenging and error-prone (cf. Section 4.2.6). Although a standardized annotation protocol was followed, annotations of vessels differed in size, location, and number. Especially due to the decreasing light fluence and SNR with tissue depth, annotation of PA images requires considerable domain expertise and can still be ambiguous.
In general, there are several ways to approach this challenge. For example, one could register PA and US images with even more imaging modalities. Alternative approaches are to develop weakly supervised [Ren et al., 2020, Pan et al., 2021], completely unsupervised [Yuan et al., 2020, Ji et al., 2019, Cho et al., 2021], or self-supervised [Caron et al., 2021, Singh et al., 2018] semantic segmentation methods for the given data that have shown impressive results in computer vision in the last couple of years. There are also developments where a pre-trained GAN [Karras et al., 2019, Karras et al., 2020a] was used to generate new images along with associated semantic segmentations. For this purpose, the feature space was analyzed semi-supervised or completely unsupervised [Li et al., 2021a, Zhang et al., 2021e, Pakhomov et al., 2021]. However, it is not trivial to expect these methods to work with PA images because signal intensity within a tissue class can attenuate greatly with depth due to the absorption and the effect of fluence.
The second experiment of the scene graph-based approach shows another promising approach to dealing with annotation uncertainty. The incorporation of PA simulation steps in the optimization, such as the DL-based optical PA simulation here, allows for training with solely reference PA images and thus eliminates the need for reference annotations. However, any additional factors contributing to the simulation must be known in advance or are expected to be approximated in a sufficiently realistic manner. The findings from the scene graph-based experiments, in conjunction with the parallel research by Bench et al., 2023, indicate the feasibility of generating PA images without relying on reference annotations. While these techniques bear certain constraints, continued research in this direction has the potential to enhance PA image synthesis and yield a more comprehensive grasp for the annotations.

Sparsity of tissue geometries One limitation is that the tissue geometries used in this thesis so far are piecewise constant and limited to a maximum of nine tissue classes. There were several reasons for this. First, these structures were relatively easy to recognize for manual annotation, and second, literature-based optical and acoustic parameters for PA simulation were (mostly) known for these classes. However, depending on the application, real images comprise more structures, such as muscles, glands, and fascia, and additional structures that might not be visible in PAI due to lack of absorbing chromophores but still affect the acoustic wave propagation. In addition, the tissue includes inhomogeneous structures that are smaller than the given image resolution, such as arterioles, venules, and capillaries within the background, connective tissue,

and the epidermis and dermis within the skin. Adding spatial variations of optical and acoustic properties within the corresponding structures according to appropriately chosen ranges of values could compensate for this limitation in the future.

Two-dimensionality of tissue geometry modeling Furthermore, the tissue modeling presented in this thesis was limited to 2D due to the fact that the MSOT only provides 2D images. Nonetheless, modeling 3D tissue context enhances the simulations due to the more realistic out-of-plane signal formation. Along these lines, 3D PA imaging would be required, which is an active area of research with many potential applications [Lee et al., 2020, Holzwarth et al., 2021b, Jiang et al., 2022]. For example, 3D imaging could overcome limitations of 2D imaging, such as the subjectivity of the operator's examination and the limited reproducibility of a particular image [Fenster et al., 1996]. Furthermore, the 3D context of PA images is expected to improve the quantification performance [Bench et al., 2020].

Validation in simulated domain One crucial limitation of this thesis is that the validation of the realism of tissue geometries was performed in the simulation domain only, which does not allow for drawing broad conclusions in a clinical setting. There are various reasons for this. Firstly, validating estimates of real measured data without GT values (chicken-and-egg dilemma) is an area of active research [Hübner et al., 2023, Gröhl et al., 2023b]. Secondly, the focus of this thesis was the automated modeling of tissue geometries, which alone are not sufficient as a basis for PA simulations. For a comparison with reconstructed real data, among others, the optical as well as acoustic tissue parameters must be calibrated. Furthermore, it must be ensured that the device-specific noise model, the Grüneisenparameter, the impulse response of the transducer, and the laser center wavelength variation, to name a few, correspond to reality. Thus, validation with real data was hampered by the lack of knowing the simulation model's factors precisely. Active research investigates these factors, but no broadly applicable simulation model have been identified so far [Bench et al., 2023, Dreher et al., 2023].

Potential of scene graphs The scene graph-based approach shows a high potential for addressing some of the mentioned limitations in future work since scene graphs can, in principle, become extremely complex with many nodes and node attributes. For example, one could extend the probabilistic grammar by including additional tissue classes and structures to generate graphs with, accordingly, additional nodes. Furthermore, additional node attributes could account for tissue inhomogeneities. One could even reformulate the approach to allow the

nodes to represent 3D rather than 2D structures or rephrase the grammar to account for 3D tissue geometries. Additional node attributes could be used to optimize optical and acoustic tissue properties simultaneously. To apply the approach to real data, one would need to add the acoustic simulation and reconstruction to the optimization pipeline, both of which are differentiable in this scenario. This could allow a comparison between the generated and the measured PA images in both the raw time series and reconstructed domains.

Clinical perspective The techniques presented in this thesis have potential significance for a clinical perspective of PAI. First, the ability to extract tissue geometries can reveal anatomical shifts indicative of various diseases. The semantic segmentation of vessels, displayed with high image contrast in PAI, could be crucial for the diagnosis of cardiovascular diseases. In addition, the graph-based method could potentially be used to determine the distribution of anatomical markers of diseases by analyzing pathological data sets. Finally and most importantly, the generation of realistic tissue geometries plays a pivotal role in advancing the realism of PA simulations. Though the scope of this thesis was limited to data from healthy subjects, the methods showcased are versatile and adaptable to pathological tissues when performed with PA images from affected patients and the corresponding annotations. Once further gaps in the sim-to-real domain gap of PAI are addressed, this work could contribute to solving the quantification problem through data-driven strategies. This quantification and, thus, determination of clinically relevant physiological parameters in vivo and non-invasively represents a unique capability that could revolutionize health care.

6. Summary

PAI is an emerging imaging modality that combines optical with acoustic (US) imaging and has the potential to provide morphological and physiological tissue properties in depths of several centimeters. Especially the physiological properties, such as sO_2 and BVF, are relevant for various diseases, e.g., for the diagnosis and therapy response monitoring of cancer. However, the quantification of the concentration of different chromophores and related physiological properties from PA images involves solving two inverse problems, namely the acoustic and the optical inverse problem, and is a topic of ongoing research in PAI.

Data-driven methods are an important part of active research for solving these inverse problems and achieving qPAI in vivo. Yet, they usually require GT labels for which there are to date no gold standard methods that provide these properties in vivo. The typical approach is, therefore, to use simulated training data that exhibit these GT properties. To date, however, these data-driven approaches have not yet been proven to be reliable for in vivo applications and, in some cases, provide estimates that poorly correlate with clinical expectations. Thus, a major hurdle to overcome is the domain gap between PA simulations and real-world measurements.

This thesis is part of a novel, larger effort to investigate whether data-driven methods can improve the realism of in silico PA images and thus enable quantitative PAI. An important innovative step of this approach was to disentangle the different factors contributing to image formation. This disentanglement allows for individual optimization and analysis of each image formation component. Specific to this thesis, the focus was solely on the realistic and automatic modeling of tissue geometries, which describe the morphologies of different tissue types and serve as the basis for PA simulations. To this end, three research questions (RQ1 - RQ3) were investigated, each by a specifically designed approach.

6.1. Summary of Contributions

This thesis presents data-driven approaches to tissue geometry generation for PAI for the first time. In particular, the following contributions have been made:

Contribution 1: Tissue Geometry Estimation with Neural Networks

The basis of this thesis constituted an approach to semantic segmentation of PA images using discriminative networks (cf. RQ1). Although there is similar work on the semantic segmentation of PA images, this work is the first on multi-label semantic segmentation of MSOT images. Two neural networks of different input granularity types, a FCNN and the nnU-Net, were trained on multi-spectral PA images and the corresponding manual reference annotations. Note that the nnU-Net, which won various biomedical segmentation challenges, was applied in the PAI domain for the first time within this approach. Although the number of available data was limited, the feasibility of the approach was successfully demonstrated, with both networks resulting in plausible segmentations and overall high overlap- and distance-based metric values. This held true for an in-distribution test data set (baseline experiment) and when applied to test images from body regions other than the ones included in the training (robustness experiment). A comparison of the network types revealed that spatial context is valuable for segmentation performance (DSCs for nnU-Net: 0.85; for FCNN: 0.66). Within this analysis, a comparison of networks trained solely on co-registered US images made particularly clear that the multispectral nature of PA images is of high importance for semantic segmentation for some tissue classes, such as blood (DSCs for PAUS nnU-Net: 0.74, for US nnU-Net: 0.32). A limitation of the approach was its high dependence on the annotation quality, and an inter-rater reliability study for the annotation class blood highlighted the inherent difficulties in annotating PA images. Overall, though, the results of this approach have indicated that DL-based semantic segmentation could replace manual annotations of PA images in the future. This work was published in the *Photoacoustics* journal [Schellenberg et al., 2022b].

Contribution 2: Tissue Geometry Generation with Generative Adversarial Networks

In order to generate any number of tissue geometries, GANs were leveraged to augment a small set of manual reference annotations ($N = 78$) in an automatic fashion. The approach of

realistically learning the anatomical component of PA imaging is itself a conceptual innovation. Even though GAN-based image synthesis is popular in other domains, the specific application of GANs for tissue geometry synthesis has not been addressed before in the field of PAI.
A comparative assessment of the performance of a downstream quantification task trained on simulated PA images based on different approaches to tissue geometry generation (annotation-, GAN-, and literature-based) successfully validated the plausibility of the generated geometries. In addition, the downstream task performances demonstrated the added benefit of a GAN-based augmentation in a realistic setting, especially when compared to a model that derived tissue geometries from literature knowledge. With this approach, it was shown that tissue geometries could be realistically learned, which might potentially redefine the current benchmarks for tissue geometry modeling. This work resulted in a publication in the *Photoacoustics* journal [Schellenberg et al., 2022a].

Contribution 3: Tissue Geometry Generation with Scene Graphs

The third contribution was two-fold: the aim was to generate realistic tissue geometries and to gain insights into pivotal geometric parameters. A novel approach was conceived by adapting a scene graph-based framework, originally developed in the field of computer vision, for PAI-oriented tissue geometry synthesis. It is noteworthy to mention that while scene graphs have proven their standing in computer vision, their use in the PA community is nascent, highlighting the innovation of this approach and introducing an entirely novel perspective previously unexplored for the purpose of generating tissue geometries.
The core of the concept lies in leveraging available prior knowledge about the hierarchical structure of the general tissue composition of the PA images. This knowledge encoded in scene graphs allowed the realistic generation of tissue geometries and, in parallel, the explicit learning of the underlying distribution of interpretable geometric quantities.
Two experiments with different levels of complexity successfully showed the feasibility of generating tissue geometries with explicitly learned geometric quantities that followed target distributions. Of significance was the image-based experiment, relying solely on reference PA images and circumventing the need for time-intensive and often ambiguous manual annotations. For this purpose, a DL-based forward simulation was included in the training paradigm. The integration of DL-based simulations into optimization paradigms has been reported in the literature before. Yet, this graph-based formulation offers a new perspective and shows the great potential of deciphering node attribute distributions facilitating analyses on annotation uncertainty, for example.

While this approach has so far performed exclusively in silico and still has some significant limitations, such as the unmet multimodality of two veins in an image and the dependence on the generalizability of the DL-based simulation model, to name a few, its potential remains substantial for one reason: its versatility. The versatility of the model, both by extending the topology of graphs and by scaling the number of node attributes, potentially allows for the synthesis of tissue geometries that are interpretable and beyond that solely rely on in vivo reference PA images - a groundbreaking step in the field of PAI.

6.2. Conclusion

An unanswered question in PAI is the quantification of the concentration of different chromophores and related physiological properties in the tissue. Active research is conducted to investigate whether data-driven methods can solve the underlying inverse problems to achieve qPAI in vivo. However, a major hurdle to overcome in this regard is the domain gap between PA simulations and real measurements. This work pursues a novel approach that disentangles the various factors contributing to PA image formation and formulates image formation and quantification as one joint data-driven framework.

In conclusion, this dissertation makes a significant contribution by introducing three data-driven techniques specific to the challenging task of modeling realistic tissue geometries. This effort not only enhances the realism of PA simulations but may also streamline PA image analysis and pave the way for the overarching goal of qPAI in the long run. Tissue geometries serve as the first step of PA image formation, and their realistic modeling is therefore of great importance. However, it is important to note that they represent only one out of several components for PA image formation. As such, a significant need for research and development remains in order to realize a robust, fully data-driven, and realistic simulation pipeline. Yet, a growing body of work under the keyword *learning to simulate* explores neural networks as a replacement for simulation building blocks, and not just in the PAI domain, indicating a promising future for this line of research. Although there are still other challenges with respect to the whole concept, such as error propagation when integrating individual neural networks and validation without GTs, the promising findings and concepts embodied in this thesis constitute a milestone in this direction and, more generally, a pioneering advance in the field of PA image analysis.

6.3. Publications

This section lists the publications authored during the time of this thesis. First, a list of first-authorship publications is given. This list is followed by a collection of co-authorship publications. Each listing is divided into peer-reviewed journal publications and *other* publications, which include non-peer-reviewed posters, talks, and patents.

Peer-reviewed first-authored journal publications

- **Schellenberg, Melanie**, Dreher, Kris K, Holzwarth, Niklas, Isensee, Fabian, Reinke, Annika, Schreck, Nicholas, Seitel, Alexander, Tizabi, Minu D, Maier-Hein, Lena, and Gröhl, Janek (2022b). "Semantic segmentation of multispectral photoacoustic images using deep learning". In: *Photoacoustics*. Vol. 26, p. 100341. DOI: 10.1016/j.pacs.2022.100341.

- **Schellenberg, Melanie**, Gröhl, Janek, Dreher, Kris K, Nölke, Jan-Hinrich, Holzwarth, Niklas, Tizabi, Minu D, Seitel, Alexander, and Maier-Hein, Lena (2022a). "Photoacoustic image synthesis with generative adversarial networks". In: *Photoacoustics*. Vol. 28, p. 100402. DOI: 10.1016/j.pacs.2022.100402.

Other first-authored publications

- **Schellenberg, Melanie**, Gröhl, Janek, Dreher, Kris, Holzwarth, Niklas, Tizabi, Minu D, Seitel, Alexander, and Maier-Hein, Lena (2021). "Generation of training data for quantitative photoacoustic imaging". In Proceedings of: *Photons Plus Ultrasound: Imaging and Sensing*. Vol. 11642. International Society for Optics and Photonics, 116421J. DOI: 10.1117/12.2578180.

- Gröhl, Janek, **Schellenberg, Melanie**, Dreher, Kris K, Holzwarth, Niklas, Tizabi, Minu D, Seitel, Alexander, and Maier-Hein, Lena (2021d). "Semantic segmentation of multispectral photoacoustic images using deep learning". In: *Photons Plus Ultrasound: Imaging and Sensing*, 116423F. DOI: 10.1117/12.2578135.

Peer-reviewed co-authored journal publications

- Gröhl, Janek, Dreher, Kris K, **Schellenberg, Melanie**, Seitel, Alexander, and Maier-Hein, Lena (2021a). "SIMPA: an open source toolkit for simulation and processing of photoacoustic images". In Proceedings of: *Photons Plus Ultrasound: Imaging and Sensing*. Vol. 11642. International Society for Optics and Photonics, p. 116423C. DOI: 10.1117/1.JBO.27.8.083010.

- Gröhl, Janek, **Schellenberg, Melanie**, Dreher, Kris, and Maier-Hein, Lena (2021b). "Deep learning for biomedical photoacoustic imaging: a review". In: *Photoacoustics*. Vol. 22, p. 100241. DOI: 10.1016/j.pacs.2021.100241.

- Holzwarth, Niklas, **Schellenberg, Melanie**, Gröhl, Janek, Dreher, Kris, Nölke, Jan-Hinrich, Seitel, Alexander, Tizabi, Minu D, Müller-Stich, Beat P, and Maier-Hein, Lena (2021b). "Tattoo tomography: freehand 3D photoacoustic image reconstruction with an optical pattern". In: *International Journal of Computer Assisted Radiology and Surgery*. Vol. 16, pp. 1101–1110. DOI: 10.1007/s11548-021-02399-w.

- Dreher, Kris K, Ayala, Leonardo, **Schellenberg, Melanie**, Hübner, Marco, Nölke, Jan-Hinrich, Adler, Tim J, Seidlitz, Silvia, Sellner, Jan, Studier-Fischer, Alexander, Gröhl, Janek, et al. (2023). "Unsupervised domain transfer with conditional invertible neural networks". In: *arXiv preprint arXiv:2303.10191*. - accepted at MICCAI2023

- Yamlahi, Amine, Tran, Thuy Nuong, Godau, Patrick, **Schellenberg, Melanie**, Michael, Dominik, Smidt, Finn-Henri, Noelke, Jan-Hinrich, Adler, Tim, Tizabi, Minu Dietlinde, Nwoye, Chinedu, et al. (2023). "Self-distillation for surgical action recognition". In: *arXiv preprint arXiv:2303.12915*. - accepted at MICCAI2023

- Rix, Tom, Dreher, Kris K, Nölke, Jan-Hinrich, **Schellenberg, Melanie**, Tizabi, Minu D, Seitel, Alexander, and Maier-Hein, Lena (2023). "Efficient photoacoustic image synthesis with deep learning". In: *Sensors*. Vol. 23, no. 16, p. 7085. DOI: 10.3390/s23167085.

- Holzwarth, Niklas, **Schellenberg, Melanie**, Gröhl, Janek, Dreher, Kris, Nölke, Jan-Hinrich, Seitel, Alexander, Tizabi, Minu D, Müller-Stich, Beat P, and Maier-Hein, Lena (2023a). "Abstract: tattoo-tomographie". In Proceedings of: *Bildverarbeitung für die Medizin*. Springer, pp. 114–114. DOI: 10.1007/978-3-658-41657-7_25.

- Nwoye, Chinedu Innocent, Yu, Tong, Sharma, Saurav, Murali, Aditya, Alapatt, Deepak, Vardazaryan, Armine, Yuan, Kun, Hajek, Jonas, Reiter, Wolfgang, Yamlahi, Amine, ..., **Schellenberg, Melanie**, ..., et al. (2023). "CholecTriplet2022: Show me a tool and tell me the triplet — An endoscopic vision challenge for surgical action triplet detection". In: *Medical Image Analysis*. Vol. 89, p. 102888. DOI: https://doi.org/10.1016/j.media.2023.102888.

- Maier-Hein, Lena, Wagner, Martin, Ross, Tobias, Reinke, Annika, Bodenstedt, Sebastian, Full, Peter M, Hempe, Hellena, Mindroc-Filimon, Diana, Scholz, Patrick, Tran, Thuy Nuong, ..., **Schellenberg, Melanie**, ..., et al. (2021). "Heidelberg colorectal data set for surgical data science in the sensor operating room". In: *Scientific Data*. Vol. 8, no. 1, p. 101. DOI: 10.1038/s41597-021-00882-2.

Other co-authored publications

- Holzwarth, Niklas, **Schellenberg, Melanie**, Gröhl, Janek, Dreher, Kris K, Nölke, Jan-Hinrich, Biegger, Philipp, Tizabi, Minu D, Seitel, Alexander, and Maier-Hein, Lena (2021a). "Tattoo tomography: an optical pattern approach for context-aware photoacoustics". In Proceedings of: *Photons Plus Ultrasound: Imaging and Sensing*. Vol. 11642. International Society for Optics and Photonics, p. 1164217. DOI: 10.1007/s11548-021-02399-w.

- Vieten, Patricia, Dreher, Kris K, Holzwarth, Niklas, **Schellenberg, Melanie**, Nölke, Jan-Hinrich, Seitel, Alexander, Gröhl, Janek, Rachel, Zoë, Siea, Andrei, Held, Thomas, et al. (2022). "Deep learning-based semantic segmentation of clinically relevant tissue structures leveraging multispectral photoacoustic images". In Proceedings of: *Photons Plus Ultrasound: Imaging and Sensing*. International Society for Optics and Photonics, PC11960oP. DOI: 10.1117/12.2608616.

- Rix, Tom, Hübner, Marco, Dreher, Kris K, Nölke, Jan-Hinrich, Ayala, Leonardo, **Schellenberg, Melanie**, Sellner, Jan, Seidlitz, Silvia, Studier-Fischer, Alexander, Müller-Stich, Beat, et al. (2022). "Deep learning for spectral image synthesis". In Proceedings of: *Multimodal Biomedical Imaging XVII*. International Society for Optics and Photonics, PC11952oI. DOI: 10.1117/12.2608622.

- Holzwarth, Niklas, Dreher, Kris, **Schellenberg, Melanie**, Nölke, Jan-Hinrich, Gröhl, Janek, and Maier-Hein, Lena (2023c). *Method and system for context-ware photoacoustic imaging*. US Patent App. 18/004,689.

- Holzwarth, Niklas, Staus, Marcella, Günther, Josefine, Calderazzo, Silvia, **Schellenberg, Melanie**, Lang, Werner, Maier-Hein, Lena, Rother, Ulrich, and Seitel, Alexander (2023b). "Clinical tattoo tomography". In Proceedings of: *Photons Plus Ultrasound: Imaging and Sensing*. International Society for Optics and Photonics, PC1237930. DOI: 10.1117/12.2649870.

7. Supplemental Material

This chapter contains additional material supplementing the thesis. The structure is similar to the thesis, meaning that the additional material belonging to the photoacoustic data (cf. Section A) or to one of the RQs (cf. Section 1.3) is presented in a corresponding chapter. First, the link to the annotation protocol is provided. Then, supplementary results on semantic segmentation of PA images are presented (cf. Section B). This section is followed by additional results and simulation specifications for the GAN-based generation of tissue geometries (cf. Section C). A final section (cf. Section D) provides details on the literature review on tissue modeling that was summarized in Section 3.2.

A. Photoacoustic Data

This section contains a link to the annotation protocol that was used for the manual annotations.

Disclosure to this work:
The annotations performed followed the annotation protocol that is part of the *Photoacoustics* journal publication by Schellenberg et al., 2022b.

The annotation protocol is accessible under: `https://ars.els-cdn.com/content/image/1-s2.0-S2213597922000118-mmc1.pdf`

B. Tissue Geometry Estimation with Neural Networks

The supplemental results of the baseline and robustness experiments (cf. Section 4.2.5) can be found in this section. For both experiments, qualitative and quantitative results are presented.

Disclosure to this work:
The following supplemental material was published in the journal *Photoacoustics* by Schellenberg et al., 2022b and the content, Figures B.1-B.4, and Tables B.1-B.5 were taken (partly modified) from this publication with permission.

Baseline experiment

Table B.1.: Dice Similarity Coefficients (DSCs) and Normalized Surface Distances (NSDs) achieved with the baseline experiment performed for the nnU-Net and Fully-Connected Neural Network (FCNN) trained on Photoacoustic (PA) images, PA and Ultrasound (US) (PAUS) images, and US images of the forearm, calf, and neck. For each test image, the metric values were calculated separately for each tissue class, and the corresponding results were averaged over all structures. The means of the metric values calculated across all test cases are shown here.

	Tissue Class	nnU-Net PA	nnU-Net PAUS	nnU-Net US	FCNN PA	FCNN PAUS
DSC	Average	0.83	0.85	0.80	0.62	0.66
	Blood	0.71	0.74	0.32	0.48	0.53
	Skin	0.89	0.89	0.87	0.77	0.79
	Tissue	0.98	0.98	0.98	0.88	0.89
	Membrane	0.91	0.91	0.91	0.77	0.83
	US gel	0.86	0.86	0.86	0.75	0.80
	Heavy water	0.99	0.99	0.99	0.69	0.76
	Artefact	0.44	0.52	0.56	0.03	0.05
	Fat	0.86	0.88	0.87	0.59	0.63
NSD	Average	0.88	0.89	0.84	0.59	0.61
	Blood	0.84	0.85	0.47	0.75	0.75
	Skin	0.98	0.98	0.97	0.87	0.89
	Tissue	0.83	0.85	0.83	0.24	0.27
	Membrane	1.00	1.00	1.00	0.89	0.92
	US gel	0.98	0.98	0.98	0.91	0.93
	Heavy water	1.00	1.00	1.00	0.27	0.31
	Artefact	0.55	0.52	0.55	0.07	0.07
	Fat	0.90	0.93	0.93	0.70	0.72

Table B.1 shows the DSC and NSD of all tissue structures averaged across the test images. Overall, the DSCs and NSDs are higher for the nnU-Net compared to the FCNN. Especially for the class blood, metric scores are higher for networks that were trained on data including PA images. Figure B.1 qualitatively shows the best, median, and worst estimations of semantic segmentation performed with the nnU-Net trained on PAUS images. In the worst case, the nnU-Net estimated superficial vessels and a larger vessel relatively deep in the tissue that were not shown in the reference annotation.

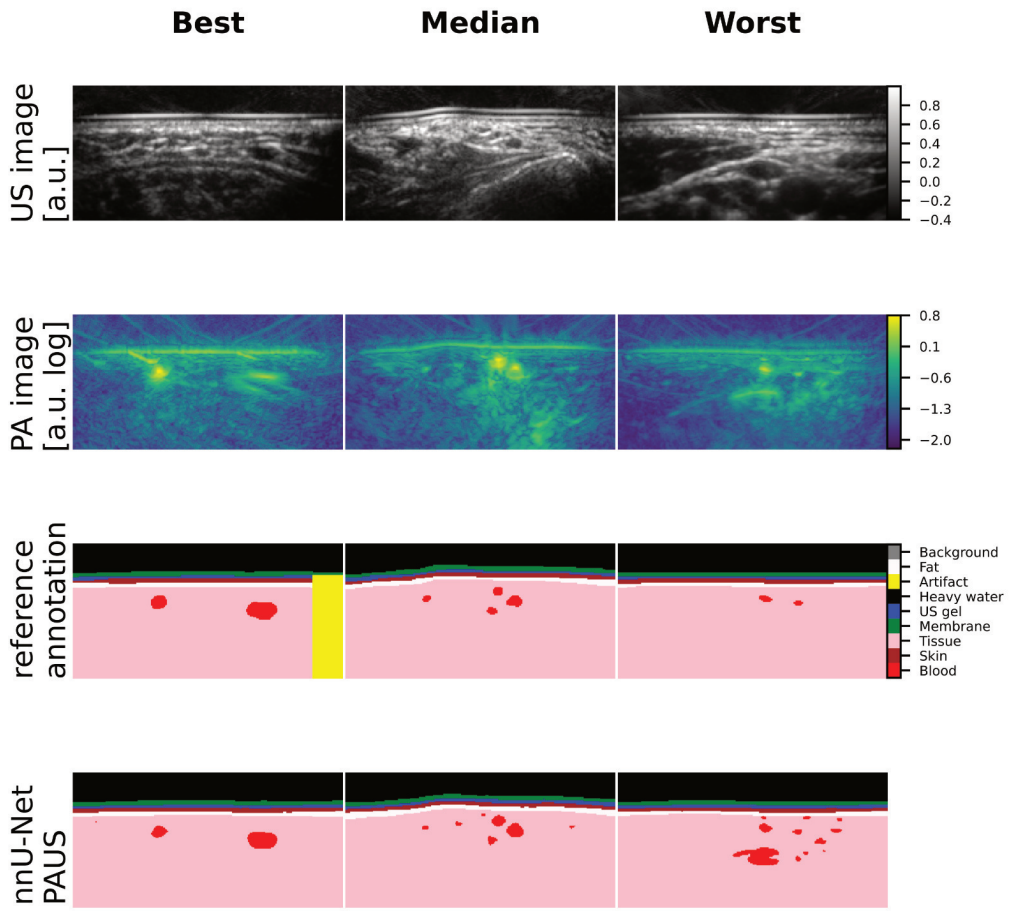

Figure B.1.: The semantic segmentation estimations of the nnU-Net trained on a combination of Photoacoustic (PA) and Ultrasound (US) (PAUS) data closely resemble the reference annotations in the (*left*) best and (*center*) median case. In the (*right*) worst case, the network over-segmented small superficial vessels and a larger deeper vessel. The US image, the PA image, the reference, and the estimated semantic segmentations are shown from top to bottom and were chosen according to the blood Dice Similarity Coefficient (DSC). Only images with at least 60 pixels of blood in the reference images were considered for choosing the median image.

Table B.2.: Linear mixed model estimates of the human annotator reliability study with the Dice Similarity Coefficient (DSC) and (log-scaled) Surface Distance. The body region was considered a fixed effect, and the annotator and image id (six scans for each body site) were considered random intercepts. The estimates, the standard error (SE), and the t-value according to Satterthwaites's method[1] for the fixed effects and the standard deviation σ of the random effects are shown.

		Fixed effect			**Random effect**
		Estimate	SE	t-value	σ
DSC	Intercept	0.66	0.05	14.19	
	Forearm vs. Calf	0.07	0.07	1.03	
	Neck vs. Calf	0.07	0.07	0.10	
	Image id				0.08
	Annotator				0.02
	Residual				0.09
(log-scaled) Surface Distance	Intercept	1.60	0.25	6.33	
	Forearm vs. Calf	-0.64	0.35	-1.83	
	Neck vs. Calf	-0.58	0.35	-1.65	
	Image id				0.36
	Annotator				0.23
	Residual				0.65

Table B.2 presents the results of the linear mixed model analysis that was used for aggregating the inter-rater reliability results (cf. Figure 4.2.5) and for setting the threshold for the *blood* NSD.

Robustness experiment

For each of the three robustness experiments, supplemental qualitative and quantitative results are presented. More specifically, additional results of the model trained on calf and neck data and tested on forearm data (cf. Figure B.2 and Table B.3), the model trained on forearm and neck data and tested on calf data (cf. Figure B.3 and Table B.4), and the model trained on forearm and calf data and tested on neck data (cf. Figure B.4 and Table B.5) are shown.

[1] `https://rdrr.io/cran/lmerTest/man/summary.lmerModLmerTest.html`

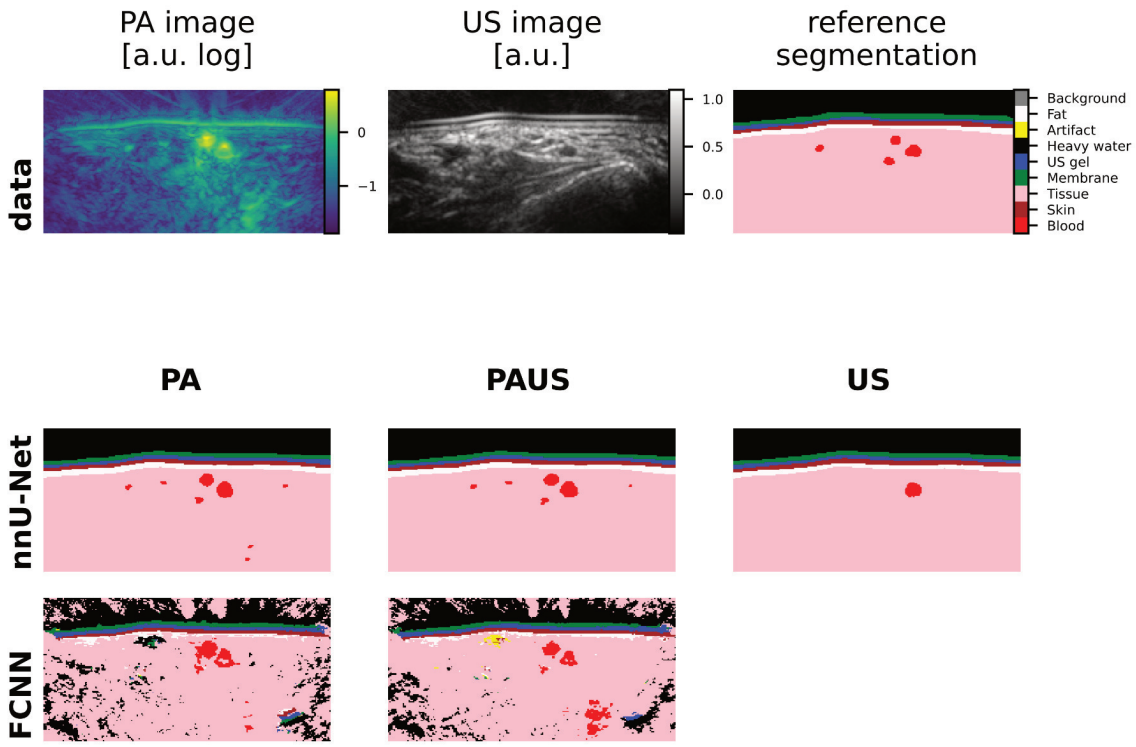

Figure B.2.: The semantic segmentation estimations of the nnU-Net and the Fully-Connected Neural Network (FCNN) **trained on calf and neck images** and **tested on forearm images** are in agreement with the reference annotations. The first row shows (*left*) the log-scaled PA image at 800 nm, (*center*) the US image, and (*right*) the reference segmentation of the representative example. The estimations of the nnU-Net and the FCNN trained and tested on (*left*) Photoacoustic (PA) images, (*center*) PA and Ultrasound (US) (PAUS) images, or (*right*) US images alone are shown below. Note that the FCNN was not trained on US images because of their one-dimensional nature. The representative image chosen was the one that achieved the median blood Dice Similarity Coefficient (DSC) with the nnU-Net trained on PAUS images and that contained at least 60 blood pixels.

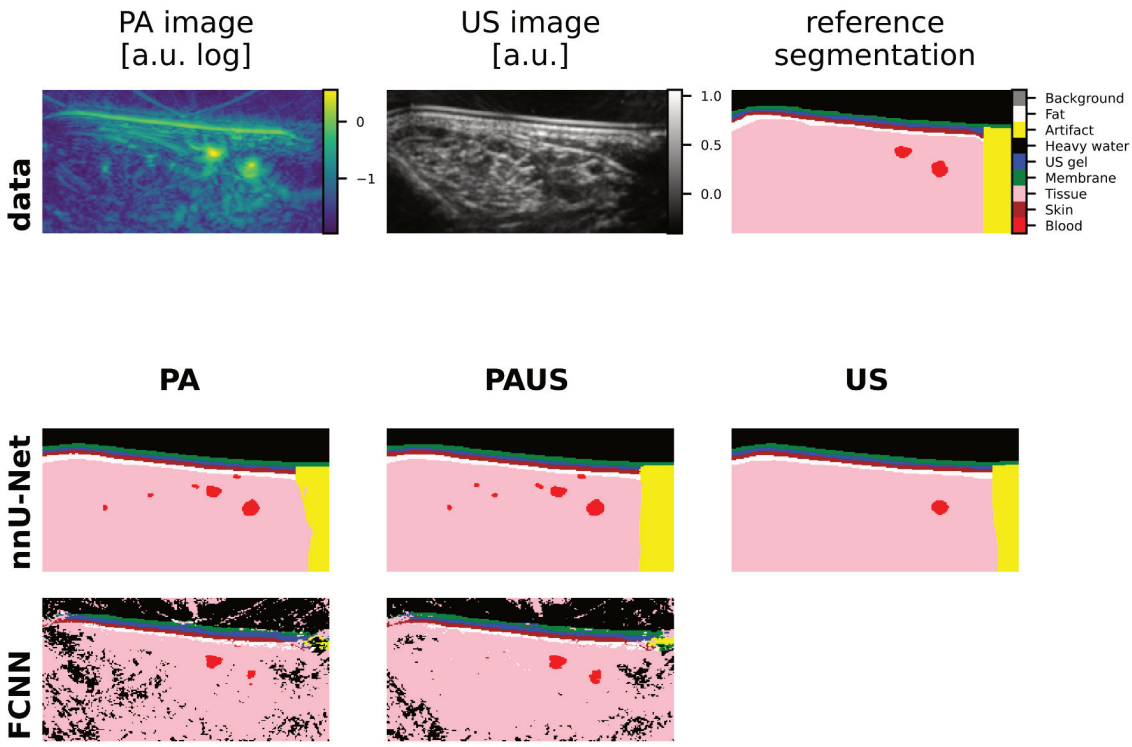

Figure B.3.: The semantic segmentation estimations of the nnU-Net and the Fully-Connected Neural Network (FCNN) **trained on forearm and neck images** and **tested on calf images** are in agreement with the reference annotations. The first row shows (*left*) the log-scaled PA image at 800 nm, (*center*) the US image, and (*right*) the reference segmentation of the representative example. The estimations of the nnU-Net and the FCNN trained and tested on (*left*) Photoacoustic (PA) images, (*center*) PA and Ultrasound (US) (PAUS) images, or (*right*) US images alone are shown below. Note that the FCNN was not trained on US images because of their one-dimensional nature. The representative image chosen was the one that achieved the median blood Dice Similarity Coefficient (DSC) with the nnU-Net trained on PAUS images and that contained at least 60 blood pixels.

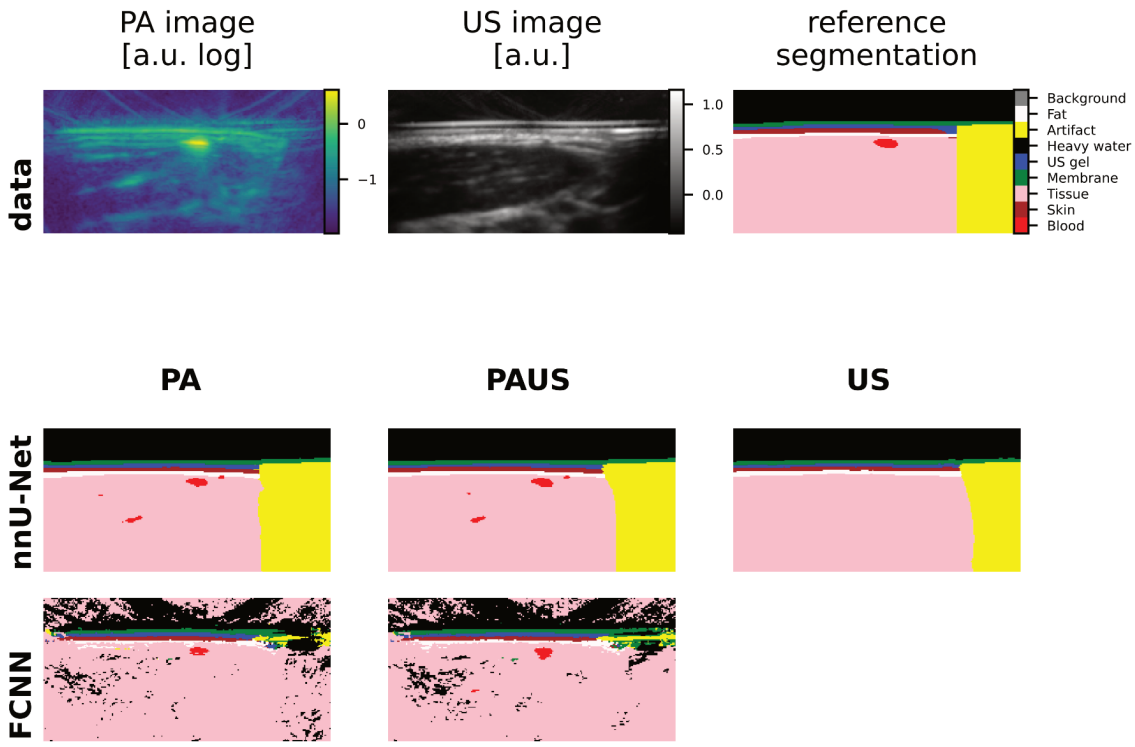

Figure B.4.: The semantic segmentation estimations of the nnU-Net and the Fully-Connected Neural Network (FCNN) **trained on forearm and calf images** and **tested on neck images** are in agreement with the reference annotations. The first row shows (*left*) the log-scaled PA image at 800 nm, (*center*) the US image, and (*right*) the reference segmentation of the representative example. The estimations of the nnU-Net and the FCNN trained and tested on (*left*) Photoacoustic (PA) images, (*center*) PA and Ultrasound (US) (PAUS) images, or (*right*) US images alone are shown below. Note that the FCNN was not trained on US images because of their one-dimensional nature. The representative image chosen was the one that achieved the median blood Dice Similarity Coefficient (DSC) with the nnU-Net trained on PAUS images and that contained at least 60 blood pixels.

Table B.3.: The Dice Similarity Coefficients (DSCs) and Normalized Surface Distances (NSDs) achieved with the nnU-Net and Fully-Connected Neural Network (FCNN) **trained on calf and neck images** and **tested on forearm images**. The models were based on either Photoacoustic (PA) images, PA and Ultrasound (US) (PAUS) images, or US images. For each test image, the metrics were calculated separately for each tissue class, and the corresponding results were averaged over all structures (class average). The metric values aggregated across all test cases are shown here for the class average and the tissue classes blood and skin.

	Tissue Class	nnU-Net PA	nnU-Net PAUS	nnU-Net US	FCNN PA	FCNN PAUS
DSC	Average	0.78	0.83	0.74	0.60	0.64
	Blood	0.66	0.72	0.21	0.46	0.49
	Skin	0.88	0.89	0.86	0.75	0.78
NSD	Average	0.83	0.88	0.81	0.58	0.60
	Blood	0.83	0.85	0.42	0.79	0.75
	Skin	0.97	0.98	0.96	0.86	0.89

Table B.4.: The Dice Similarity Coefficients (DSCs) and Normalized Surface Distances (NSDs) achieved with the nnU-Net and Fully-Connected Neural Network (FCNN) **trained on forearm and neck images** and **tested on calf images**. The models were based on either Photoacoustic (PA) images, PA and Ultrasound (US) (PAUS) images, or US images. For each test image, the metrics were calculated separately for each tissue class, and the corresponding results were averaged over all structures (class average). The metric values aggregated across all test cases are shown here for the class average and the tissue classes blood and skin.

	Tissue Class	nnU-Net PA	nnU-Net PAUS	nnU-Net US	FCNN PA	FCNN PAUS
DSC	Average	0.84	0.86	0.79	0.62	0.66
	Blood	0.69	0.74	0.34	0.46	0.54
	Skin	0.90	0.90	0.86	0.77	0.80
NSD	Average	0.88	0.89	0.84	0.58	0.61
	Blood	0.82	0.83	0.48	0.73	0.78
	Skin	0.98	0.98	0.96	0.85	0.86

Table B.5.: The Dice Similarity Coefficients (DSCs) and Normalized Surface Distances (NSDs) achieved with the nnU-Net and Fully-Connected Neural Network (FCNN) **trained on forearm and calf images** and **tested on neck images**. The models were based on either Photoacoustic (PA) images, PA and Ultrasound (US) (PAUS) images, or US images. For each test image, the metrics were calculated separately for each tissue class, and the corresponding results were averaged over all structures (class average). The metric values aggregated across all test cases are shown here for the class average and the tissue classes blood and skin.

	Tissue Class	nnU-Net PA	nnU-Net PAUS	nnU-Net US	FCNN PA	FCNN PAUS
DSC	Average	0.81	0.82	0.77	0.62	0.65
	Blood	0.68	0.70	0.12	0.53	0.52
	Skin	0.86	0.86	0.83	0.73	0.75
NSD	Average	0.88	0.87	0.82	0.59	0.59
	Blood	0.83	0.84	0.34	0.75	0.68
	Skin	0.96	0.97	0.95	0.86	0.87

C. Tissue Geometry Generation with Generative Adversarial Networks

For the three experiments performed with the GAN-based approach to tissue geometry generation (cf. Section 4.3.5), qualitative and quantitative results are shown in this section.

Disclosure to this work:
The following supplemental material was published in the journal *Photoacoustics* by Schellenberg et al., 2022a and the content, Figures C.3-C.7, and Table C.1 were taken (partly modified) from this publication with permission.

Forearm experiment

The performance of the GAN-based augmentation strategy to generate tissue geometries is further validated in Figure C.1. The model trained on a combination of GAN- and annotation-based data achieves the best rank for most of the target structures and clearly outperforms the

model trained solely on literature- or annotation-based data. The distributions of the absolute errors of the competing quantification downstream task models tested on the annotation-based target test data set are shown in Figure C.2. The distributions of the model trained on GAN-based tissue geometries are narrowest and closest to zero for most of the tissue classes. In addition, examples of the tissue geometries generated during training of the GAN based on forearm data are shown in Figure C.3.

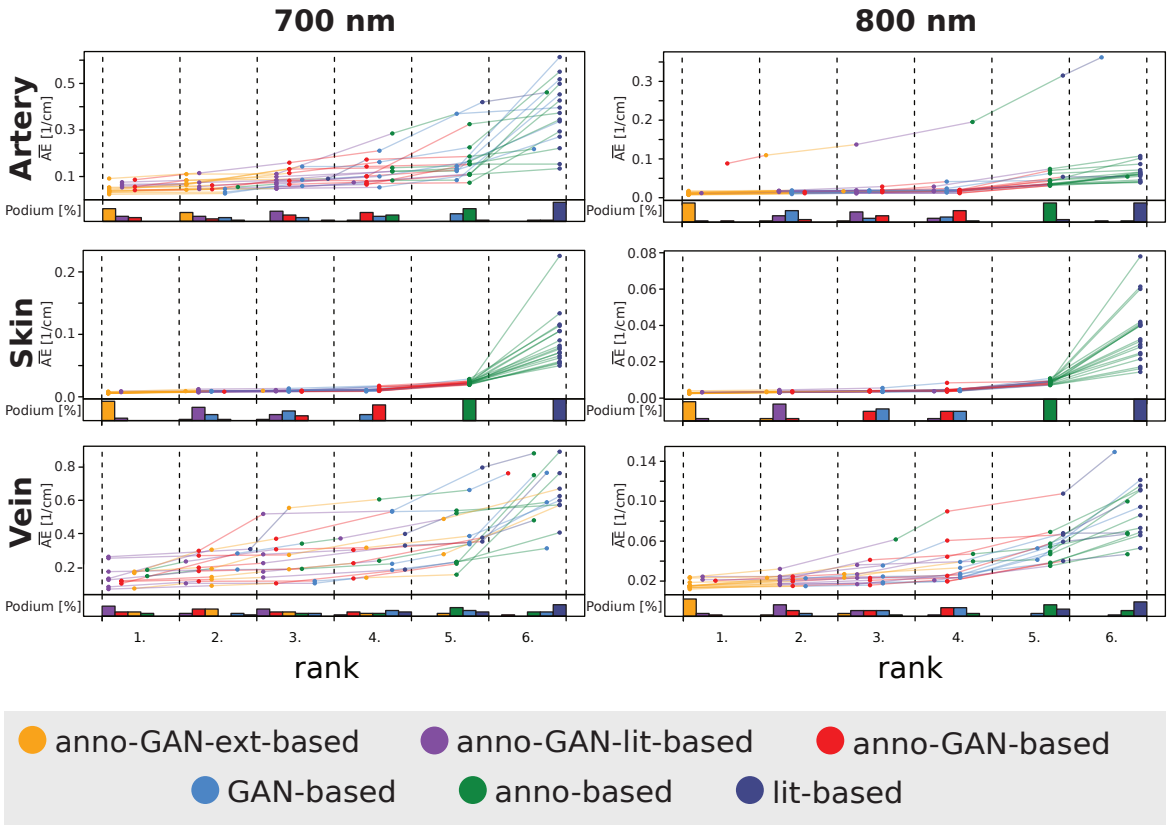

Figure C.1.: Comparative analysis and ranking of the six forearm quantification downstream task models using the absolute errors (AE) per target structures. The models were trained on simulated Photoacoustic (PA) images based on different combinations of annotation (anno)-, Generative Adversarial Network (GAN)-, and literature-based forearm tissue geometries (cf. Table 4.3.1) and tested on the identical annotation-based test data set. The rankings were computed for the mean absolute error (AE) (more specifically, $AE_{x,c=1,\lambda}$, $AE_{x,c=2,\lambda}$, and $AE_{x,c=7,\lambda}$) within the tissue classes (*top*) artery, (*center*) skin, and (*bottom*) vein at wavelengths of (*left*) 700 nm and (*right*) 800 nm using the challengeR concept [Wiesenfarth et al., 2021]. Top parts: For every test case, the AE per model and per test image is plotted color-coded and ordered by the achieved ranks (best 1, worst 6). Lower parts: The bar charts represent the relative frequency at which each model achieved the rank encoded by the podium place.

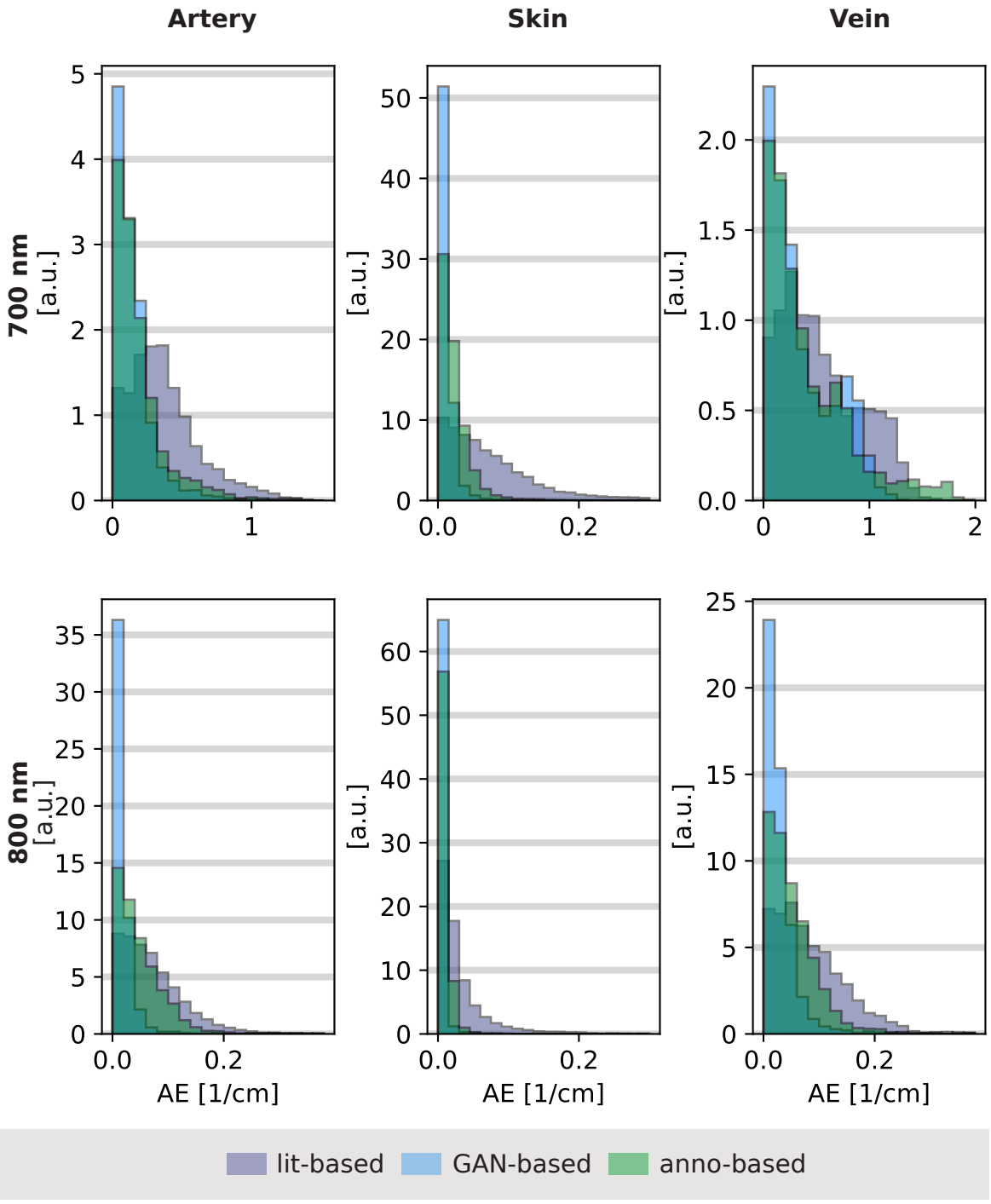

Figure C.2.: Comparative validation of three forearm quantification downstream task models using the distributions of the absolute errors (AE). The models were trained with simulated Photoacoustic (PA) images that were based on literature (lit)-, Generative Adversarial Network (GAN)-, and annotation (anno)-based forearm tissue geometries, respectively and tested on the identical annotation-based test data set. The distributions of the estimated AE at (*top*) 700 nm and (*bottom*) 800 nm for the tissue classes: (*left*) artery, (*center*) skin, and (*right*) vein are shown. The distributions of the GAN-based model are narrowest and closest to zero for most of the tissue classes.

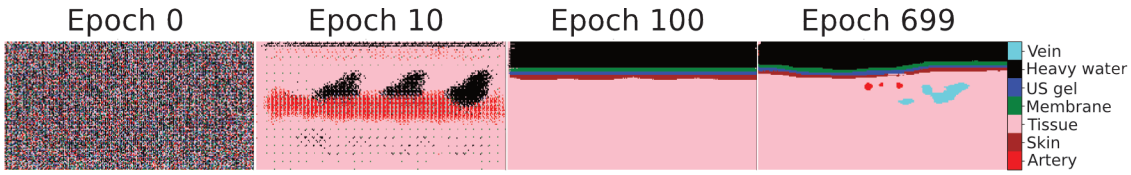

Figure C.3.: Representative examples of tissue geometries generated during training of the Generative Adversarial Network (GAN) based on reference forearm tissue geometries.

Calf experiment

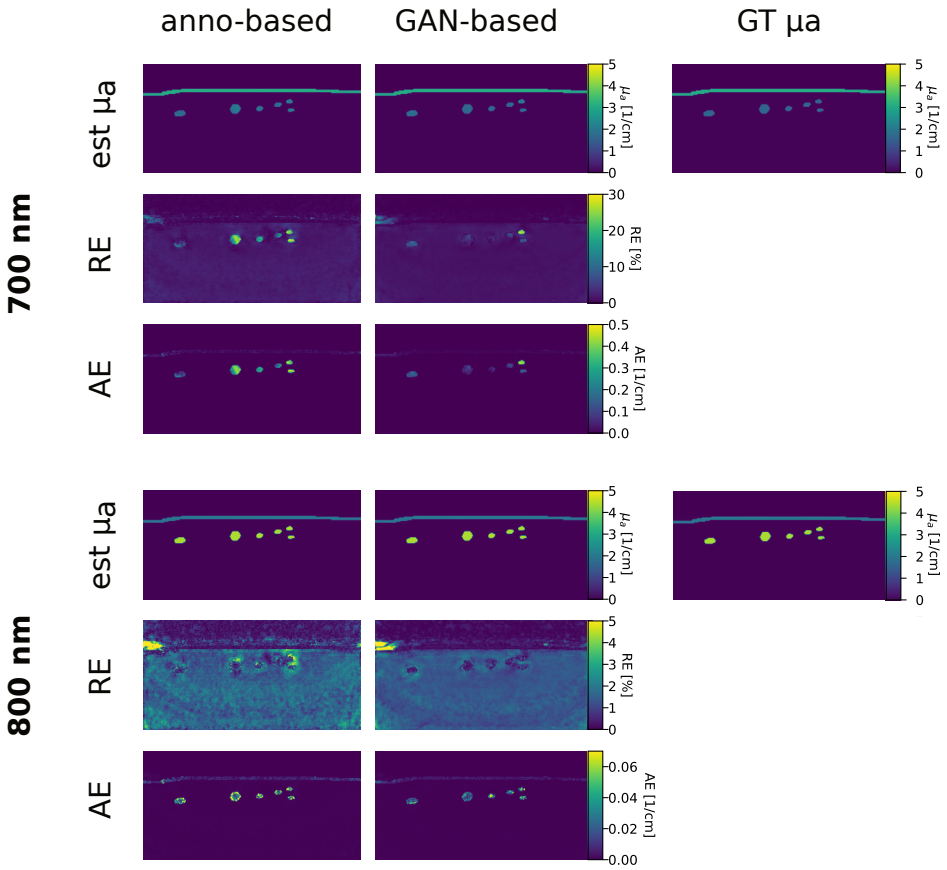

Figure C.4.: Qualitative quantification results on a representative annotation-based calf test case for the models trained on (*left*) annotation-based (anno) and (*right*) Generative Adversarial Network (GAN)-based data. The estimated absorption coefficient (est μ_a), the relative error (RE), the absolute error (AE), and the corresponding ground truth (Ground Truth (GT) μ_a) at (*top*) 700 nm and (*bottom*) 800 nm reveal that the GAN-based models more closely resemble the μ_a GTs than the annotation-based model. The example image was chosen according to the median of the per-image mean absolute errors at 700 nm ($AE_{x,c=0,\lambda=700nm}$) for the model trained on the annotation-based data set.

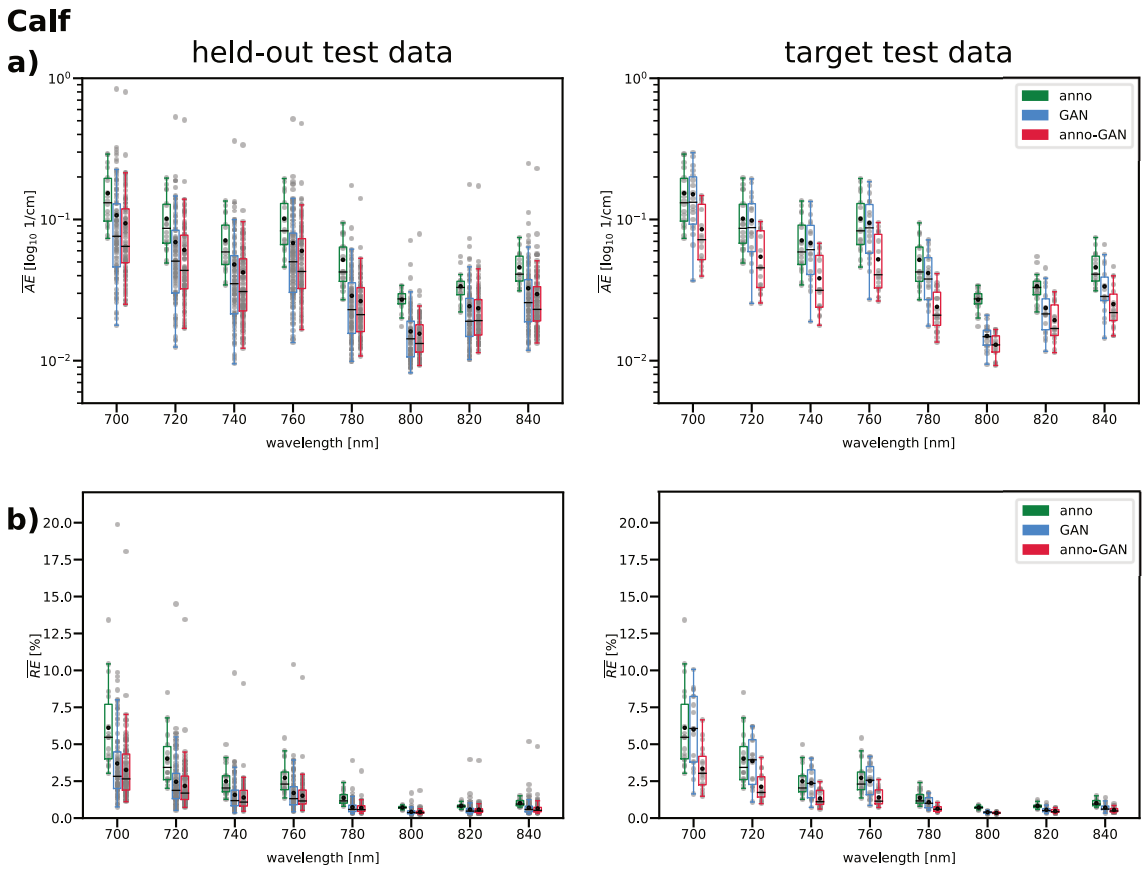

Figure C.5.: Quantitative results of the calf experiment with three quantification models trained on the different data configurations as shown in Table 4.3.1. The (**a**) absolute and (**b**) relative errors of the models (*left*) tested on the in-distribution held-out test set are in the same order of magnitude as when the models (*right*) were applied on target annotation-based test data. The per-image and per-wavelength absolute and relative errors ($AE_{x,c=1,2,7,\lambda}$ and $RE_{x,c=1,2,7,\lambda}$) aggregated over the target classes artery, skin, and vein (gray dots) are shown. The median, the interquartile range, and the mean values per respective wavelength are indicated as a black bar, colored box, and black dot, respectively.

Qualitative and quantitative results of the competing quantification downstream tasks for calf data are given in Figure C.4 and Figure C.5, respectively. As shown in Figure C.4, the models trained on annotation- or GAN-based data and tested on the identical annotation-based test data set closely resemble the ground truth coefficients. However, the estimation errors are largest in target structures. The example image was chosen according to the median of the mean absolute errors averaged over the whole images at 700 nm ($AE_{x,c=0,\lambda=700\text{nm}}$) estimated with the annotation-based model. The quantitative plots analyzing the absolute and relative errors show that the three competing downstream task models perform similarly on held-out and target test data (cf. Figure C.5).

Neck experiment

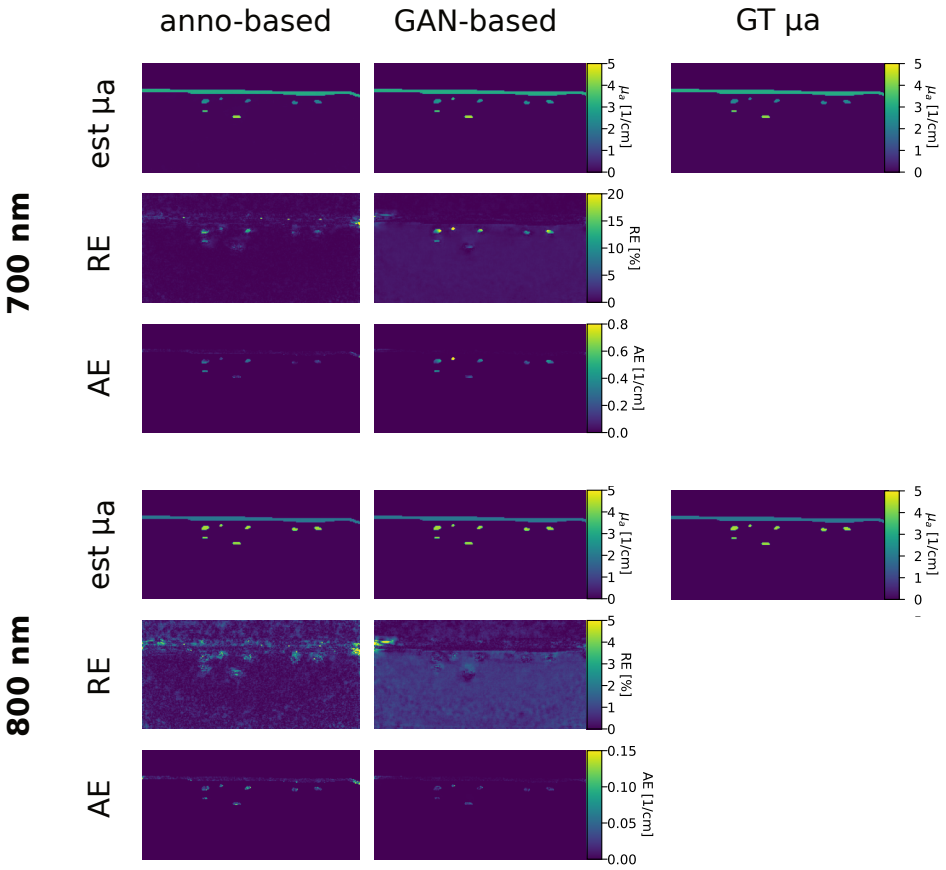

Figure C.6.: Qualitative quantification results on a representative annotation-based neck test case for the models trained on (*left*) annotation-based (anno) and (*right*) Generative Adversarial Network (GAN)-based data. The estimated absorption coefficient (est μ_a), the relative error (RE), the absolute error (AE), and the corresponding ground truth (Ground Truth (GT) μ_a) at (*top*) 700 nm and (*bottom*) 800 nm reveal that the GAN-based models resemble the μ_a GTs comparably well to the annotation-based model. The example image was chosen according to the median of the per-image mean absolute errors at 700 nm ($\text{AE}_{x,c=0,\lambda=700\text{nm}}$) for the model trained on the annotation-based data set.

Qualitative and quantitative results of the quantification downstream task for the neck models are given in Figure C.6 and Figure C.7, respectively. Overall, the results of the calf experiment hold true. The models trained on annotation- or GAN-based data closely resemble the GT, but the error is largest in target structures. Analogous to the calf experiment, the example image was chosen according to the median of the mean absolute errors averaged over the whole images at 700 nm ($\text{AE}_{x,c=0,\lambda=700\text{nm}}$) for the annotation-based model. Overall, similar to the calf results,

the performance on held-out test data and target annotation-based test data is comparable for the neck downstream task models, and the absolute and relative errors are in the same order of magnitude (cf. Figure C.7).

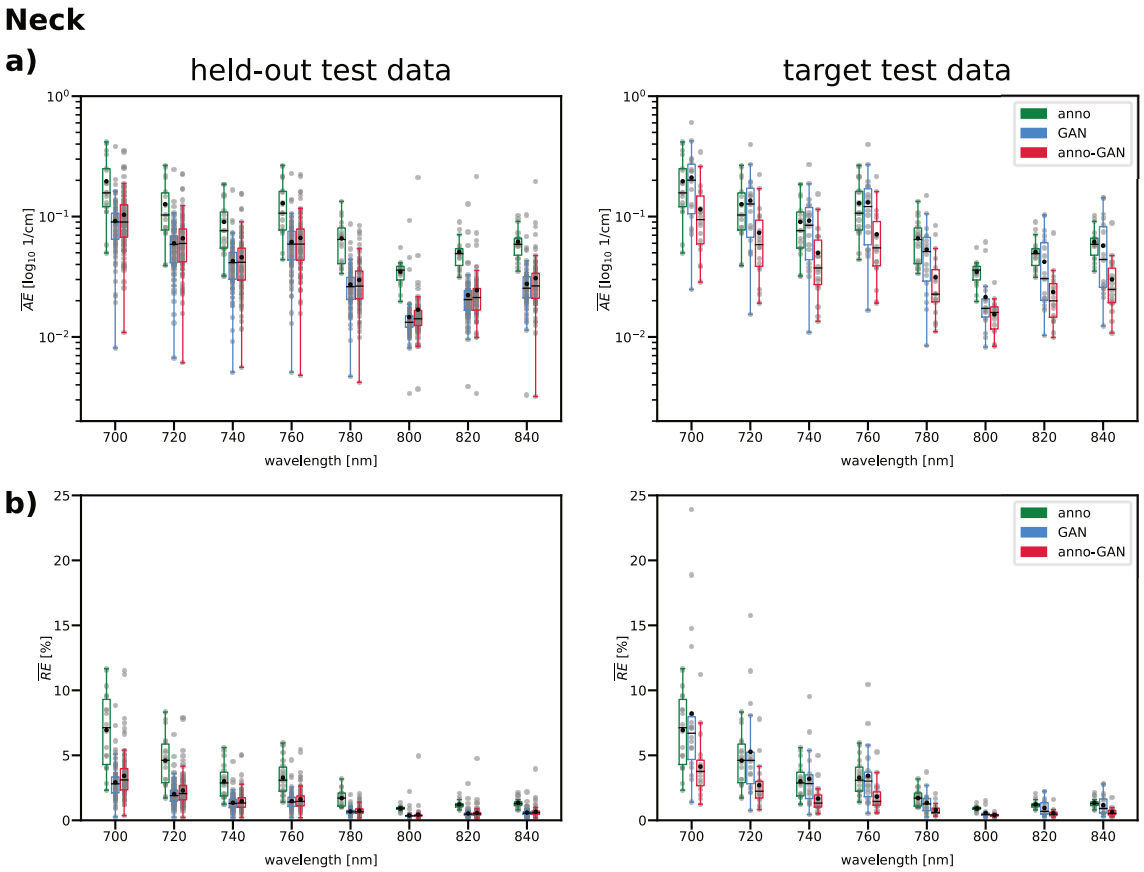

Figure C.7.: Quantitative results of the neck experiment with three quantification models trained on the different data configurations as shown in Table 4.3.1. The (**a**) absolute and (**b**) relative errors of the models (*left*) tested on the in-distribution held-out test set are in the same order of magnitude as when the models (*right*) were applied on target annotation-based test data. The per-image and per-wavelength absolute and relative errors ($\mathrm{AE}_{x,c=1,2,7,\lambda}$ and $\mathrm{RE}_{x,c=1,2,7,\lambda}$) aggregated over the target classes artery, skin, and vein (gray dots) are shown. The median, the interquartile range, and the mean values per respective wavelength are indicated as a black bar, colored box, and black dot, respectively.

Literature-based forearm model

For the simulation of the literature-based human forearm model (cf. Figure 4.3.2), volumes of the same size compared to the annotation-based simulations (cf. Section 4.3.2) were used.

In other words, the 3D volume was of size 75.0 mm x 20.0 mm x 68.2 mm $+ \ t_{\text{USgel}}$ along the x-, y-, and z-axis, where t_{USgel} denotes the thickness of the US gel layer. The 3D volume was assembled in three steps:

1. One 2D cross-section of a right human forearm (x-z-axis) was generated, including the seven tissue classes heavy water, membrane, US gel, skin, background tissue, artery (ulnar, interosseous, radial, and random), and vein (accompanying, random). A schematic is shown in Figure 4.3.2. The size from the top of the skin layer to the bottom of the forearm model was 75 mm x 25 mm. The 2D geometrical properties of these tissue structures are explained in detail in the following section. The optical properties were assigned per tissue class according to SIMPA's internal tissue library (cf. Section 4.3.2).

2. The 2D cross-section was placed in the center of the y-axis, such that the highest point of the US gel layer was positioned at the bottom end of the probe origin, at z = 43.2 mm. Copies of the 2D plane were stacked along the y-axis.

3. A deformation algorithm included in SIMPA was applied to the 3D volume. It consists of six steps. First, the surface of the skin was raised at four to six (sampled from $\mathcal{U}\ (4, 6)$) equally distributed points along the x- and y-axis by a respective elevation amplitude of $z_{\text{Gauss}} \sim \mathcal{U}(0.0,\ 1.0)$ mm. Second, the elevation amplitudes of the surface were normalized by dividing each amplitude by the maximum elevation amplitude. Third, for every four to six positions along the x-axis x_i and every four to six positions along the y-axis y_j, a scaling factor z_{cos} was calculated by:

$$z_{\text{cos}} = \cos\left(x_i\pi - \frac{\pi}{2}\right)^2 \cdot \cos\left(y_j\pi - \frac{\pi}{2}\right)^2. \tag{7.1}$$

   Fourth, the resulting scaling factors for each of the positions were rescaled such that the maximum factor was equal to 2. Subsequently, the scaling factors were multiplied by the elevation amplitudes, respectively. Fifth, the maximum elevation along the z-axis was subtracted from all elevation amplitudes to ensure that the surface maximum is at a z-position of 43.2 mm $+ \ t_{\text{USgel}}$. Lastly, the simulation surface was deformed by applying a 2D cubic interpolation to the elevation amplitudes, and the volume underneath the surface was deformed accordingly.

In the following, the 2D geometrical properties of the seven tissue classes are presented. Note that the origin for the following description is $o(x, z)$ = $(0.0, 43.2 \text{ mm} + t_{\text{USgel}})$. Specifically for vessels, an assembling strategy is presented.

Heavy water The area between the top of the 2D cross-section and the membrane layer was assigned to heavy water.

Memebrane The membrane was modeled on top of the US gel layer with a constant thickness of 1 mm.

Ultrasound gel The US gel layer was modeled on top of the skin layer, and its thickness t_{USgel} was sampled from $\mathcal{N}(0.4,\ 0.1)$ mm.

Skin The skin layer was located such that the top part of the layer was at $z = 0$ mm. Based on findings in the work by Oltulu et al., 2018, the skin layer's thickness was sampled from a positive normal distribution $\mathcal{N}_+(0.2\text{ mm},\ 0.1\text{ mm})$. $\mathcal{N}_+$ was defined such that only non-negative values were accepted during sampling.

Tissue The tissue filled the space between the skin and the bottom of the simulation plane. This tissue class was assigned the lowest priority, meaning other tissue classes, such as arteries and veins, could replace corresponding pixels.

Artery The arteries were modeled as ellipses of radius r and eccentricity ε elongated along the x-axis. This stretching was introduced to account for any compression of the vessels along this axis due to pressure on the tissue by the probe and the operator. The eccentricity was assumed to follow a uniform distribution $\varepsilon \sim \mathcal{U}(0.0,\ 0.8)$. Following the works by Ashraf et al., 2010 and Hubmer et al., 2004 that studied the radii of arteries in the human forearm, the radii of the radial, ulnar, and interosseous arteries were sampled from positive normal distributions $\mathcal{N}_+(1.1\text{ mm},\ 0.2\text{ mm})$, $\mathcal{N}_+(\text{mm})$, and $\mathcal{N}_+(0.3\text{ mm},\ 0.1\text{ mm})$, respectively.
The x positions of the radial and ulnar arteries were defined with a relative shift s_{rel} and a random shift s_{random} as $x \sim 37.5\text{ mm} + s_{\text{rel}} \pm s_{\text{random}}$, respectively. The relative shift was introduced to make the simulation more realistic, with target structures not necessarily located centrally in the PA image. It was sampled from $\mathcal{U}(-15.0\text{ mm},\ 15.0\text{ mm})$ to vary the x positions of the vessels, which is equivalent to moving the PA probe along the x-axis in reverse order. The random shift was implemented to vary the x-positions of the ulnar and radial arteries but to keep an average distance of the two arteries following $\mathcal{N}_+(15.0\text{ mm},\ 2.5\text{ mm})$. The z positions of the ulnar and radial arteries were sampled from $\mathcal{N}_+(4.0\text{ mm},\ 0.1\text{ mm})$ where μ and σ were empirically

determined by an in-house analysis of PA images from healthy volunteers.
The x position of the interosseous artery was chosen to be central to the ulnar and radial arteries on average. It was assumed to follow $37.5\,\text{mm} + s_{rel} \pm \mathcal{N}(0.0\,\text{mm},\ 2.5\,\text{mm})$. The height z of the interosseous artery was assumed as $z \sim \mathcal{N}_+(\text{mm})$.

Vein Veins were modeled to accompany arteries following the publication by Standring, 2021. While the radii of veins accompanying the ulnar or radial artery were assumed to be similar, they were sampled from $\mathcal{N}_+(0.5\,\text{mm},\ 0.1\,\text{mm})$ based on the work by Yang et al., 2018. The radius of veins accompanying the interosseous artery was assumed to be half as large and sampled from $\mathcal{N}(0.3\,\text{mm},\ 0.1\,\text{mm})$. The eccentricity was assumed to follow a uniform distribution $\varepsilon \sim \mathcal{U}(0.3,\ 0.9)$. The x and z positions were determined relative to the parent artery and sampled from $\pm\,\mathcal{N}(2.5\,\text{mm},\ 0.4\,\text{mm})$ and $\mathcal{N}(0.0\,\text{mm},\ 0.8\,\text{mm})$, respectively.

Table C.1.: Probabilities (P) of the number (#) of arteries and veins, respectively. The probabilities were extracted from the measured PAI data set (cf. Section 4.1), consisting of 96 forearm images.

# vessels	P(arteries)	P(veins)
0	3/96	18/96
1	8/96	22/96
2	21/96	20/96
3	26/96	19/96
4	14/96	5/96
5	10/96	6/96
6	4/96	2/96
7	3/96	2/96
8	4/96	2/96
9	2/96	-
10	1/96	-

Assembling of vessels As investigated in an empirical analysis of the acquired human forearm data set, individual volunteers have different numbers of vessels. To account for this diversity, random vessels were added. The radius of random vessels was sampled from $\mathcal{U}(0.3\,\text{mm},\ 0.6\,\text{mm})$ equivalent to the range between 25 % and 50 % of the size of the radial artery. The x position was sampled from $\mathcal{U}(18.8\,\text{mm},\ 56.4\,\text{mm})$. If the random vessel was an artery, the eccentricity

was, as for arteries, sampled from a uniform distribution $\varepsilon \sim \mathcal{U}(0.0,\ 0.8)$ and the vessel was located according to $z \sim \mathcal{U}(0.0\ \text{mm},\ 12.5\ \text{mm})$. If the random vessel was a vein, the eccentricity was, as for veins, sampled from a uniform distribution $\varepsilon \sim \mathcal{U}(0.3,\ 0.9)$ and the vessels' height was located with equal probability superficially or as arterial random vessels following $\mathcal{U}(0.0\ \text{mm},\ 12.5\ \text{mm})$. The height of superficial random veins was determined following $\mathcal{U}(t_{\text{Skin}} + 2 \cdot r_{\text{RandomVessel}},\ t_{\text{Skin}} + 4 \cdot r_{\text{RandomVessel}})$ with thickness t and radius r.
Aiming to achieve a realistic distribution of vessels, acquired PAI measurements were analyzed. The probabilities of the number of arteries and veins in the data set were analyzed (cf. Table C.1), and a Poisson distribution was fitted to each distribution, respectively. To assemble one forearm simulation, the following procedure was performed until all vessels were modeled:

1. The number of vessels $\#V_{\text{arteries}}$ and $\#V_{\text{veins}}$ of one forearm model was sampled from the two distributions, respectively.

2. The relative shift (s_{rel}) was determined.

3. The arteries were modeled. Depending on the number of arteries (cf. step 1.), one of the following scenarios was applied:

    a) $\#V_{\text{arteries}} == 1$: The first artery modeled was the ulnar or radial artery dependent on the positive or negative sign of the $shift_{rel}$.

    b) $\#V_{\text{arteries}} == 2$: The first artery was modeled as in (a). The second artery was a random vessel.

    c) $\#V_{\text{arteries}} == 3$: The three arteries modeled were the ulnar, radial, and interosseous arteries.

    d) $\#V_{\text{arteries}} \geq 3$: The first three arteries were modeled as in (c). The remaining ones were modeled as random ones.

4. Vein modeling.

    a) V_{radial}, V_{ulnar}, or $V_{\text{interosseous}}$ existed: For each of the available arteries zero, one (probability $50:50$ for left or right), or both accompanying veins (probability of $1/3$ each) were added.

    b) V_{radial}, V_{ulnar}, or $V_{\text{interosseous}}$ did not exist: A random vein was modeled.

Note that for steps 3. and 4. only new vessels were accepted if they did not overlap with already modeled vessels. If there was an overlap, the vessel was rejected, and new positioning (x- and z-position) and geometrical (radius, eccentricity) values were sampled.

D. Literature Review: Tissue Geometry Generation in Deep Learning-based Photoacoustic Imaging

An extensive analysis of the concepts for modeling tissue geometries in publications on Deep Learning (DL)-based Photoacoustic Imaging (PAI) between January 2017 and June 2023 was performed. The search string on Google Scholar was defined as ("Deep Learning" OR "Neural Network") AND ("Photoacoustic" OR "Optoacoustic") following Gröhl et al., 2021b. The initial over 300 papers were refined based on screening the abstracts and figures if available, resulting in 217 publications. Only papers in the field of DL-based PA imaging/angiography/microscopy were included in the analysis, i.e., pure spectroscopy papers were excluded. Scientific letters were also included unless they were below one page in length. The peer-reviewed version was preferred if there were multiple versions of a paper. Six publications could not be accessed, so either another version was used or the respective publication was excluded. Approximately 60 % of the papers based their work on virtual data, as shown in Table D. This table also shows the assignment of the papers into seven tissue modeling categories.

Table D.1.: Analysis of concepts for modeling tissue geometries in publications on Deep Learning (DL)-based Photoacoustic (PA) image analysis between January 2017 and June 2023. Compared to the publications highlighted in red, the ones highlighted in greed did use virtual data ($\sim$ 60 %). The works based on virtual data were categorized according to their concept of tissue geometry modeling. Seven classes were defined, basing tissue geometry modeling on random geometric shapes, pattern phantoms, model-based vasculatures or tissue, segmentation (seg.)-based vasculature or tissue, or DL.

#	reference	geometric shapes	pattern phantom	model-based vasculature	model-based tissue	seg.-based vasculature	seg.-based tissue	DL-based tissue
1	Aggrawal et al., 2022							
2	Aggrawal et al., 2023							
3	Agrawal et al., 2021b	x			x			
4	Agrawal et al., 2021c				x			
5	Agrawal et al., 2021a	x						
6	Allman et al., 2018	x						
7	Allman et al., 2019	x						
8	Anas et al., 2018b							
9	Anas et al., 2018a	x						
10	Antholzer et al., 2018b		x			x		
11	Antholzer et al., 2019b					x		
12	Antholzer et al., 2021	x						
13	Antholzer et al., 2018a	x						
14	Antholzer et al., 2019a	x	x					
15	Athira et al., 2022							
16	Awasthi et al., 2020		x	x		x		
17	Awasthi et al., 2019		x	x				
18	Bell, 2019	x						
19	Bench et al., 2023	x						x
20	Bench et al., 2020				x	x		

#	reference	geometric shapes	pattern phantom	model-based vasculature	model-based tissue	seg.-based vasculature	seg.-based tissue	DL-based tissue
21	Boink et al., 2019					x		
22	Cai et al., 2018	x						
23	Chen et al., 2020a						x	
24	Chen et al., 2020b	x						
25	Chen et al., 2019							
26	Cheng et al., 2022							
27	Chlis et al., 2020							
28	Choi et al., 2023							
29	Czuchnowski et al., 2021							
30	Davoudi et al., 2019	x					x	
31	Davoudi et al., 2021							
32	Dehner et al., 2022a							
33	Dehner et al., 2022b							
34	Deng et al., 2019	x						
35	Dhengre et al., 2020							
36	DiSpirito et al., 2020							
37	Dreher et al., 2023				x			
38	Durairaj et al., 2020	x						
39	Farnia et al., 2020					x		
40	Feng et al., 2022	x						
41	Feng et al., 2020	x	x	x				
42	Gao et al., 2022					x		
43	Gerl et al., 2020			x		x		
44	Godefroy et al., 2021					x		
45	Gong et al., 2021					x		

#	reference	geometric shapes	pattern phantom	model-based vasculature	model-based tissue	seg.-based vasculature	seg.-based tissue	DL-based tissue
46	González et al., 2023		x			x		
47	González et al., 2022		x			x		
48	Gopalan et al., 2023	x						
49	Grasso et al., 2022							
50	Gröhl et al., 2018	x						
51	Gröhl et al., 2021c	x			x			
52	Gu et al., 2023							
53	Guan et al., 2020			x		x		
54	Guan et al., 2021b	x		x		x		
55	Guan et al., 2021a		x	x			x	
56	Gubbi et al., 2021	x						
57	Gulenko et al., 2022							
58	Guo et al., 2022					x		
59	Gutta et al., 2017		x	x				
60	Hakakzadeh et al., 2022			x				
61	Hakimnejad et al., 2023							
62	Hariri et al., 2020							
63	Hauptmann et al., 2023					x		
64	Hauptmann et al., 2018					x		
65	He et al., 2022							
66	Hoffmann et al., 2022							
67	Hsu et al., 2023			x		x		
68	Hu et al., 2022							
69	Hwang et al., 2023		x					
70	Jeon et al., 2021	x						

#	reference	geometric shapes	pattern phantom	model-based vasculature	model-based tissue	seg.-based vasculature	seg.-based tissue	DL-based tissue
71	Jeon et al., 2020	x						
72	Jiang et al., 2023							
73	Jnawali et al., 2019a							
74	Jnawali et al., 2020							
75	Jnawali et al., 2019b							
76	Johnstonbaugh et al., 2020	x						
77	Johnstonbaugh et al., 2019	x						
78	Joseph et al., 2021					x		
79	Kenhagho et al., 2021					x		
80	Kikkawa et al., 2021							
81	Kim et al., 2022a							
82	Kim et al., 2022b							
83	Kim et al., 2020					x		
84	Kirchner et al., 2021	x						
85	Lafci et al., 2020a							
86	Lafci et al., 2020b							
87	Lan et al., 2020					x		
88	Lan et al., 2021c			x				
89	Lan et al., 2021b	x						
90	Lan et al., 2023b		x			x		
91	Lan et al., 2019a	x				x		
92	Lan et al., 2023a							
93	Lan et al., 2021a					x		
94	Lan et al., 2019b	x				x		
95	Lan et al., 2019c					x		

#	reference	geometric shapes	pattern phantom	model-based vasculature	model-based tissue	seg.-based vasculature	seg.-based tissue	DL-based tissue
96	Le et al., 2021							
97	Le et al., 2022b							
98	Leng et al., 2021a							
99	Leng et al., 2021b							
100	Li et al., 2020b	x	x	x				
101	Li et al., 2022a	x					x	
102	Li et al., 2021d			x				
103	Li et al., 2021c			x				
104	Li et al., 2021b			x				
105	Li et al., 2021e	x				x		
106	Li et al., 2021f							
107	Li et al., 2021g					x		
108	Lin et al., 2019				x			
109	Lin et al., 2023							
110	Lu et al., 2021a							
111	Lu et al., 2021b	x		x				
112	Luke et al., 2019	x						
113	Lunz et al., 2020	x				x		
114	Ly et al., 2022							
115	Ma et al., 2020							
116	Ma et al., 2022					x		x
117	Ma et al., 2023b			x				
118	Ma et al., 2023a							
119	Madasamy et al., 2022					x	x	
120	Madhumathy et al., 2022							
121	Maneas et al., 2023	x						

#	reference	geometric shapes	pattern phantom	model-based vasculature	model-based tissue	seg.-based vasculature	seg.-based tissue	DL-based tissue
122	Manwar et al., 2020							
123	Meng et al., 2022							
124	Moustakidis et al., 2019							
125	Nitkunanantharajah et al., 2020							
126	Nölke et al., 2021	x						
127	Olefir et al., 2020	x		x				
128	Orozco et al., 2021			x				
129	Ozdemir et al., 2022				x			
130	Pan et al., 2023	x				x		
131	Park et al., 2021							
132	Patil et al., 2021							
133	Rajendran et al., 2020	x		x				
134	Rajendran et al., 2021	x		x				
135	Rajendran et al., 2022	x		x				
136	Rajendran et al., 2023	x		x				
137	Ramos-Vega et al., 2022							
138	Refaee et al., 2021							
139	Reiter et al., 2017	x						
140	Ren et al., 2021b							
141	Sahlström et al., 2023a					x	x	
142	Sahlström et al., 2023b		x			x		
143	Schellenberg et al., 2022b							
144	Schellenberg et al., 2022a			x		x		x
145	Schlereth et al., 2022							
146	Schwab et al., 2019b		x					

#	reference	geometric shapes	pattern phantom	model-based vasculature	model-based tissue	seg.-based vasculature	seg.-based tissue	DL-based tissue
147	Schwab et al., 2019a		x					
148	Schwab et al., 2018					x		
149	Seong et al., 2023							
150	Shahid et al., 2022							
151	Shahid et al., 2021b							
152	Shahid et al., 2021a							
153	Shan et al., 2019b		x					
154	Shan et al., 2019a							
155	Sharma et al., 2020	x						
156	Shen et al., 2021	x						
157	Shi et al., 2022	x						
158	Singh et al., 2020							
159	Song et al., 2021							
160	Song et al., 2022	x						
161	Song et al., 2020							
162	Sun et al., 2021a	x				x		
163	Sun et al., 2021b				x			
164	Susmelj et al., 2022				x			
165	Sweeney et al., 2023			x				
166	Tang et al., 2023					x		
167	Tong et al., 2020					x		
168	Tserevelakis et al., 2023							
169	Tserevelakis et al., 2022							
170	Vera et al., 2023					x		

#	reference	geometric shapes	pattern phantom	model-based vasculature	model-based tissue	seg.-based vasculature	seg.-based tissue	DL-based tissue
171	Vousten et al., 2023							
172	Vu et al., 2021							
173	Vu et al., 2020	x				x		
174	Waibel et al., 2018	x						
175	Wang et al., 2022b					x		
176	Wang et al., 2021							
177	Warrier et al., 2022							
178	Wu et al., 2017							
179	Xiao et al., 2022	x						
180	Yang et al., 2019c	x						
181	Yang et al., 2019a						x	
182	Yang et al., 2019b					x		
183	Yazdani et al., 2021	x						
184	Yu et al., 2021							
185	Yuan et al., 2020							
186	Yue et al., 2022			x				
187	Zhang et al., 2023a						x	
188	Zhang et al., 2020			x				
189	Zhang et al., 2021a	x		x				
190	Zhang et al., 2018a							
191	Zhang et al., 2018b					x		
192	Zhang et al., 2021c							
193	Zhang et al., 2022a					x		
194	Zhang et al., 2021b				x			
195	Zhang et al., 2022b							
196	Zhang et al., 2021d							

#	reference	geometric shapes	pattern phantom	model-based vasculature	model-based tissue	seg.-based vasculature	seg.-based tissue	DL-based tissue
197	Zhang et al., 2022e							
198	Zhang et al., 2021f							
199	Zhang et al., 2023b							
200	Zhang et al., 2022c							
201	Zhang et al., 2022d			x				
202	Zhao et al., 2021							
203	Zhao et al., 2022							
204	Zhao et al., n.d.							
205	Zhao et al., 2020	x	x			x		
206	Zhao et al., 2023							
208	Zheng et al., 2022a				x			
209	Zheng et al., 2022b				x			
210	Zheng et al., 2022c			x				
211	Zhou et al., 2021a							
212	Zhou et al., 2019							
213	Zhou et al., 2022							
214	Zhou et al., 2021b							
215	Zhu et al., 2022b							
216	Zou et al., 2022	x						
217	Zou et al., 2023		x					

Bibliography

Ackerman, Michael J (1998). "The visible human project". In: *Proceedings of the IEEE*. Vol. 86, no. 3, pp. 504–511. DOI: 10.1109/5.662875.

Aggarwal, Charu (2023). "Correction to: neural networks and deep learning". In: *Neural Networks and Deep Learning: A Textbook*. Springer. DOI: 10.1007/978-3-031-29642-0.

Aggrawal, Deepika, Zafar, Mohsin, Schonfeld, Dan, and Avanaki, Kamran (2022). "Deep learning-boosted photoacoustic microscopy with an extremely low energy laser". In Proceedings of: *Photons Plus Ultrasound: Imaging and Sensing*. Vol. 11960. International Society for Optics and Photonics, pp. 256–265. DOI: 10.1117/12.2613061.

Aggrawal, Deepika et al. (2023). "E-Unet: a deep learning method for photoacoustic signal enhancement". In Proceedings of: *Photons Plus Ultrasound: Imaging and Sensing*. Vol. 12379. International Society for Optics and Photonics, pp. 189–193. DOI: 10.1117/12.2651217.

Agrawal, Sumit, Gaddale, Prameth, Karri, Sri Phani Krishna, and Kothapalli, Sri-Rajasekhar (2021a). "Learning optical scattering through symmetrical orthogonality enforced independent components for unmixing deep tissue photoacoustic signals". In: *IEEE Sensors Letters*. Vol. 5, no. 5, pp. 1–4. DOI: 10.1109/LSENS.2021.3073927.

Agrawal, Sumit, Suresh, Thaarakh, Garikipati, Ankit, Dangi, Ajay, and Kothapalli, Sri-Rajasekhar (2021b). "Modeling combined ultrasound and photoacoustic imaging: simulations aiding device development and artificial intelligence". In: *Photoacoustics*. Vol. 24, p. 100304. DOI: 10.1016/j.pacs.2021.100304.

Agrawal, Sumit et al. (2021c). "In vivo demonstration of reflection artifact reduction in LED-based photoacoustic imaging using deep learning". In Proceedings of: *Photons Plus Ultrasound: Imaging and Sensing*. Vol. 11642. International Society for Optics and Photonics, pp. 107–116. DOI: 10.1117/12.2579082.

Allen-Zhu, Zeyuan and Li, Yuanzhi (2020). "Towards understanding ensemble, knowledge distillation and self-distillation in deep learning". In: *arXiv preprint arXiv:2012.09816*.

Allman, Derek, Assis, Fabrizio, Chrispin, Jonathan, and Bell, Muyinatu A Lediju (2019). "A deep learning-based approach to identify in vivo catheter tips during photoacoustic-guided cardiac interventions". In Proceedings of: *Photons Plus Ultrasound: Imaging and Sensing*. Vol. 10878. International Society for Optics and Photonics, pp. 454–460. DOI: 10.1117/12.2510993.

Allman, Derek, Reiter, Austin, and Bell, Muyinatu A Lediju (2018). "Photoacoustic source detection and reflection artifact removal enabled by deep learning". In: *IEEE Transactions on Medical Imaging*. Vol. 37, no. 6, pp. 1464–1477. DOI: 10.1109/TMI.2018.2829662.

Anas, Emran Mohammad Abu, Zhang, Haichong K, Audigier, Chloé, and Boctor, Emad M (2018a). "Robust photoacoustic beamforming using dense convolutional neural networks". In Proceedings of: *Simulation, Image Processing, and Ultrasound Systems for Assisted Diagnosis and Navigation*. Springer, pp. 3–11. DOI: 10.1007/978-3-030-01045-4_1.

Anas, Emran Mohammad Abu, Zhang, Haichong K, Kang, Jin, and Boctor, Emad (2018b). "Enabling fast and high quality LED photoacoustic imaging: a recurrent neural networks based approach". In: *Biomedical Optics Express*. Vol. 9, no. 8, pp. 3852–3866. DOI: 10.1364/BOE.9.003852.

Antholzer, Stephan and Haltmeier, Markus (2021). "Discretization of learned NETT regularization for solving inverse problems". In: *Journal of Imaging*. Vol. 7, no. 11, p. 239. DOI: 10.3390/jimaging7110239.

Antholzer, Stephan, Haltmeier, Markus, Nuster, Robert, and Schwab, Johannes (2018a). "Photoacoustic image reconstruction via deep learning". In Proceedings of: *Photons Plus Ultrasound: Imaging and Sensing*. Vol. 10494. International Society for Optics and Photonics, pp. 433–442. DOI: doi.org/10.1117/12.2290676.

Antholzer, Stephan, Haltmeier, Markus, and Schwab, Johannes (2019a). "Deep learning for photoacoustic tomography from sparse data". In: *Inverse Problems in Science and Engineering*. Vol. 27, no. 7, pp. 987–1005. DOI: 10.1080/17415977.2018.1518444.

Antholzer, Stephan, Schwab, Johannes, Bauer-Marschallinger, Johnnes, Burgholzer, Peter, and Haltmeier, Markus (2019b). "NETT regularization for compressed sensing photoacoustic tomography". In Proceedings of: *Photons Plus Ultrasound: Imaging and Sensing*. Vol. 10878. International Society for Optics and Photonics, pp. 272–282. DOI: 10.1117/12.2508486.

Antholzer, Stephan, Schwab, Johannes, and Haltmeier, Markus (2018b). "Deep learning versus L1-minimization for compressed sensing photoacoustic tomography". In Proceedings of: *IEEE International Ultrasonics Symposium*. IEEE, pp. 206–212. DOI: 10.1109/ULTSYM.2018.8579737.

Arjovsky, Martin, Chintala, Soumith, and Bottou, Léon (2017). "Wasserstein generative adversarial networks". In Proceedings of: *International Conference on Machine Learning*. PMLR, pp. 214–223. URL: https://proceedings.mlr.press/v70/arjovsky17a.html.

Asgari Taghanaki, Saeid, Abhishek, Kumar, Cohen, Joseph Paul, Cohen-Adad, Julien, and Hamarneh, Ghassan (2021). "Deep semantic segmentation of natural and medical images: a review". In: *Artificial Intelligence Review*. Vol. 54, pp. 137–178. DOI: 10.1007/s10462-020-09854-1.

Ashraf, Tariq et al. (2010). "Size of radial and ulnar artery in local population". In: *JPMA-Journal of the Pakistan Medical Association*. Vol. 60, no. 10, p. 817.

Assi, Hisham et al. (2023). "A review of a strategic roadmapping exercise to advance clinical translation of photoacoustic imaging: From current barriers to future adoption". In: *Photoacoustics*, p. 100539. DOI: 10.1016/j.pacs.2023.100539.

Athira, S and Anoop, S (2022). "Image enhancement in reconstructed photoacoustic microscopy images using deep learning". In Proceedings of: *International Conference on Innovations in Science and Technology for Sustainable Development*. IEEE, pp. 269–274. DOI: 10.1109/ICISTSD55159.2022.10010536.

Attia, Amalina Binte Ebrahim et al. (2019). "A review of clinical photoacoustic imaging: current and future trends". In: *Photoacoustics*. Vol. 16, p. 100144. DOI: 10.1016/j.pacs.2019.100144.

Awasthi, Navchetan, Jain, Gaurav, Kalva, Sandeep Kumar, Pramanik, Manojit, and Yalavarthy, Phaneendra K (2020). "Deep neural network-based sinogram super-resolution and bandwidth enhancement for limited-data photoacoustic tomography". In: *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*. Vol. 67, no. 12, pp. 2660–2673. DOI: 10.1109/TUFFC.2020.2977210.

Awasthi, Navchetan et al. (2019). "PA-Fuse: deep supervised approach for the fusion of photoacoustic images with distinct reconstruction characteristics". In: *Biomedical Optics Express*. Vol. 10, no. 5, pp. 2227–2243. DOI: 10.1364/BOE.10.002227.

Beard, Paul (2011). "Biomedical photoacoustic imaging". In: *Interface Focus*. Vol. 1, no. 4, pp. 602–631. DOI: 10.1098/rsfs.2011.0028.

Bell, Muyinatu A Lediju (2019). "Deep learning the sound of light to guide surgeries". In Proceedings of: *Advanced Biomedical and Clinical Diagnostic and Surgical Guidance Systems XVII*. Vol. 10868. International Society for Optics and Photonics, pp. 34–39. DOI: 10.1117/12.2521315.

Bench, Ciaran and Cox, Ben T (2023). "Enhancing synthetic training data for quantitative photoacoustic tomography with generative deep learning". In: *arXiv preprint arXiv:2305.04714*.

Bench, Ciaran, Hauptmann, Andreas, and Cox, Ben (2020). "Toward accurate quantitative photoacoustic imaging: learning vascular blood oxygen saturation in three dimensions". In: *Journal of Biomedical Optics*. Vol. 25, no. 8, pp. 085003–085003. DOI: 10.1117/1.JBO.25.8.085003.

Biggs, Felix, Schrab, Antonin, and Gretton, Arthur (2023). "MMD-FUSE: learning and combining kernels for two-sample testing without data splitting". In: *arXiv preprint arXiv:2306.08777*.

Boink, Yoeri E, Manohar, Srirang, and Brune, Christoph (2019). "A partially-learned algorithm for joint photo-acoustic reconstruction and segmentation". In: *IEEE Transactions on Medical Imaging*. Vol. 39, no. 1, pp. 129–139. DOI: 10.1109/TMI.2019.2922026.

Brochu, Frederic M et al. (2016). "Towards quantitative evaluation of tissue absorption coefficients using light fluence correction in optoacoustic tomography". In: *IEEE Transactions on Medical Imaging*. Vol. 36, no. 1, pp. 322–331. DOI: 10.1109/TMI.2016.2607199.

Bronstein, Michael M, Bruna, Joan, LeCun, Yann, Szlam, Arthur, and Vandergheynst, Pierre (2017). "Geometric deep learning: going beyond euclidean data". In: *IEEE Signal Processing Magazine*. Vol. 34, no. 4, pp. 18–42. DOI: 10.1109/MSP.2017.2693418.

Cai, Chuangjian, Deng, Kexin, Ma, Cheng, and Luo, Jianwen (2018). "End-to-end deep neural network for optical inversion in quantitative photoacoustic imaging". In: *Optics Letters*. Vol. 43, no. 12, pp. 2752–2755. DOI: 10.1364/OL.43.002752.

Cao, Yingchun et al. (2017). "Spectral analysis assisted photoacoustic imaging for lipid composition differentiation". In: *Photoacoustics*. Vol. 7, pp. 12–19. DOI: 10.1016/j.pacs.2017.05.002.

Caron, Mathilde et al. (2021). "Emerging properties in self-supervised vision transformers". In Proceedings of: *IEEE/CVF International Conference on Computer Vision*, pp. 9650–9660. URL: https://openaccess.thecvf.com/content/ICCV2021/papers/Caron_Emerging_Properties_in_Self-Supervised_Vision_Transformers_ICCV_2021_paper.pdf.

Cassidy, Paul J and Radda, George K (2005). "Molecular imaging perspectives". In: *Journal of the Royal Society Interface*. Vol. 2, no. 3, pp. 133–144. DOI: 10.1098/rsif.2005.0040.

Chartsias, Agisilaos et al. (2019). "Disentangled representation learning in cardiac image analysis". In: *Medical Image Analysis*. Vol. 58, p. 101535. DOI: 10.1016/j.media.2019.101535.

Chatterji, Niladri S, Long, Philip M, and Bartlett, Peter L (2021). "When does gradient descent with logistic loss find interpolating two-layer networks?" In: *The Journal of Machine Learning Research*. Vol. 22, no. 1, pp. 7135–7182. URL: https://jmlr.csail.mit.edu/papers/volume22/20-1372/20-1372.pdf.

Chen, Panpan, Liu, Chengcheng, Feng, Ting, Li, Yong, and Ta, Dean (2020a). "Improved photoacoustic imaging of numerical bone model based on attention block U-Net deep learning network". In: *Applied Sciences*. Vol. 10, no. 22, p. 8089. DOI: 10.3390/app10228089.

Chen, Tingting et al. (2020b). "A deep learning method based on U-Net for quantitative photoacoustic imaging". In Proceedings of: *Photons Plus Ultrasound: Imaging and Sensing*. Vol. 11240. International Society for Optics and Photonics, pp. 216–223. DOI: 10.1117/12.2543173.

Chen, Xingxing, Qi, Weizhi, and Xi, Lei (2019). "Deep-learning-based motion-correction algorithm in optical resolution photoacoustic microscopy". In: *Visual Computing for Industry, Biomedicine, and Art*. Vol. 2, pp. 1–6. DOI: 10.1186/s42492-019-0022-9.

Cheng, Shengfu et al. (2022). "High-resolution photoacoustic microscopy with deep penetration through learning". In: *Photoacoustics*. Vol. 25, p. 100314. DOI: 10.1016/j.pacs.2021.100314.

Chlis, Nikolaos-Kosmas et al. (2020). "A sparse deep learning approach for automatic segmentation of human vasculature in multispectral optoacoustic tomography". In: *Photoacoustics*. Vol. 20, p. 100203. DOI: 10.1016/j.pacs.2020.100203.

Cho, Jang Hyun, Mall, Utkarsh, Bala, Kavita, and Hariharan, Bharath (2021). "PiCIE: Unsupervised semantic segmentation using invariance and equivariance in clustering". In Proceedings of: *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16794–16804. URL: https://openaccess.thecvf.com/content/CVPR2021/papers/Cho_PiCIE_Unsupervised_Semantic_Segmentation_Using_Invariance_and_Equivariance_in_Clustering_CVPR_2021_paper.pdf.

Choi, Seongwook et al. (2023). "Deep learning enhances multiparametric dynamic volumetric photoacoustic computed tomography in vivo (DL-PACT)". In: *Advanced Science*. Vol. 10, no. 1, p. 2202089. DOI: 10.1002/advs.202202089.

Chowdhury, Kaushik Basak, Bader, Maximilian, Dehner, Christoph, Jüstel, Dominik, and Ntziachristos, Vasilis (2021). "Individual transducer impulse response characterization method to improve image quality of array-based handheld optoacoustic tomography". In: *Optics Letters*. Vol. 46, no. 1, pp. 1–4. DOI: 10.1364/OL.412661.

Chowdhury, Kaushik Basak, Prakash, Jaya, Karlas, Angelos, Jüstel, Dominik, and Ntziachristos, Vasilis (2020). "A synthetic total impulse response characterization method for correction of hand-held optoacoustic images". In: *IEEE Transactions on Medical Imaging*. Vol. 39, no. 10, pp. 3218–3230. DOI: 10.1109/TMI.2020.2989236.

Çiçek, Özgün, Abdulkadir, Ahmed, Lienkamp, Soeren S, Brox, Thomas, and Ronneberger, Olaf (2016). "3D U-Net: learning dense volumetric segmentation from sparse annotation".

In Proceedings of: *Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 424–432. DOI: 10.1007/978-3-319-46723-8_49.

Costa, Pedro et al. (2017a). "End-to-end adversarial retinal image synthesis". In: *IEEE Transactions on Medical Imaging*. Vol. 37, no. 3, pp. 781–791. DOI: 10.1109/TMI.2017.2759102.

Costa, Pedro et al. (2017b). "Towards adversarial retinal image synthesis". In: *arXiv preprint arXiv:1701.08974*.

Cox, Ben T and Beard, Paul C (2005). "Fast calculation of pulsed photoacoustic fields in fluids using k-space methods". In: *The Journal of the Acoustical Society of America*. Vol. 117, no. 6, pp. 3616–3627. DOI: 10.1121/1.1920227.

Cox, Benjamin T, Arridge, Simon R, Köstli, Kornel P, and Beard, Paul C (2006). "Two-dimensional quantitative photoacoustic image reconstruction of absorption distributions in scattering media by use of a simple iterative method". In: *Applied Optics*. Vol. 45, no. 8, pp. 1866–1875. DOI: 10.1364/AO.45.001866.

Cox, BT, Arridge, SR, and Beard, PC (2007). "Gradient-based quantitative photoacoustic image reconstruction for molecular imaging". In Proceedings of: *Photons Plus Ultrasound: Imaging and Sensing*. Vol. 6437. International Society for Optics and Photonics, pp. 445–454. DOI: 10.1117/12.700031.

Czuchnowski, Jakub and Prevedel, Robert (2021). "Adaptive optics enhanced sensitivity in Fabry-Pérot based photoacoustic tomography". In: *Photoacoustics*. Vol. 23, p. 100276. DOI: 10.1016/j.pacs.2021.100276.

Davoudi, Neda, Deán-Ben, Xosé Luís, and Razansky, Daniel (2019). "Deep learning optoacoustic tomography with sparse data". In: *Nature Machine Intelligence*. Vol. 1, no. 10, pp. 453–460. DOI: 10.1038/s42256-019-0095-3.

Davoudi, Neda, Lafci, Berkan, Özbek, Ali, Deán-Ben, Xosé Luís, and Razansky, Daniel (2021). "Deep learning of image-and time-domain data enhances the visibility of structures in optoacoustic tomography". In: *Optics Letters*. Vol. 46, no. 13, pp. 3029–3032. DOI: 10.1364/OL.424571.

Dehghani, Hamid et al. (2009). "Near infrared optical tomography using NIRFAST: algorithm for numerical model and image reconstruction". In: *Communications in Numerical Methods in Engineering*. Vol. 25, no. 6, pp. 711–732. DOI: 10.1002/cnm.1162.

Dehner, Christoph, Olefir, Ivan, Chowdhury, Kaushik Basak, Jüstel, Dominik, and Ntziachristos, Vasilis (2022a). "Deep-learning-based electrical noise removal enables high spectral optoacoustic contrast in deep tissue". In: *IEEE Transactions on Medical Imaging*. Vol. 41, no. 11, pp. 3182–3193. DOI: 10.1109/TMI.2022.3180115.

Dehner, Christoph, Zahnd, Guillaume, Ntziachristos, Vasilis, and Jüstel, Dominik (2022b). "DeepMB: deep neural network for real-time model-based optoacoustic image reconstruction with adjustable speed of sound". In: *arXiv preprint arXiv:2206.14485*.

Deng, Handi, Wang, Xuanhao, Cai, Chuangjian, Luo, Jianwen, and Ma, Cheng (2019). "Machine-learning enhanced photoacoustic computed tomography in a limited view configuration". In Proceedings of: *Advanced Optical Imaging Technologies II*. Vol. 11186. International Society for Optics and Photonics, pp. 52–59. DOI: 10.1117/12.2539148.

Devlin, Jacob, Chang, Ming-Wei, Lee, Kenton, and Toutanova, Kristina (2018). "BERT: pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805*.

Dhengre, Nikhil, Sinha, Saugata, Chinni, Bhargava, Dogra, Vikram, and Rao, Navalgund (2020). "Computer aided detection of prostate cancer using multiwavelength photoacoustic data with convolutional neural network". In: *Biomedical Signal Processing and Control*. Vol. 60, p. 101952. DOI: 10.1016/j.bspc.2020.101952.

Dice, Lee R (1945). "Measures of the amount of ecologic association between species". In: *Ecology*. Vol. 26, no. 3, pp. 297–302. DOI: 10.2307/1932409.

DiSpirito, Anthony et al. (2020). "Reconstructing undersampled photoacoustic microscopy images using deep learning". In: *IEEE Transactions on Medical Imaging*. Vol. 40, no. 2, pp. 562–570. DOI: 10.1109/TMI.2020.3031541.

Dreher, K. K., Gröhl, J., Adler, T., Krichner, T., and Maier-Hein, L. (2020). "Towards realistic simulation of photoacoustic images." In Proceedings of: *Photons Plus Ultrasound: Imaging and Sensing*. International Society for Optics and Photonics.

Dreher, Kris K et al. (2023). "Unsupervised domain transfer with conditional invertible neural networks". In: *arXiv preprint arXiv:2303.10191*.

Drozdzal, Michal, Vorontsov, Eugene, Chartrand, Gabriel, Kadoury, Samuel, and Pal, Chris (2016). "The importance of skip connections in biomedical image segmentation". In Proceedings of: *Deep Learning in Medical Image Analysis, Large-Scale Annotation of Biomedical Data and Expert Label Synthesis*. Springer, pp. 179–187. DOI: 10.1007/978-3-319-46976-8_19.

Du, Getao, Cao, Xu, Liang, Jimin, Chen, Xueli, and Zhan, Yonghua (2020). "Medical image segmentation based on U-Net: a review." In: *Journal of Imaging Science & Technology*. Vol. 64, no. 2. DOI: 10.1088/1742-6596/2547/1/012010.

Durairaj, Deepit Abhishek et al. (2020). "Unsupervised deep learning approach for photoacoustic spectral unmixing". In Proceedings of: *Photons Plus Ultrasound: Imaging and*

Sensing. Vol. 11240. International Society for Optics and Photonics, pp. 173–181. DOI: 10.1117/12.2546964.

Dziugaite, Gintare Karolina, Roy, Daniel M, and Ghahramani, Zoubin (2015). "Training generative neural networks via maximum mean discrepancy optimization". In: *arXiv preprint arXiv:1505.03906*.

Else, Thomas, Gröhl, Janek, Hacker, Lina, and Bohndiek, Sarah E (2023). "PATATO: a Python photoacoustic tomography analysis toolkit". In Proceedings of: *Photons Plus Ultrasound: Imaging and Sensing*. International Society for Optics and Photonics, PC123790T. DOI: 10.1117/12.2648830.

Esser, Patrick, Sutter, Ekaterina, and Ommer, Björn (2018). "A variational U-Net for conditional appearance and shape generation". In Proceedings of: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8857–8866. URL: https://openaccess.thecvf.com/content_cvpr_2018/papers/Esser_A_Variational_U-Net_CVPR_2018_paper.pdf.

Fadden, Christopher and Kothapalli, Sri-Rajasekhar (2018). "A single simulation platform for hybrid photoacoustic and RF-acoustic computed tomography". In: *Applied Sciences*. Vol. 8, no. 9, p. 1568. DOI: 10.3390/app8091568.

Fang, Qianqian and Boas, David A (2009). "Monte Carlo simulation of photon migration in 3D turbid media accelerated by graphics processing units". In: *Optics express*. Vol. 17, no. 22, pp. 20178–20190. DOI: 10.1364/OE.17.020178.

Farnia, Parastoo et al. (2020). "High-quality photoacoustic image reconstruction based on deep convolutional neural network: towards intra-operative photoacoustic imaging". In: *Biomedical Physics & Engineering Express*. Vol. 6, no. 4, p. 045019. DOI: 10.1088/2057-1976/ab9a10.

Feng, Fei, Liang, Siqi, Luo, Jiajia, and Chen, Sung-Liang (2022). "High-fidelity deconvolution for acoustic-resolution photoacoustic microscopy enabled by convolutional neural networks". In: *Photoacoustics*. Vol. 26, p. 100360. DOI: 10.1016/j.pacs.2022.100360.

Feng, Jinchao et al. (2020). "End-to-end Res-Unet based reconstruction algorithm for photoacoustic imaging". In: *Biomedical Optics Express*. Vol. 11, no. 9, pp. 5321–5340. DOI: 10.1364/BOE.396598.

Fenster, Aaron and Downey, Donal B (1996). "3-D ultrasound imaging: a review". In: *IEEE Engineering in Medicine and Biology magazine*. Vol. 15, no. 6, pp. 41–51. DOI: 10.1109/51.544511.

Fey, Matthias and Lenssen, Jan Eric (2019). "Fast graph representation learning with PyTorch Geometric". In: *arXiv preprint arXiv:1903.02428*.

Frangi, Alejandro F, Niessen, Wiro J, Vincken, Koen L, and Viergever, Max A (1998). "Multiscale vessel enhancement filtering". In Proceedings of: *Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 130–137. DOI: 10.1007/BFb0056195.

Frangi, Alejandro F, Tsaftaris, Sotirios A, and Prince, Jerry L (2018). "Simulation and synthesis in medical imaging". In: *IEEE Transactions on Medical Imaging*. Vol. 37, no. 3, pp. 673–679. DOI: 10.1109/TMI.2018.2800298.

Gao, Ya, Xu, Wenyi, Chen, Yiming, Xie, Weiya, and Cheng, Qian (2022). "Deep learning-based photoacoustic imaging of vascular network through thick porous media". In: *IEEE Transactions on Medical Imaging*. Vol. 41, no. 8, pp. 2191–2204. DOI: 10.1109/TMI.2022.3158474.

Gerl, Stefan et al. (2020). "A distance-based loss for smooth and continuous skin layer segmentation in optoacoustic images". In Proceedings of: *Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 309–319. DOI: 10.1007/978-3-030-59725-2_30.

Godefroy, Guillaume, Arnal, Bastien, and Bossy, Emmanuel (2021). "Compensating for visibility artefacts in photoacoustic imaging with a deep learning approach providing prediction uncertainties". In: *Photoacoustics*. Vol. 21, p. 100218. DOI: 10.1016/j.pacs.2020.100218.

Gong, Jiali, Lan, Hengrong, Gao, Feng, and Gao, Fei (2021). "Deep learning regularized acceleration for photoacoustic image reconstruction". In Proceedings of: *IEEE International Ultrasonics Symposium*. IEEE, pp. 1–4. DOI: 10.1109/IUS52206.2021.9593560.

Gonzalez, Eduardo A, Graham, Camryn A, and Lediju Bell, Muyinatu A (2021). "Acoustic frequency-based approach for identification of photoacoustic surgical biomarkers". In: *Frontiers in Photonics*. Vol. 2, p. 6. DOI: 10.3389/fphot.2021.716656.

González, Martín G and Vega, Leonardo Rey (2022). "Model-based fully dense UNet for image enhancement in software-defined optoacoustic tomography". In Proceedings of: *IEEE Biennial Congress of Argentina*. IEEE, pp. 1–6. DOI: 10.1109/ARGENCON55245.2022.9940135.

González, Martín G, Vera, Matias, and Vega, Leonardo J Rey (2023). "Combining band-frequency separation and deep neural networks for optoacoustic imaging". In: *Optics and Lasers in Engineering*. Vol. 163, p. 107471. DOI: 10.1016/j.optlaseng.2022.107471.

Goodfellow, Ian, Bengio, Yoshua, and Courville, Aaron (2016). *Deep learning*. http://www.deeplearningbook.org. MIT Press.

Goodfellow, Ian et al. (2014). "Generative adversarial nets". In: *Advances in Neural Information Processing Systems*. Vol. 27. URL: https://proceedings.neurips.cc/paper_files/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf.

Gopalan, Anitha et al. (2023). "Reconstructing the photoacoustic image with high quality using the deep neural network model". In: *Contrast Media & Molecular Imaging*. Vol. 2023. DOI: 10.1155/2023/1172473.

Grasso, Valeria, Willumeit-Römer, Regine, and Jose, Jithin (2022). "Superpixel spectral unmixing framework for the volumetric assessment of tissue chromophores: a photoacoustic data-driven approach". In: *Photoacoustics*. Vol. 26, p. 100367. DOI: 10.1016/j.pacs.2022.100367.

Gretton, Arthur, Borgwardt, Karsten, Rasch, Malte, Schölkopf, Bernhard, and Smola, Alex (2006). "A kernel method for the two-sample-problem". In: *Advances in Neural Information Processing Systems*. Vol. 19. URL: https://proceedings.neurips.cc/paper_files/paper/2006/file/e9fb2eda3d9c55a0d89c98d6c54b5b3e-Paper.pdf.

Gretton, Arthur, Borgwardt, Karsten M, Rasch, Malte J, Schölkopf, Bernhard, and Smola, Alexander (2012). "A kernel two-sample test". In: *The Journal of Machine Learning Research*. Vol. 13, no. 1, pp. 723–773. URL: https://www.jmlr.org/papers/volume13/gretton12a/gretton12a.pdf.

Gröhl, Janek, Dreher, Kris K, Schellenberg, Melanie, Seitel, Alexander, and Maier-Hein, Lena (2021a). "SIMPA: an open source toolkit for simulation and processing of photoacoustic images". In Proceedings of: *Photons Plus Ultrasound: Imaging and Sensing*. Vol. 11642. International Society for Optics and Photonics, p. 116423C. DOI: 10.1117/1.JBO.27.8.083010.

Gröhl, Janek, Hacker, Lina, and Cox, Ben (2023a). "Open-source implementation and systematic comparison of image reconstruction algorithms for linear array transducers". In Proceedings of: *Photons Plus Ultrasound: Imaging and Sensing*. International Society for Optics and Photonics, PC123790Y. DOI: 10.1117/12.2649934.

Gröhl, Janek, Kirchner, Thomas, Adler, Tim, and Maier-Hein, Lena (2018). "Confidence estimation for machine learning-based quantitative photoacoustics". In: *Journal of Imaging*. Vol. 4, no. 12, p. 147. DOI: 10.3390/jimaging4120147.

Gröhl, Janek, Schellenberg, Melanie, Dreher, Kris, and Maier-Hein, Lena (2021b). "Deep learning for biomedical photoacoustic imaging: a review". In: *Photoacoustics*. Vol. 22, p. 100241. DOI: 10.1016/j.pacs.2021.100241.

Gröhl, Janek et al. (2021c). "Learned spectral decoloring enables photoacoustic oximetry". In: *Scientific Reports*. Vol. 11, no. 1, p. 6565. DOI: https://doi.org/10.1038/s41598-021-83405-8.

Gröhl, Janek et al. (2021d). "Semantic segmentation of multispectral photoacoustic images using deep learning". In: *Photons Plus Ultrasound: Imaging and Sensing*, 116423F. DOI: 10.1117/12.2578135.

Gröhl, Janek et al. (2023b). "Moving beyond simulation: data-driven quantitative photoacoustic imaging using tissue-mimicking phantoms". In: *arXiv preprint arXiv:2306.06748*.

Grün, H, Paltauf, Guenther, Haltmeier, Markus, and Burgholzer, Peter (2007). "Photoacoustic tomography using a fiber based Fabry-Perot interferometer as an integrating line detector and image reconstruction by model-based time reversal method". In Proceedings of: *European Conference on Biomedical Optics*. Optica Publishing Group, 6631_6. DOI: 10.1364/ECBO.2007.6631_6.

Gu, Liujie et al. (2023). "Sentinel lymph node mapping in patients with breast cancer using a photoacoustic/ultrasound dual-modality imaging system with carbon nanoparticles as the contrast agent: a pilot study". In: *Biomedical Optics Express*. Vol. 14, no. 3, pp. 1003–1014. DOI: 10.1364/BOE.482126.

Guan, Steven, Hsu, Ko-Tsung, and Chitnis, Parag V (2021a). "Fourier neural operator networks: a fast and general solver for the photoacoustic wave equation". In: *arXiv preprint arXiv:2108.09374*.

Guan, Steven, Hsu, Ko-Tsung, Eyassu, Matthias, and Chitnis, Parag V (2021b). "Dense dilated UNet: deep learning for 3D photoacoustic tomography image reconstruction". In: *arXiv preprint arXiv:2104.03130*.

Guan, Steven, Khan, Amir A, Sikdar, Siddhartha, and Chitnis, Parag V (2020). "Limited-view and sparse photoacoustic tomography for neuroimaging with deep learning". In: *Scientific Reports*. Vol. 10, no. 1, p. 8510. DOI: 10.1038/s41598-020-65235-2.

Gubbi, Mardava R and Bell, Muyinatu A Lediju (2021). "Deep learning-based photoacoustic visual servoing: using outputs from raw sensor data as inputs to a robot controller". In Proceedings of: *IEEE International Conference on Robotics and Automation*. IEEE, pp. 14261–14267. DOI: 10.1109/ICRA48506.2021.9561369.

Guibas, John T, Virdi, Tejpal S, and Li, Peter S (2017). "Synthetic medical images from dual generative adversarial networks". In: *arXiv preprint arXiv:1709.01872*.

Gulenko, Oleksandra et al. (2022). "Deep-learning-based algorithm for the removal of electromagnetic interference noise in photoacoustic endoscopic image processing". In: *Sensors*. Vol. 22, no. 10, p. 3961. DOI: 10.3390/s22103961.

Guo, Mengjie, Lan, Hengrong, Yang, Changchun, Liu, Jiang, and Gao, Fei (2022). "AS-Net: fast photoacoustic reconstruction with multi-feature fusion from sparse data". In: *IEEE*

Transactions on Computational Imaging. Vol. 8, pp. 215–223. DOI: 10.1109/TCI.2022.3155379.

Gurney, Kevin (1997). *An introduction to neural networks*. CRC press.

Gutta, Sreedevi et al. (2017). "Deep neural network-based bandwidth enhancement of photoacoustic data". In: *Journal of Biomedical Optics*. Vol. 22, no. 11, pp. 116001–116001. DOI: 10.1117/1.JBO.22.11.116001.

Hakakzadeh, Soheil, Kavehvash, Zahra, and Pramanik, Manojit (2022). "Artifact removal factor for circular-view photoacoustic tomography". In Proceedings of: *IEEE International Ultrasonics Symposium*. IEEE, pp. 1–4. DOI: 10.1109/IUS54386.2022.9958228.

Hakimnejad, Hesam, Azimifar, Zohreh, and Nazemi, Mohammad Sadegh (2023). "Unsupervised photoacoustic tomography image reconstruction from limited-view unpaired data using an improved CycleGAN". In Proceedings of: *International Computer Conference, Computer Society of Iran*. IEEE, pp. 1–6. DOI: 10.1109/CSICC58665.2023.10105363.

Hariri, Ali, Alipour, Kamran, Mantri, Yash, Schulze, Jurgen P, and Jokerst, Jesse V (2020). "Deep learning improves contrast in low-fluence photoacoustic imaging". In: *Biomedical Optics Express*. Vol. 11, no. 6, pp. 3360–3373. DOI: 10.1364/BOE.395683.

Hauptmann, Andreas and Poimala, Jenni (2023). "Model-corrected learned primal-dual models for fast limited-view photoacoustic tomography". In: *arXiv preprint arXiv:2304.01963*.

Hauptmann, Andreas et al. (2018). "Model-based learning for accelerated, limited-view 3-D photoacoustic tomography". In: *IEEE Transactions on Medical Imaging*. Vol. 37, no. 6, pp. 1382–1393. DOI: 10.1109/TMI.2018.2820382.

He, Da, Zhou, Jiasheng, Shang, Xiaoyu, Luo, Jiajia, and Chen, Sung-Liang (2022). "De-noising of photoacoustic microscopy images by deep learning". In: *arXiv preprint arXiv:2201.04302*.

Hesamian, Mohammad Hesam, Jia, Wenjing, He, Xiangjian, and Kennedy, Paul (2019). "Deep learning techniques for medical image segmentation: achievements and challenges". In: *Journal of Digital Imaging*. Vol. 32, pp. 582–596. DOI: 10.1007/s10278-019-00227-x.

Hoffmann, Bianca et al. (2022). "Spatial quantification of clinical biomarker pharmacokinetics through deep learning-based segmentation and signal-oriented analysis of MSOT data". In: *Photoacoustics*. Vol. 26, p. 100361. DOI: 10.1016/j.pacs.2022.100361.

Holzwarth, Niklas et al. (2021a). "Tattoo tomography: an optical pattern approach for context-aware photoacoustics". In Proceedings of: *Photons Plus Ultrasound: Imaging and Sensing*. Vol. 11642. International Society for Optics and Photonics, p. 1164217. DOI: 10.1007/s11548-021-02399-w.

Holzwarth, Niklas et al. (2021b). "Tattoo tomography: freehand 3D photoacoustic image reconstruction with an optical pattern". In: *International Journal of Computer Assisted Radiology and Surgery*. Vol. 16, pp. 1101–1110. DOI: 10.1007/s11548-021-02399-w.

— (2023a). "Abstract: tattoo-tomographie". In Proceedings of: *Bildverarbeitung für die Medizin*. Springer, pp. 114–114. DOI: 10.1007/978-3-658-41657-7_25.

Holzwarth, Niklas et al. (2023b). "Clinical tattoo tomography". In Proceedings of: *Photons Plus Ultrasound: Imaging and Sensing*. International Society for Optics and Photonics, PC1237930. DOI: 10.1117/12.2649870.

Holzwarth, Niklas et al. (2023c). *Method and system for context-ware photoacoustic imaging*. US Patent App. 18/004,689.

Hsu, Ko-Tsung, Guan, Steven, and Chitnis, Parag V (2023). "Fast iterative reconstruction for photoacoustic tomography using learned physical model: theoretical validation". In: *Photoacoustics*. Vol. 29, p. 100452. DOI: 10.1016/j.pacs.2023.100452.

Hu, Yexing et al. (2022). "Deep learning facilitates fully automated brain image registration of optoacoustic tomography and magnetic resonance imaging". In: *Biomedical Optics Express*. Vol. 13, no. 9, pp. 4817–4833. DOI: 10.1364/BOE.458182.

Hubmer, Martin G et al. (2004). "The posterior interosseous artery in the distal part of the forearm. Is the term 'recurrent branch of the anterior interosseous artery' justified?" In: *British Journal of Plastic Surgery*. Vol. 57, no. 7, pp. 638–644. DOI: 10.1016/j.bjps.2004.06.011.

Hübner, Marco et al. (2023). "How to assess the realism of synthetic spectral images". In Proceedings of: *Molecular-Guided Surgery: Molecules, Devices, and Applications IX*. International Society for Optics and Photonics, PC1236104. DOI: 10.1117/12.2648461.

Hwang, Gyeongha, Jeon, Gihyeon, and Moon, Sunghwan (2023). "Self-supervised learning for a nonlinear inverse problem with forward operator involving an unknown function arising in photoacoustic tomography". In: *arXiv preprint arXiv:2301.08693*.

Isensee, Fabian, Jaeger, Paul F, Kohl, Simon AA, Petersen, Jens, and Maier-Hein, Klaus H (2021). "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation". In: *Nature Methods*. Vol. 18, no. 2, pp. 203–211. DOI: 10.1038/s41592-020-01008-z.

Isola, Phillip, Zhu, Jun-Yan, Zhou, Tinghui, and Efros, Alexei A (2017). "Image-to-image translation with conditional adversarial networks". In Proceedings of: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1125–1134. URL: https://openaccess.thecvf.com/content_cvpr_2017/papers/Isola_Image-To-Image_Translation_With_CVPR_2017_paper.pdf.

iThera Medical GmbH (2021). *MSOT Acuity Echo*. URL: https://ithera-medical.com/products/clinical-research/msot-acuity-echo/ (visited on 08/15/2023).

Jacques, Steven L (2014). "Coupling 3D Monte Carlo light transport in optically heterogeneous tissues to photoacoustic signal generation". In: *Photoacoustics*. Vol. 2, no. 4, pp. 137–142. DOI: 10.1016/j.pacs.2014.09.001.

— (2015). *Generic tissue optical properties*. URL: https://omlc.org/news/feb15/generic_optics/index.html (visited on 11/11/2019).

Jeon, Seungwan, Choi, Wonseok, Park, Byullee, and Kim, Chulhong (2021). "A deep learning-based model that reduces speed of sound aberrations for improved in vivo photoacoustic imaging". In: *IEEE Transactions on Image Processing*. Vol. 30, pp. 8773–8784. DOI: 10.1109/TIP.2021.3120053.

Jeon, Seungwan and Kim, Chulhong (2020). "Deep learning-based speed of sound aberration correction in photoacoustic images". In Proceedings of: *Photons Plus Ultrasound: Imaging and Sensing*. Vol. 11240. International Society for Optics and Photonics, pp. 24–27. DOI: /10.1117/12.2543440.

Ji, Xu, Henriques, Joao F, and Vedaldi, Andrea (2019). "Invariant information clustering for unsupervised image classification and segmentation". In Proceedings of: *IEEE/CVF International Conference on Computer Vision*, pp. 9865–9874. URL: https://openaccess.thecvf.com/content_ICCV_2019/papers/Ji_Invariant_Information_Clustering_for_Unsupervised_Image_Classification_and_Segmentation_ICCV_2019_paper.pdf.

Jiang, Daohuai et al. (2022). "Hand-held 3D photoacoustic imager with GPS". In: *arXiv preprint arXiv:2203.09048*.

Jiang, Zhuoran et al. (2023). "Radiation-induced acoustic signal denoising using a supervised deep learning framework for imaging and therapy monitoring". In: *arXiv preprint arXiv:2304.13868*.

Jnawali, Kamal, Chinni, Bhargava, Dogra, Vikram, and Rao, Navalgund (2019a). "Transfer learning for automatic cancer tissue detection using multispectral photoacoustic imaging". In Proceedings of: *Medical Imaging 2019: Computer-Aided Diagnosis*. Vol. 10950. International Society for Optics and Photonics, pp. 982–987. DOI: 10.1117/12.2506950.

— (2020). "Automatic cancer tissue detection using multispectral photoacoustic imaging". In: *International Journal of Computer Assisted Radiology and Surgery*. Vol. 15, pp. 309–320. DOI: 10.1007/s11548-019-02101-1.

Jnawali, Kamal, Chinni, Bhargava, Dogra, Vikram, Sinha, Saugata, and Rao, Navalgund (2019b). "Deep 3D convolutional neural network for automatic cancer tissue detection using multispectral photoacoustic imaging". In Proceedings of: *Medical Imaging 2019: Ultrasonic*

Imaging and Tomography. Vol. 10955. International Society for Optics and Photonics, pp. 299–306. DOI: 10.1117/12.2518686.

Johnson, Justin (2019). *Deep learning for computer vision (UMich EECS 498-007)*. URL: https://www.youtube.com/playlist?list=PLLhQgjrONLVFP1E7p2jWMMeM2FWUf2Qc7 (visited on 09/12/2023).

Johnson, Justin, Gupta, Agrim, and Fei-Fei, Li (2018). "Image generation from scene graphs". In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1219–1228. URL: https://openaccess.thecvf.com/content_cvpr_2018/papers/Johnson_Image_Generation_From_CVPR_2018_paper.pdf.

Johnstonbaugh, Kerrick et al. (2019). "Novel deep learning architecture for optical fluence dependent photoacoustic target localization". In Proceedings of: *Photons Plus Ultrasound: Imaging and Sensing*. Vol. 10878. International Society for Optics and Photonics, pp. 95–102. DOI: 10.1117/12.2511015.

Johnstonbaugh, Kerrick et al. (2020). "A deep learning approach to photoacoustic wavefront localization in deep-tissue medium". In: *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*. Vol. 67, no. 12, pp. 2649–2659. DOI: 10.1109/TUFFC.2020.2964698.

Joseph, Francis Kalloor et al. (2021). "Generative adversarial network-based photoacoustic image reconstruction from bandlimited and limited-view data". In Proceedings of: *Photons Plus Ultrasound: Imaging and Sensing*. Vol. 11642. International Society for Optics and Photonics, pp. 208–213. DOI: 10.1117/12.2577750.

Joyce, Thomas and Kozerke, Sebastian (2019). "3D medical image synthesis by factorised representation and deformable model learning". In Proceedings of: *Simulation and Synthesis in Medical Imaging*. Springer, pp. 110–119. DOI: 10.1007/978-3-030-32778-1_12.

Jumper, John et al. (2021). "Highly accurate protein structure prediction with AlphaFold". In: *Nature*. Vol. 596, no. 7873, pp. 583–589. DOI: 10.1038/s41586-021-03819-2.

Kar, Amlan et al. (2019). "Meta-sim: learning to generate synthetic datasets". In Proceedings of: *IEEE/CVF International Conference on Computer Vision*, pp. 4551–4560. URL: https://openaccess.thecvf.com/content_ICCV_2019/papers/Kar_Meta-Sim_Learning_to_Generate_Synthetic_Datasets_ICCV_2019_paper.pdf.

Kar, Mithun Kumar, Nath, Malaya Kumar, and Neog, Debanga Raj (2021). "A review on progress in semantic image segmentation and its application to medical images". In: *SN Computer Science*. Vol. 2, no. 5, p. 397. DOI: 10.1007/s42979-021-00784-5.

Karras, Tero, Laine, Samuli, and Aila, Timo (2019). "A style-based generator architecture for generative adversarial networks". In Proceedings of: *IEEE/CVF Conference on Computer Vi-*

sion and Pattern Recognition, pp. 4401–4410. URL: https://openaccess.thecvf.com/content_CVPR_2019/papers/Karras_A_Style-Based_Generator_Architecture_for_Generative_Adversarial_Networks_CVPR_2019_paper.pdf.

Karras, Tero et al. (2020a). "Analyzing and improving the image quality of StyleGAN". In Proceedings of: *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8110–8119. URL: https://openaccess.thecvf.com/content_CVPR_2020/papers/Karras_Analyzing_and_Improving_the_Image_Quality_of_StyleGAN_CVPR_2020_paper.pdf.

Karras, Tero et al. (2020b). "Training generative adversarial networks with limited data". In: *Advances in Neural Information Processing Systems*. Vol. 33, pp. 12104–12114. URL: https://papers.nips.cc/paper/2020/file/8d30aa96e72440759f74bd2306c1fa3d-Paper.pdf.

Keiser, Gerd (2016). *Biophotonics*. Springer. DOI: 10.1007/978-981-10-0945-7.

Kenhagho, Nguendon Hervé et al. (2021). "Machine learning-based optoacoustic tissue classification method for laser osteotomes using an air-coupled transducer". In: *Lasers in Surgery and Medicine*. Vol. 53, no. 3, pp. 377–389. DOI: 10.1002/lsm.23290.

Kerkhof, Faes D, Van Leeuwen, Timo, and Vereecke, Evie E (2018). "The digital human forearm and hand". In: *Journal of Anatomy*. Vol. 233, no. 5, pp. 557–566. DOI: 10.1111/joa.12877.

Kikkawa, Ryo, Kajita, Hiroki, Imanishi, Nobuaki, Aiso, Sadakazu, and Bise, Ryoma (2021). "Unsupervised body hair detection by positive-unlabeled learning in photoacoustic image". In Proceedings of: *IEEE Engineering in Medicine & Biology Society*. IEEE, pp. 3349–3352. DOI: 10.1109/EMBC46164.2021.9630720.

Kim, Jongbeom et al. (2022a). "Deep learning acceleration of multiscale superresolution localization photoacoustic imaging". In: *Light: Science & Applications*. Vol. 11, no. 1, p. 131. DOI: 10.1038/s41377-022-00820-w.

Kim, Jongbeom et al. (2022b). "Deep learning alignment of bidirectional raster scanning in high speed photoacoustic microscopy". In: *Scientific Reports*. Vol. 12, no. 1, p. 16238. DOI: 10.1038/s41598-022-20378-2.

Kim, MinWoo, Jeng, Geng-Shi, Pelivanov, Ivan, and O'Donnell, Matthew (2020). "Deep-learning image reconstruction for real-time photoacoustic system". In: *IEEE Transactions on Medical Imaging*. Vol. 39, no. 11, pp. 3379–3390. DOI: 10.1109/TMI.2020.2993835.

Kingma, Diederik P and Ba, Jimmy (2014). "Adam: a method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980*.

Kipf, Thomas N and Welling, Max (2016). "Semi-supervised classification with graph convolutional networks". In: *arXiv preprint arXiv:1609.02907*.

Kirchner, Thomas and Frenz, Martin (2021). "Multiple illumination learned spectral decoloring for quantitative optoacoustic oximetry imaging". In: *Journal of Biomedical Optics*. Vol. 26, no. 8, pp. 085001–085001. DOI: 10.3390/jimaging4100121.

Kirchner, Thomas, Gröhl, Janek, and Maier-Hein, Lena (2018a). "Context encoding enables machine learning-based quantitative photoacoustics". In: *Journal of Biomedical Optics*. Vol. 23, no. 5, pp. 056008–056008. DOI: 10.1117/1.JBO.23.5.056008.

Kirchner, Thomas, Sattler, Franz, Gröhl, Janek, and Maier-Hein, Lena (2018b). "Signed real-time delay multiply and sum beamforming for multispectral photoacoustic imaging". In: *Journal of Imaging*. Vol. 4, no. 10, p. 121.

Knieling, Ferdinand et al. (2017). "Multispectral optoacoustic tomography for assessment of Crohn's disease activity". In: *New England Journal of Medicine*. Vol. 376, no. 13, pp. 1292–1294. DOI: 10.1056/NEJMc1612455.

Koonce, Brett (2021). "EfficientNet". In: *Convolutional neural networks with swift for Tensorflow: image recognition and dataset categorization*. Apress, pp. 109–123. DOI: 10.1007/978-1-4842-6168-2_10.

Kose, O Deniz and Shen, Yanning (2022). "Fairnorm: fair and fast graph neural network training". In: *arXiv preprint arXiv:2205.09977*.

Lafci, Berkan, Merćep, Elena, Morscher, Stefan, Deán-Ben, Xosé Luís, and Razansky, Daniel (2020a). "Efficient segmentation of multi-modal optoacoustic and ultrasound images using convolutional neural networks". In Proceedings of: *Photons Plus Ultrasound: Imaging and Sensing*. Vol. 11240. International Society for Optics and Photonics, pp. 123–128. DOI: 10.1117/12.2543970.

Lafci, Berkan, Merčep, Elena, Morscher, Stefan, Deán-Ben, Xosé Luís, and Razansky, Daniel (2020b). "Deep learning for automatic segmentation of hybrid optoacoustic ultrasound (OPUS) images". In: *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*. Vol. 68, no. 3, pp. 688–696. DOI: 10.1109/TUFFC.2020.3022324.

Lan, Hengrong, Huang, Lijie, Nie, Liming, and Luo, Jianwen (2023a). "Cross-domain unsupervised reconstruction with equivariance for photoacoustic computed tomography". In: *arXiv preprint arXiv:2301.06681*.

Lan, Hengrong, Jiang, Daohuai, and Gao, Fei (2021a). "The limited-view compensation of photoacoustic tomography via deep learning". In Proceedings of: *Photons Plus Ultrasound: Imaging and Sensing*. Vol. 11642. International Society for Optics and Photonics, pp. 267–271. DOI: 10.1117/12.2577973.

Lan, Hengrong, Jiang, Daohuai, Gao, Feng, and Gao, Fei (2021b). "Deep learning enabled real-time photoacoustic tomography system via single data acquisition channel". In: *Photoacoustics*. Vol. 22, p. 100270. DOI: 10.1016/j.pacs.2021.100270.

Lan, Hengrong, Jiang, Daohuai, Yang, Changchun, Gao, Feng, and Gao, Fei (2020). "Y-Net: hybrid deep learning image reconstruction for photoacoustic tomography in vivo". In: *Photoacoustics*. Vol. 20, p. 100197. DOI: 10.1016/j.pacs.2020.100197.

Lan, Hengrong, Yang, Changchun, and Gao, Fei (2023b). "A jointed feature fusion framework for photoacoustic image reconstruction". In: *Photoacoustics*. Vol. 29, p. 100442. DOI: 10.1016/j.pacs.2022.100442.

Lan, Hengrong, Yang, Changchun, Jiang, Daohuai, and Gao, Fei (2019a). "Deep learning approach to reconstruct the photoacoustic image using multi-frequency data". In Proceedings of: *IEEE International Ultrasonics Symposium*. IEEE, pp. 487–489. DOI: 10.1109/ULTSYM.2019.8926287.

— (2019b). "Reconstruct the photoacoustic image based on deep learning with multi-frequency ring-shape transducer array". In Proceedings of: *IEEE Engineering in Medicine and Biology Society*. IEEE, pp. 7115–7118. DOI: 10.1109/EMBC.2019.8856590.

Lan, Hengrong, Zhang, Juze, Yang, Changchun, and Gao, Fei (2021c). "Compressed sensing for photoacoustic computed tomography based on an untrained neural network with a shape prior". In: *Biomedical Optics Express*. Vol. 12, no. 12, pp. 7835–7848. DOI: 10.1364/BOE.441901.

Lan, Hengrong et al. (2019c). "Ki-GAN: knowledge infusion generative adversarial network for photoacoustic image reconstruction in vivo". In Proceedings of: *Medical Image Computing and Computer Assisted Intervention*. Springer, pp. 273–281. DOI: 10.1007/978-3-030-32239-7_31.

Laufer, Jan et al. (2012). "In vivo preclinical photoacoustic imaging of tumor vasculature development and therapy". In: *Journal of Biomedical Optics*. Vol. 17, no. 5, pp. 056016–056016. DOI: 10.1117/1.JBO.17.5.056016.

Le, Thanh Dat, Kwon, Seong Young, and Lee, Changho (2021). "Performance comparison of feature generation algorithms for mosaic photoacoustic microscopy". In Proceedings of: *Photonics*. Vol. 8. no. 9. MDPI, p. 352. DOI: 10.3390/photonics8090352.

Le, Thanh Dat, Kwon, Seong-Young, and Lee, Changho (2022a). "Segmentation and quantitative analysis of photoacoustic imaging: a review". In Proceedings of: *Photonics*. Vol. 9. no. 3. MDPI, p. 176. DOI: 10.3390/photonics9030176.

Le, Thanh Dat and Lee, Changho (2022b). "Comparative study of feature generation algorithms for mosaic photoacoustic microscopy". In Proceedings of: *Photons Plus Ultrasound:*

Imaging and Sensing. Vol. 11960. International Society for Optics and Photonics, pp. 59–63. DOI: 10.1117/12.2608100.

LeCun, Yann, Bengio, Yoshua, and Hinton, Geoffrey (2015). "Deep learning". In: *Nature*. Vol. 521, no. 7553, pp. 436–444. DOI: 10.1038/nature14539.

Lee, Changyeop, Choi, Wonseok, Kim, Jeesu, and Kim, Chulhong (2020). "Three-dimensional clinical handheld photoacoustic/ultrasound scanner". In: *Photoacoustics*. Vol. 18, p. 100173. DOI: 10.1016/j.pacs.2020.100173.

Leino, Aleksi A, Pulkkinen, Aki, and Tarvainen, Tanja (2019). "ValoMC: a Monte Carlo software and MATLAB toolbox for simulating light transport in biological tissue". In: *Osa Continuum*. Vol. 2, no. 3, pp. 957–972. DOI: 10.1364/OSAC.2.000957.

Leng, Xiandong et al. (2021a). "Assessing rectal cancer treatment response using coregistered endorectal photoacoustic and US imaging paired with deep learning". In: *Radiology*. Vol. 299, no. 2, pp. 349–358. DOI: 10.1148/radiol.2021202208.

Leng, Xiandong et al. (2021b). "Rectal cancer treatment management: deep-learning neural network based on photoacoustic microscopy image outperforms histogram-feature-based classification". In: *Frontiers in Oncology*. Vol. 11, p. 715332. DOI: 10.3389/fonc.2021.715332.

Li, Daiqing, Kar, Amlan, Ravikumar, Nishant, Frangi, Alejandro F, and Fidler, Sanja (2020a). "Fed-Sim: federated simulation for medical imaging". In: *arXiv preprint arXiv:2009.00668*.

Li, Daiqing, Yang, Junlin, Kreis, Karsten, Torralba, Antonio, and Fidler, Sanja (2021a). "Semantic segmentation with generative models: semi-supervised learning and strong out-of-domain generalization". In Proceedings of: *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8300–8311. URL: https://openaccess.thecvf.com/content/CVPR2021/papers/Li_Semantic_Segmentation_With_Generative_Models_Semi-Supervised_Learning_and_Strong_Out-of-Domain_CVPR_2021_paper.pdf.

Li, Housen, Schwab, Johannes, Antholzer, Stephan, and Haltmeier, Markus (2020b). "NETT: solving inverse problems with deep neural networks". In: *Inverse Problems*. Vol. 36, no. 6, p. 065005. DOI: 10.1088/1361-6420/ab6d57.

Li, Jiao et al. (2022a). "Deep learning-based quantitative optoacoustic tomography of deep tissues in the absence of labeled experimental data". In: *Optica*. Vol. 9, no. 1, pp. 32–41. DOI: 10.1364/OPTICA.438502.

Li, Kang, Yu, Lequan, Wang, Shujun, and Heng, Pheng-Ann (2019). "Unsupervised retina image synthesis via disentangled representation learning". In Proceedings of: *Simulation*

and Synthesis in Medical Imaging. Springer, pp. 32–41. DOI: 10.1007/978-3-030-32778-1_4.

Li, Michelle M, Huang, Kexin, and Zitnik, Marinka (2022b). "Graph representation learning in biomedicine and healthcare". In: *Nature Biomedical Engineering*. Vol. 6, no. 12, pp. 1353–1369. DOI: https://doi.org/10.1038/s41551-022-00942-x.

Li, Sihang, Wang, Zhuangzhuang, Cao, Xiongjun, Li, Zhihui, and Song, Xianlin (2021b). "Large volumetric optical-resolution photoacoustic microscopy with image fusion based on CNN feature extraction". In Proceedings of: *Current Developments in Lens Design and Optical Engineering XXII*. Vol. 11814. International Society for Optics and Photonics, pp. 95–100. DOI: 10.1117/12.2592123.

Li, Sihang, Wang, Zhuangzhuang, Gu, Chenghao, and Song, Xianlin (2021c). "Extended depth of field photoacoustic microscopy using image fusion based on deep learning". In Proceedings of: *Illumination Optics VI*. Vol. 11874. International Society for Optics and Photonics, pp. 164–169. DOI: 10.1117/12.2600748.

Li, Sihang et al. (2021d). "High-resolution 3D fusion method for optical-resolution photoacoustic microscopy using deep learning". In Proceedings of: *SPIE Future Sensing Technologies*. Vol. 11914. International Society for Optics and Photonics, pp. 107–113. DOI: 10.1117/12.2612466.

Li, Wei-Xiang, Qin, Ze-Zheng, Liu, Guang-Xing, and Sun, Ming-Jian (2021e). "Deep learning reconstruction algorithm based on sparse photoacoustic tomography system". In Proceedings of: *IEEE International Ultrasonics Symposium*. IEEE, pp. 1–4. DOI: 10.1109/IUS52206.2021.9593711.

Li, Xiuting et al. (2021f). "Atopic dermatitis classification models of 3D optoacoustic mesoscopic images". In Proceedings of: *European Conference on Biomedical Optics*. Optica Publishing Group, ETu5B–7. DOI: 10.1364/ECBO.2021.ETu5B.7.

Li, Yujia, Swersky, Kevin, and Zemel, Rich (2015). "Generative moment matching networks". In Proceedings of: *International Conference on Machine Learning*. PMLR, pp. 1718–1727. DOI: http://proceedings.mlr.press/v37/li15.pdf.

Li, Zhuoan, Lan, Hengrong, and Gao, Fei (2021g). "Learned parameters and increment for iterative photoacoustic image reconstruction via deep learning". In Proceedings of: *IEEE Engineering in Medicine & Biology Society*. IEEE, pp. 2989–2992. DOI: 10.1109/EMBC46164.2021.9630545.

Li, Zongyi et al. (2020c). "Fourier neural operator for parametric partial differential equations". In: *arXiv preprint arXiv:2010.08895*.

— (2020d). "Neural operator: graph kernel network for partial differential equations". In: *arXiv preprint arXiv:2003.03485*.

Liang, Zhichao et al. (2022). "Automatic 3-D segmentation and volumetric light fluence correction for photoacoustic tomography based on optimal 3-D graph search". In: *Medical Image Analysis*. Vol. 75, p. 102275. DOI: 10.1016/j.media.2021.102275.

Lin, Li and Wang, Lihong V (2022). "The emerging role of photoacoustic imaging in clinical oncology". In: *Nature Reviews Clinical Oncology*. Vol. 19, no. 6, pp. 365–384. DOI: 10.1038/s41571-022-00615-3.

Lin, Yixiao et al. (2023). "Deep learning based on co-registered ultrasound and photoacoustic imaging improves the assessment of rectal cancer treatment response". In: *Biomedical Optics Express*. Vol. 14, no. 5, pp. 2015–2027. DOI: 10.1364/BOE.487647.

Lin, Yongping et al. (2019). "Computer-aided classification system for early endometrial cancer of co-registered photoacoustic and ultrasonic signals". In Proceedings of: *Optics in Health Care and Biomedical Optics IX*. Vol. 11190. International Society for Optics and Photonics, pp. 136–142. DOI: 10.1117/12.2536709.

Lindholm, Andreas, Wahlström, Niklas, Lindsten, Fredrik, and Schön, Thomas B. (2022). *Machine learning - a first course for engineers and scientists*. Cambridge University Press. URL: https://smlbook.org.

Liu, Liyuan et al. (2019). "On the variance of the adaptive learning rate and beyond". In: *arXiv preprint arXiv:1908.03265*.

Lu, Mengyang et al. (2021a). "Artifact removal in photoacoustic tomography with an unsupervised method". In: *Biomedical Optics Express*. Vol. 12, no. 10, pp. 6284–6299. DOI: 10.1364/BOE.434172.

Lu, Tong et al. (2021b). "LV-GAN: a deep learning approach for limited-view optoacoustic imaging based on hybrid datasets". In: *Journal of Biophotonics*. Vol. 14, no. 2, e202000325. DOI: 10.1002/jbio.202000325.

Luke, Geoffrey P, Hoffer-Hawlik, Kevin, Van Namen, Austin C, and Shang, Ruibo (2019). "O-Net: a convolutional neural network for quantitative photoacoustic image segmentation and oximetry". In: *arXiv preprint arXiv:1911.01935*.

Lunz, Sebastian, Hauptmann, Andreas, Tarvainen, Tanja, Schönlieb, Carola-Bibiane, and Arridge, Simon (2020). "On learned operator correction". In: *arXiv preprint arXiv:2005.07069*.

Lutzweiler, Christian, Meier, Reinhard, and Razansky, Daniel (2015). "Optoacoustic image segmentation based on signal domain analysis". In: *Photoacoustics*. Vol. 3, no. 4, pp. 151–158. DOI: 10.1016/j.pacs.2015.11.002.

Ly, Cao Duong et al. (2022). "Full-view in vivo skin and blood vessels profile segmentation in photoacoustic imaging based on deep learning". In: *Photoacoustics*. Vol. 25, p. 100310. DOI: 10.1016/j.pacs.2021.100310.

Ma, Yaxin et al. (2020). "Human breast numerical model generation based on deep learning for photoacoustic imaging". In Proceedings of: *IEEE Engineering in Medicine & Biology Society*. IEEE, pp. 1919–1922. DOI: 10.1109/EMBC44109.2020.9176298.

Ma, Yiming, Lei, Zhigang, Wu, Dongjian, Shen, Yi, and Sun, Mingjian (2023a). "A neural network estimation model based light dose control method and system for low-temperature photothermal therapy". In: *Biomedical Signal Processing and Control*. Vol. 85, p. 104935. DOI: 10.1016/j.bspc.2023.104935.

Ma, Yuanzheng et al. (2022). "Cascade neural approximating for few-shot super-resolution photoacoustic angiography". In: *Applied Physics Letters*. Vol. 121, no. 10, p. 103701. DOI: 10.1063/5.0100424.

Ma, Yuanzheng et al. (2023b). "Self-similarity-based super-resolution of photoacoustic angiography from hand-drawn doodles". In: *arXiv preprint arXiv:2305.01165*.

Madasamy, Arumugaraj, Gujrati, Vipul, Ntziachristos, Vasilis, and Prakash, Jaya (2022). "Deep learning methods hold promise for light fluence compensation in three-dimensional optoacoustic imaging". In: *Journal of Biomedical Optics*. Vol. 27, no. 10, pp. 106004–106004. DOI: 10.1117/1.JBO.27.10.106004.

Madhumathy, P and Pandey, Digvijay (2022). "Deep learning based photo acoustic imaging for non-invasive imaging". In: *Multimedia Tools and Applications*. Vol. 81, no. 5, pp. 7501–7518. DOI: 10.1007/s11042-022-11903-6.

Mai, Thi Thao et al. (2021). "In vivo quantitative vasculature segmentation and assessment for photodynamic therapy process monitoring using photoacoustic microscopy". In: *Sensors*. Vol. 21, no. 5, p. 1776. DOI: 10.3390/s21051776.

Maier-Hein, Lena, Menze, Bjoern, et al. (2022). "Metrics reloaded: pitfalls and recommendations for image analysis validation". In: *arXiv.org*, no. 2206.01653.

Maier-Hein, Lena et al. (2021). "Heidelberg colorectal data set for surgical data science in the sensor operating room". In: *Scientific Data*. Vol. 8, no. 1, p. 101. DOI: 10.1038/s41597-021-00882-2.

Mallidi, Srivalleesha, Luke, Geoffrey P, and Emelianov, Stanislav (2011). "Photoacoustic imaging in cancer detection, diagnosis, and treatment guidance". In: *Trends in Biotechnology*. Vol. 29, no. 5, pp. 213–221. DOI: 10.1016/j.tibtech.2011.01.006.

Mandal, Subhamoy, Viswanath, PS, Yeshaswini, N, Dean-Ben, X Luís, and Razansky, Daniel (2015). "Multiscale edge detection and parametric shape modeling for boundary delineation

in optoacoustic images". In Proceedings of: *IEEE Engineering in Medicine and Biology Society*. IEEE, pp. 707–710. DOI: 10.1109/EMBC.2015.7318460.

Maneas, Efthymios et al. (2023). "Enhancement of instrumented ultrasonic tracking images using deep learning". In: *International Journal of Computer Assisted Radiology and Surgery*. Vol. 18, no. 2, pp. 395–399. DOI: 10.1007/s11548-022-02728-7.

Manwar, Rayyan et al. (2020). "Deep learning protocol for improved photoacoustic brain imaging". In: *Journal of Biophotonics*. Vol. 13, no. 10, e202000212. DOI: 10.1002/jbio.202000212.

Matrone, Giulia, Savoia, Alessandro Stuart, Caliano, Giosuè, and Magenes, Giovanni (2014). "The delay multiply and sum beamforming algorithm in ultrasound B-mode medical imaging". In: *IEEE Transactions on Medical Imaging*. Vol. 34, no. 4, pp. 940–949. DOI: 10.1109/TMI.2014.2371235.

Meng, Jing et al. (2022). "Depth-extended acoustic-resolution photoacoustic microscopy based on a two-stage deep learning network". In: *Biomedical Optics Express*. Vol. 13, no. 8, pp. 4386–4397. DOI: 10.1364/BOE.461183.

Mongan, John, Moy, Linda, and Kahn Jr, Charles E (2020). *Checklist for artificial intelligence in medical imaging (CLAIM): a guide for authors and reviewers*. DOI: 10.1148/ryai.2020200029.

Moore, Michael J et al. (2019). "Photoacoustic F-mode imaging for scale specific contrast in biological systems". In: *Communications Physics*. Vol. 2, no. 1, p. 30. DOI: 10.1038/s42005-019-0131-y.

Morse, Philip M., Ingard, K. Uno, and Shankland, R. S. (May 1969). "Theoretical acoustics". In: *Physics Today*. Vol. 22, no. 5, pp. 98–99. DOI: 10.1063/1.3035602.

Moustakidis, Serafeim, Omar, Murad, Aguirre, Juan, Mohajerani, Pouyan, and Ntziachristos, Vasilis (2019). "Fully automated identification of skin morphology in raster-scan optoacoustic mesoscopy using artificial intelligence". In: *Medical Physics*. Vol. 46, no. 9, pp. 4046–4056. DOI: 10.1002/mp.13725.

Nikolov, Stanislav et al. (2021). "Clinically applicable segmentation of head and neck anatomy for radiotherapy: deep learning algorithm development and validation study". In: *Journal of Medical Internet Research*. Vol. 23, no. 7, e26151. DOI: 10.2196/26151.

Nitkunanantharajah, Suhanyaa et al. (2019). "Skin surface detection in 3D optoacoustic mesoscopy based on dynamic programming". In: *IEEE Transactions on Medical Imaging*. Vol. 39, no. 2, pp. 458–467. DOI: 10.1109/TMI.2019.2928393.

Nitkunanantharajah, Suhanyaa et al. (2020). "Three-dimensional optoacoustic imaging of nailfold capillaries in systemic sclerosis and its potential for disease differentiation using

deep learning". In: *Scientific Reports*. Vol. 10, no. 1, p. 16444. DOI: 10.1038/s41598-020-73319-2.

Nolden, Marco et al. (2013). "The Medical Imaging Interaction Toolkit: challenges and advances: 10 years of open-source development". In: *International Journal of Computer Assisted Radiology and Surgery*. Vol. 8, pp. 607–620. DOI: 10.1007/s11548-013-0840-8.

Nölke, Jan-Hinrich et al. (2021). "Invertible neural networks for uncertainty quantification in photoacoustic imaging". In Proceedings of: *Bildverarbeitung für die Medizin*. Springer, pp. 330–335. DOI: 10.1007/978-3-658-33198-6_80.

Nwoye, Chinedu Innocent et al. (2023). "CholecTriplet2022: Show me a tool and tell me the triplet — An endoscopic vision challenge for surgical action triplet detection". In: *Medical Image Analysis*. Vol. 89, p. 102888. DOI: https://doi.org/10.1016/j.media.2023.102888.

Olefir, Ivan et al. (2020). "Deep learning-based spectral unmixing for optoacoustic imaging of tissue oxygen saturation". In: *IEEE Transactions on Medical Imaging*. Vol. 39, no. 11, pp. 3643–3654. DOI: 10.1109/TMI.2020.3001750.

Oliveira, Dario Augusto Borges and Viana, Matheus Palhares (2018). "Lung nodule synthesis using CNN-based latent data representation". In Proceedings of: *Simulation and Synthesis in Medical Imaging*. Springer, pp. 111–118. DOI: 10.1007/978-3-030-00536-8_12.

Oltulu, Pembe, Ince, Bilsev, Kokbudak, Naile, Findik, Sidika, and Kilinc, Fahriye (2018). "Measurement of epidermis, dermis, and total skin thicknesses from six different body regions with a new ethical histometric technique". In: *Turkish Journal of Plastic Surgery*. Vol. 26, no. 2, pp. 56–61. DOI: 10.4103/tjps.TJPS_2_17.

Orozco, Rafael, Siahkoohi, Ali, Rizzuti, Gabrio, Leeuwen, Tristan van, and Herrmann, Felix Johan (2021). "Photoacoustic imaging with conditional priors from normalizing flows". In Proceedings of: *NeurIPS 2021 Workshop on Deep Learning and Inverse Problems*. URL: https://openreview.net/forum?id=woi1OTvR001.

Ozdemir, Firat, Lafci, Berkan, Deán-Ben, Xosé Luís, Razansky, Daniel, and Perez-Cruz, Fernando (2022). "OADAT: experimental and synthetic clinical optoacoustic data for standardized image processing". In: *arXiv preprint arXiv:2206.08612*.

Pakhomov, Daniil, Hira, Sanchit, Wagle, Narayani, Green, Kemar E, and Navab, Nassir (2021). "Segmentation in style: unsupervised semantic image segmentation with StyleGAN and clip". In: *arXiv preprint arXiv:2107.12518*.

Pan, Bolin and Betcke, Marta M (2023). "On learning the invisible in photoacoustic tomography with flat directionally sensitive detector". In: *SIAM Journal on Imaging Sciences*. Vol. 16, no. 2, pp. 770–801. DOI: doi.org/10.1137/22M148793X.

Pan, Shun-Yi, Lu, Cheng-You, Lee, Shih-Po, and Peng, Wen-Hsiao (2021). "Weakly-supervised image semantic segmentation using graph convolutional networks". In Proceedings of: *IEEE International Conference on Multimedia and Expo*. IEEE, pp. 1–6. DOI: 10.1109/ICME51207.2021.9428116.

Park, Sojeong et al. (2021). "Model learning analysis of 3D optoacoustic mesoscopy images for the classification of atopic dermatitis". In: *Biomedical Optics Express*. Vol. 12, no. 6, pp. 3671–3683. DOI: 10.1364/BOE.415105.

Park, Suhyun, Karpiouk, Andrei B, Aglyamov, Salavat R, and Emelianov, Stanislav Y (2008). "Adaptive beamforming for photoacoustic imaging". In: *Optics Letters*. Vol. 33, no. 12, pp. 1291–1293. DOI: 10.1364/OL.33.001291.

Park, Taesung, Liu, Ming-Yu, Wang, Ting-Chun, and Zhu, Jun-Yan (2019). "Semantic image synthesis with spatially-adaptive normalization". In Proceedings of: *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2337–2346. URL: https://openaccess.thecvf.com/content_CVPR_2019/papers/Park_Semantic_Image_Synthesis_With_Spatially-Adaptive_Normalization_CVPR_2019_paper.pdf.

Paszke, Adam et al. (2019). "Pytorch: an imperative style, high-performance deep learning library". In: *Advances in Neural Information Processing Systems*. Vol. 32.

Patil, Megha et al. (2021). "Evaluation of auto-encoder network with photoacoustic signal for unsupervised classification of prostate cancer". In Proceedings of: *Computer Vision and Image Processing*. Springer, pp. 420–429. DOI: 10.1007/978-981-16-1086-8_37.

Pham, Dzung L et al. (2020). "Contrast adaptive tissue classification by alternating segmentation and synthesis". In Proceedings of: *Simulation and Synthesis in Medical Imaging*. Springer, pp. 1–10. DOI: 10.1007/978-3-030-59520-3_1.

Prahl, Scott and Jacques, Steven L (2017). *Optical properties*. URL: https://omlc.org/classroom/ece532/class3/index.html (visited on 08/15/2023).

Pulkkinen, Aki, Cox, Bejamin T, Arridge, Simon R, Kaipio, Jari P, and Tarvainen, Tanja (2014). "A Bayesian approach to spectral quantitative photoacoustic tomography". In: *Inverse Problems*. Vol. 30, no. 6, p. 065012. DOI: 10.1088/0266-5611/30/6/065012.

Radford, Alec, Metz, Luke, and Chintala, Soumith (2015). "Unsupervised representation learning with deep convolutional generative adversarial networks". In: *arXiv preprint arXiv:1511.06434*.

Rajendran, Praveenbalaji and Pramanik, Manojit (2020). "Deep learning approach to improve tangential resolution in photoacoustic tomography". In: *Biomedical Optics Express*. Vol. 11, no. 12, pp. 7311–7323. DOI: 10.1364/BOE.410145.

Rajendran, Praveenbalaji and Pramanik, Manojit (2021). "Deep-learning-based multi-transducer photoacoustic tomography imaging without radius calibration". In: *Optics Letters*. Vol. 46, no. 18, pp. 4510–4513. DOI: 10.1364/OL.434513.

— (2022). "High frame rate (∼3 Hz) circular photoacoustic tomography using single-element ultrasound transducer aided with deep learning". In: *Journal of Biomedical Optics*. Vol. 27, no. 6, pp. 066005–066005. DOI: 10.1117/1.JBO.27.6.066005.

— (2023). "Deep learning based high frame rate photoacoustic tomography". In Proceedings of: *Photons Plus Ultrasound: Imaging and Sensing*. Vol. 12379. International Society for Optics and Photonics, pp. 271–277. DOI: 10.1117/12.2648035.

Ramesh, Aditya, Dhariwal, Prafulla, Nichol, Alex, Chu, Casey, and Chen, Mark (2023). *DALL-E 2*. URL: https://openai.com/dall-e-2 (visited on 09/07/2023).

Ramos-Vega, Marta et al. (2022). "Mapping of neuroinflammation-induced hypoxia in the spinal cord using optoacoustic imaging". In: *Acta Neuropathologica Communications*. Vol. 10, no. 1, pp. 1–13. DOI: 10.1186/s40478-022-01337-4.

Raumonen, Pasi and Tarvainen, Tanja (2018). "Segmentation of vessel structures from photoacoustic images with reliability assessment". In: *Biomedical Optics Express*. Vol. 9, no. 7, pp. 2887–2904. DOI: 10.1364/BOE.9.002887.

Refaee, Amir, Kelly, Corey J, Moradi, Hamid, and Salcudean, Septimiu E (2021). "Denoising of pre-beamformed photoacoustic data using generative adversarial networks". In: *Biomedical Optics Express*. Vol. 12, no. 10, pp. 6184–6204. DOI: 10.1364/BOE.431997.

Reiter, Austin and Bell, Muyinatu A Lediju (2017). "A machine learning approach to identifying point source locations in photoacoustic data". In Proceedings of: *Photons Plus Ultrasound: Imaging and Sensing*. Vol. 10064. International Society for Optics and Photonics, pp. 504–509. DOI: 10.1117/12.2255098.

Ren, Wuwei, Deán-Ben, Xosé Luís, Augath, Mark-Aurel, and Razansky, Daniel (2021a). "Feasibility study on concurrent optoacoustic tomography and magnetic resonance imaging". In Proceedings of: *Photons Plus Ultrasound: Imaging and Sensing*. Vol. 11642. International Society for Optics and Photonics, pp. 14–18. DOI: 10.1117/12.2577519.

Ren, Zhong, Liu, Tao, and Liu, Guodong (2021b). "Classification and discrimination of real and fake blood based on photoacoustic spectroscopy combined with particle swarm optimized wavelet neural networks". In: *Photoacoustics*. Vol. 23, p. 100278. DOI: 10.1016/j.pacs.2021.100278.

Ren, Zhongzheng et al. (2020). "UFO 2: a unified framework towards omni-supervised object detection". In Proceedings of: *European Conference on Computer Vision*. Springer, pp. 288–313. DOI: 10.1007/978-3-030-58529-7_18.

Rix, Tom et al. (2022). "Deep learning for spectral image synthesis". In Proceedings of: *Multimodal Biomedical Imaging XVII*. International Society for Optics and Photonics, PC119520I. DOI: 10.1117/12.2608622.

Rix, Tom et al. (2023). "Efficient photoacoustic image synthesis with deep learning". In: *Sensors*. Vol. 23, no. 16, p. 7085. DOI: 10.3390/s23167085.

Rombach, Robin, Blattmann, Andreas, Lorenz, Dominik, Esser, Patrick, and Ommer, Björn (2022). "High-resolution image synthesis with latent diffusion models". In Proceedings of: *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695. URL: https://openaccess.thecvf.com/content/CVPR2022/papers/Rombach_High-Resolution_Image_Synthesis_With_Latent_Diffusion_Models_CVPR_2022_paper.pdf.

Ronneberger, Olaf, Fischer, Philipp, and Brox, Thomas (2015). "U-Net: convolutional networks for biomedical image segmentation". In Proceedings of: *Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 234–241. DOI: 10.1007/978-3-319-24574-4_28.

Roß, Tobias et al. (2023). "Beyond rankings: learning (more) from algorithm validation". In: *Medical Image Analysis*. Vol. 86, p. 102765. DOI: 10.1016/j.media.2023.102765.

Rusak, Filip et al. (2020). "3D brain MRI GAN-based synthesis conditioned on partial volume maps". In Proceedings of: *Simulation and Synthesis in Medical Imaging*. Springer, pp. 11–20. DOI: 10.1007/978-3-030-59520-3_2.

Sahlström, Teemu and Tarvainen, Tanja (2023a). "Utilizing variational autoencoders in photoacoustic tomography". In Proceedings of: *Photons Plus Ultrasound: Imaging and Sensing*. Vol. 12379. International Society for Optics and Photonics, pp. 233–241. DOI: 10.1117/12.2644801.

— (2023b). "Utilizing variational autoencoders in the Bayesian inverse problem of photoacoustic tomography". In: *SIAM Journal on Imaging Sciences*. Vol. 16, no. 1, pp. 89–110. DOI: 10.1137/22M1489897.

Schellenberg, Melanie et al. (2021). "Generation of training data for quantitative photoacoustic imaging". In Proceedings of: *Photons Plus Ultrasound: Imaging and Sensing*. Vol. 11642. International Society for Optics and Photonics, 116421J. DOI: 10.1117/12.2578180.

Schellenberg, Melanie et al. (2022a). "Photoacoustic image synthesis with generative adversarial networks". In: *Photoacoustics*. Vol. 28, p. 100402. DOI: 10.1016/j.pacs.2022.100402.

Schellenberg, Melanie et al. (2022b). "Semantic segmentation of multispectral photoacoustic images using deep learning". In: *Photoacoustics*. Vol. 26, p. 100341. DOI: 10.1016/j.pacs.2022.100341.

Schlereth, Maja et al. (2022). "Automatic classification of neuromuscular diseases in children using photoacoustic imaging". In Proceedings of: *Bildverarbeitung für die Medizin*. Springer, pp. 285–290. DOI: 10.1007/978-3-658-36932-3_60.

Schober, Otmar, Kiessling, Fabian, and Debus, Jürgen (2020). *Molecular imaging in oncology*. Vol. 216. Springer Nature. DOI: 10.3322/caac.21713.

Schwab, Johannes, Antholzer, Stephan, and Haltmeier, Markus (2019a). "Learned backprojection for sparse and limited view photoacoustic tomography". In Proceedings of: *Photons Plus Ultrasound: Imaging and Sensing*. Vol. 10878. International Society for Optics and Photonics, pp. 263–271. DOI: 10.1117/12.2508438.

Schwab, Johannes, Antholzer, Stephan, Nuster, Robert, and Haltmeier, Markus (2018). "Real-time photoacoustic projection imaging using deep learning". In: *arXiv preprint arXiv:1801.06693*.

Schwab, Johannes, Antholzer, Stephan, Nuster, Robert, Paltauf, Günther, and Haltmeier, Markus (2019b). "Deep learning of truncated singular values for limited view photoacoustic tomography". In Proceedings of: *Photons Plus Ultrasound: Imaging and Sensing*. Vol. 10878. International Society for Optics and Photonics, pp. 254–262. DOI: 10.1117/12.2508418.

Schweiger, Martin and Arridge, Simon (2014). "The Toast++ software suite for forward and inverse modeling in optical tomography". In: *Journal of Biomedical Optics*. Vol. 19, no. 4, pp. 040801–040801. DOI: 10.1117/1.JBO.19.4.040801.

Segars, W Paul, Sturgeon, G, Mendonca, S, Grimes, Jason, and Tsui, Benjamin MW (2010). "4D XCAT phantom for multimodality imaging research". In: *Medical Physics*. Vol. 37, no. 9, pp. 4902–4915. DOI: 10.1118/1.3480985.

Seong, Daewoon et al. (2023). "Three-dimensional reconstructing undersampled photoacoustic microscopy images using deep learning". In: *Photoacoustics*. Vol. 29, p. 100429. DOI: 10.1016/j.pacs.2022.100429.

Shahid, Husnain, Khalid, Adnan, Liu, Xin, Irfan, Muhammad, and Ta, Dean (2021a). "A deep learning approach for the photoacoustic tomography recovery from undersampled measurements". In: *Frontiers in Neuroscience*. Vol. 15, p. 598693. DOI: 10.3389/fnins.2021.598693.

Shahid, Husnain, Khalid, Adnan, Yue, Yaoting, Liu, Xin, and Ta, Dean (2022). "Feasibility of a generative adversarial network for artifact removal in experimental photoacoustic imaging". In: *Ultrasound in Medicine & Biology*. Vol. 48, no. 8, pp. 1628–1643. DOI: 10.1016/j.ultrasmedbio.2022.04.008.

Shahid, Husnain, Yue, Yaoting, Khalid, Adnan, Liu, Xin, and Ta, Dean (2021b). "Batch renormalization accumulated residual U-Network for artifacts removal in photoacoustic imaging". In Proceedings of: *IEEE International Ultrasonics Symposium*. IEEE, pp. 1–4.

Shan, Hongming, Wang, Ge, and Yang, Yang (2019a). "Accelerated correction of reflection artifacts by deep neural networks in photo-acoustic tomography". In: *Applied Sciences*. Vol. 9, no. 13, p. 2615. DOI: 10.3390/app9132615.

Shan, Hongming, Wiedeman, Christopher, Wang, Ge, and Yang, Yang (2019b). "Simultaneous reconstruction of the initial pressure and sound speed in photoacoustic tomography using a deep-learning approach". In Proceedings of: *Novel Optical Systems, Methods, and Applications XXII*. Vol. 11105. International Society for Optics and Photonics, pp. 18–27. DOI: 10.1117/12.2529984.

Shao, Peng, Cox, Ben, and Zemp, Roger J (2011). "Estimating optical absorption, scattering, and Grueneisen distributions with multiple-illumination photoacoustic tomography". In: *Applied Optics*. Vol. 50, no. 19, pp. 3145–3154. DOI: 10.1364/AO.50.003145.

Sharma, Arunima and Pramanik, Manojit (2020). "Convolutional neural network for resolution enhancement and noise reduction in acoustic resolution photoacoustic microscopy". In: *Biomedical Optics Express*. Vol. 11, no. 12, pp. 6826–6839. DOI: 10.1364/BOE.411257.

Shen, Kang and Tian, Chao (2021). "Deep filtered back projection for photoacoustic image reconstruction". In Proceedings of: *Asia Communications and Photonics Conference*. Optica Publishing Group, M4G–3. DOI: 10.1364/ACPC.2021.M4G.3.

Shi, Mengjie et al. (2022). "Improving needle visibility in LED-based photoacoustic imaging using deep learning with semi-synthetic datasets". In: *Photoacoustics*. Vol. 26, p. 100351. DOI: 10.1016/j.pacs.2022.100351.

Shi, Yunsheng et al. (2020). "Masked label prediction: unified message passing model for semi-supervised classification". In: *arXiv preprint arXiv:2009.03509*.

Shin, Hoo-Chang et al. (2018). "Medical image synthesis for data augmentation and anonymization using generative adversarial networks". In Proceedings of: *Simulation and Synthesis in Medical Imaging*. Springer, pp. 1–11. DOI: 10.1007/978-3-030-00536-8_1.

Singh, Mithun Kuniyil Ajith et al. (2020). "Deep learning-enhanced LED-based photoacoustic imaging". In Proceedings of: *Photons Plus Ultrasound: Imaging and Sensing*. Vol. 11240. International Society for Optics and Photonics, pp. 161–166. DOI: 0.1117/12.2545654.

Singh, Suriya et al. (2018). "Self-supervised feature learning for semantic segmentation of overhead imagery." In Proceedings of: *British Machine Vision Conference*. Vol. 1. no. 2, p. 4. URL: http://bmvc2018.org/contents/papers/0345.pdf.

Song, Xianlin, Tang, Kanggao, Wei, Jianshuang, and Song, Lingfang (2021). "Deep-learning denoising convolutional neural network for photoacoustic microscopy". In Proceedings of: *Optics Young Scientist Summit*. Vol. 11781. International Society for Optics and Photonics, pp. 140–145. DOI: 10.1117/12.2591380.

Song, Xianlin et al. (2022). "Improvement of spatial resolution of photoacoustic microscopy based on physical model and deep learning". In Proceedings of: *Real-Time Image Processing and Deep Learning*. Vol. 12102. International Society for Optics and Photonics, pp. 199–203. DOI: 10.1117/12.2636267.

Song, Ziran, Fu, Yuting, Qu, Jiaqi, Liu, Qi, and Wei, Xunbin (2020). "Application of convolutional neural network in signal classification for in vivo photoacoustic flow cytometry". In Proceedings of: *Optics in Health Care and Biomedical Optics X*. Vol. 11553. International Society for Optics and Photonics, pp. 363–369. DOI: 10.1117/12.2576784.

Sowmiya, C and Thittai, Arun K (2017). "Simulation of photoacoustic tomography (PAT) system in COMSOL (R) and comparison of two popular reconstruction techniques". In Proceedings of: *Medical Imaging 2017: Biomedical Applications in Molecular, Structural, and Functional Imaging*. Vol. 10137. International Society for Optics and Photonics, pp. 435–445. DOI: 10.1117/12.2254450.

Standring, Susan (2021). *Gray's anatomy e-book: the anatomical basis of clinical practice*. Elsevier Health Sciences.

Sun, Ming-Jian, Li, Wei-Xiang, Liu, Zi-Chao, and Liu, Guang-Xing (2021a). "Tumor photoacoustic image reconstruction method based on deep learning". In Proceedings of: *IEEE International Ultrasonics Symposium*. IEEE, pp. 1–4. DOI: 10.1109/IUS52206.2021.9593677.

Sun, Mingjian et al. (2020). "Full three-dimensional segmentation and quantification of tumor vessels for photoacoustic images". In: *Photoacoustics*. Vol. 20, p. 100212. DOI: 10.1016/j.pacs.2020.100212.

Sun, Zheng, Wang, Xinyu, and Yan, Xiangyang (2021b). "An iterative gradient convolutional neural network and its application in endoscopic photoacoustic image formation from incomplete acoustic measurement". In: *Neural Computing and Applications*. Vol. 33, pp. 8555–8574. DOI: 10.1007/s00521-020-05607-x.

Susmelj, Anna Klimovskaia et al. (2022). "Signal domain learning approach for optoacoustic image reconstruction from limited view data". In Proceedings of: *International Conference on Medical Imaging with Deep Learning*. PMLR, pp. 1173–1191. URL: https://proceedings.mlr.press/v172/susmelj22a/susmelj22a.pdf.

Sweeney, Paul W et al. (2023). "Segmentation of 3D blood vessel networks using unsupervised deep learning". In: *bioRxiv*, pp. 2023–04. DOI: 10.1101/2023.04.30.538453.

Tang, Yuqi et al. (2023). "High-fidelity deep functional photoacoustic tomography enhanced by virtual point sources". In: *Photoacoustics*. Vol. 29, p. 100450. DOI: 10.1016/j.pacs.2023.100450.

Tong, Tong et al. (2020). "Domain transform network for photoacoustic tomography from limited-view and sparsely sampled data". In: *Photoacoustics*. Vol. 19, p. 100190. DOI: 10.1016/j.pacs.2020.100190.

Treeby, Bradley E and Cox, Benjamin T (2010). "k-Wave: MATLAB toolbox for the simulation and reconstruction of photoacoustic wave fields". In: *Journal of Biomedical Optics*. Vol. 15, no. 2, pp. 021314–021314. DOI: 10.1117/1.3360308.

Tserevelakis, George J et al. (2022). "Hybrid confocal fluorescence and photoacoustic microscopy for the label-free investigation of melanin accumulation in fish scales". In: *Scientific Reports*. Vol. 12, no. 1, p. 7173. DOI: 10.1038/s41598-022-11262-0.

Tserevelakis, George J et al. (2023). "Deep learning-assisted frequency-domain photoacoustic microscopy". In: *Optics Letters*. Vol. 48, no. 10, pp. 2720–2723. DOI: 10.1364/OL.486624.

Tudosiu, Petru-Daniel, Graham, Mark S, Vercauteren, Tom, and Cardoso, M Jorge (2022). "Can segmentation models be trained with fully synthetically generated data?" In Proceedings of: *Simulation and Synthesis in Medical Imaging*. Vol. 13570. Springer Nature, p. 79. DOI: 10.1007/978-3-031-16980-9_8.

Unberath, Mathias et al. (2018). "DeepDRR–a catalyst for machine learning in fluoroscopy-guided procedures". In Proceedings of: *Medical Image Computing and Computer Assisted Intervention*. Springer, pp. 98–106. DOI: 10.1007/978-3-030-00937-3_12.

Vaswani, Ashish et al. (2017). "Attention is all you need". In: *Advances in Neural Information Processing Systems*. Vol. 30. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Vera, Matias, Gonzalez, Martin G, and Vega, L Rey (2023). "Invariant representations in deep learning for optoacoustic imaging". In: *Review of Scientific Instruments*. Vol. 94, no. 5. DOI: 10.1063/5.0139286.

Vieten, Patricia et al. (2022). "Deep learning-based semantic segmentation of clinically relevant tissue structures leveraging multispectral photoacoustic images". In Proceedings of: *Photons Plus Ultrasound: Imaging and Sensing*. International Society for Optics and Photonics, PC11960OP. DOI: 10.1117/12.2608616.

Vousten, Vincent, Moradi, Hamid, Wu, Zijian, Boctor, Emad M, and Salcudean, Septimiu E (2023). "Laser diode photoacoustic point source detection: machine learning-based denoising and reconstruction". In: *Optics Express*. Vol. 31, no. 9, pp. 13895–13910. DOI: 10.1364/OE.483892.

Vu, Tri, Li, Mucong, Humayun, Hannah, Zhou, Yuan, and Yao, Junjie (2020). "A generative adversarial network for artifact removal in photoacoustic computed tomography with a

linear-array transducer". In: *Experimental Biology and Medicine*. Vol. 245, no. 7, pp. 597–605. DOI: 10.1177/1535370220914285.

Vu, Tri et al. (2021). "Deep image prior for undersampling high-speed photoacoustic microscopy". In: *Photoacoustics*. Vol. 22, p. 100266. DOI: 10.1016/j.pacs.2021.100266.

Waibel, Dominik et al. (2018). "Reconstruction of initial pressure from limited view photoacoustic images using deep learning". In Proceedings of: *Photons Plus Ultrasound: Imaging and Sensing*. Vol. 10494. International Society for Optics and Photonics, pp. 196–203. DOI: doi.org/10.1117/12.2288353.

Wang, Lihong V and Wu, Hsin-i (2012). *Biomedical optics: principles and imaging*. John Wiley & Sons. DOI: 10.1002/9780470177013.

Wang, Risheng et al. (2022a). "Medical image segmentation using deep learning: a survey". In: *IET Image Processing*. Vol. 16, no. 5, pp. 1243–1267. DOI: 10.1049/ipr2.12419.

Wang, Ting-Chun et al. (2018). "High-resolution image synthesis and semantic manipulation with conditional GANs". In Proceedings of: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8798–8807. URL: https://openaccess.thecvf.com/content_cvpr_2018/papers/Wang_High-Resolution_Image_Synthesis_CVPR_2018_paper.pdf.

Wang, Tong, He, Menghui, Shen, Kang, Liu, Wen, and Tian, Chao (2022b). "Learned regularization for image reconstruction in sparse-view photoacoustic tomography". In: *Biomedical Optics Express*. Vol. 13, no. 11, pp. 5721–5737. DOI: 10.1364/BOE.469460.

Wang, Zhou, Bovik, Alan C, Sheikh, Hamid R, and Simoncelli, Eero P (2004). "Image quality assessment: from error visibility to structural similarity". In: *IEEE Transactions on Image Processing*. Vol. 13, no. 4, pp. 600–612. DOI: 10.1109/TIP.2003.819861.

Wang, Zhuangzhuang, Li, Sihang, and Song, Xianlin (2021). "Super-resolution photoacoustic microscopy based on deep learning". In Proceedings of: *Real-Time Image Processing and Deep Learning 2021*. Vol. 11736. International Society for Optics and Photonics, pp. 46–52. DOI: 10.1117/12.2589655.

Warrier, Gayathry Sobhanan et al. (2022). "Automated recognition of cancer tissues through deep learning framework from the photoacoustic specimen". In: *Contrast Media & Molecular Imaging*. Vol. 2022. DOI: 10.1155/2022/4356744.

Wiesenfarth, Manuel et al. (2021). "Methods and open-source toolkit for analyzing and visualizing challenge results". In: *Scientific Reports*. Vol. 11, no. 1, p. 2369. DOI: 10.1038/s41598-021-82017-6.

Winzeck, Stefan et al. (2018). "ISLES 2016 and 2017-benchmarking ischemic stroke lesion outcome prediction based on multispectral MRI". In: *Frontiers in Neurology*. Vol. 9, p. 679. DOI: 10.3389/fneur.2018.00679.

Wu, Xun, Sanders, Jean, Dundar, Murat, and Oralkan, Ömer (2017). "Multi-wavelength photoacoustic imaging for monitoring lesion formation during high-intensity focused ultrasound therapy". In Proceedings of: *Photons Plus Ultrasound: Imaging and Sensing*. Vol. 10064. International Society for Optics and Photonics, pp. 633–639. DOI: 10.1117/12.2248739.

Wu, Zonghan et al. (2020). "A comprehensive survey on graph neural networks". In: *IEEE Transactions on Neural Networks and Learning Systems*. Vol. 32, no. 1, pp. 4–24. DOI: 10.1109/TNNLS.2020.2978386.

Xiao, Jiaying, Jiang, Jinsheng, Zhang, Jiaxi, Wang, Yongjun, and Wang, Bo (2022). "Acoustic-resolution-based spectroscopic photoacoustic endoscopy towards molecular imaging in deep tissues". In: *Optics Express*. Vol. 30, no. 19, pp. 35014–35028. DOI: 10.1364/OE.469550.

Xu, Minghua and Wang, Lihong V (2005). "Universal back-projection algorithm for photoacoustic computed tomography". In: *Physical Review E*. Vol. 71, no. 1, p. 016706. DOI: 10.1103/PhysRevE.71.016706.

Xu, Yuan, Feng, Dazi, and Wang, Lihong V (2002). "Exact frequency-domain reconstruction for thermoacoustic tomography. I. Planar geometry". In: *IEEE Transactions on Medical Imaging*. Vol. 21, no. 7, pp. 823–828. DOI: 10.1109/TMI.2002.801172.

Xu, Ziyue et al. (2019). "Tunable CT lung nodule synthesis conditioned on background image and semantic features". In Proceedings of: *Simulation and Synthesis in Medical Imaging*. Springer, pp. 62–70. DOI: 10.1007/978-3-030-32778-1_7.

Yamlahi, Amine et al. (2023). "Self-distillation for surgical action recognition". In: *arXiv preprint arXiv:2303.12915*.

Yang, Changchun and Gao, Fei (2019a). "EDA-Net: dense aggregation of deep and shallow information achieves quantitative photoacoustic blood oxygenation imaging deep in human breast". In Proceedings of: *Medical Image Computing and Computer Assisted Intervention*. Springer, pp. 246–254. DOI: 10.1007/978-3-030-32239-7_28.

Yang, Changchun, Lan, Hengrong, and Gao, Fei (2019b). "Accelerated photoacoustic tomography reconstruction via recurrent inference machines". In Proceedings of: *IEEE Engineering in Medicine and Biology Society*. IEEE, pp. 6371–6374. DOI: 10.1109/EMBC.2019.8856290.

Yang, Changchun, Lan, Hengrong, Zhong, Hongtao, and Gao, Fei (2019c). "Quantitative photoacoustic blood oxygenation imaging using deep residual and recurrent neural network". In

Proceedings of: *IEEE International Symposium on Biomedical Imaging*. IEEE, pp. 741–744. DOI: 10.1109/ISBI.2019.8759438.

Yang, Guang and Chung, Kevin C (2018). "Ulnar artery to superficial arch bypass with a vein graft". In: *Operative Techniques: Hand and Wrist Surgery*, pp. 732–737.

Yang, Zhenyuan et al. (2014). "Multi-parametric quantitative microvascular imaging with optical-resolution photoacoustic microscopy in vivo". In: *Optics Express*. Vol. 22, no. 2, pp. 1500–1511. DOI: 10.1364/OE.22.001500.

Yazdani, Amirsaeed, Agrawal, Sumit, Johnstonbaugh, Kerrick, Kothapalli, Sri-Rajasekhar, and Monga, Vishal (2021). "Simultaneous denoising and localization network for photoacoustic target localization". In: *IEEE Transactions on Medical Imaging*. Vol. 40, no. 9, pp. 2367–2379. DOI: doi={10.1109/TMI.2021.3077187}.

Yi, Xin, Walia, Ekta, and Babyn, Paul (2019). "Generative adversarial network in medical imaging: a review". In: *Medical Image Analysis*. Vol. 58, p. 101552. DOI: 10.1016/j.media.2019.101552.

Yu, Ziming, Tang, Kanggao, and Song, Xianlin (2021). "Denoising method for image quality improvement in photoacoustic microscopy using deep learning". In Proceedings of: *Computational Optics*. Vol. 11875. International Society for Optics and Photonics, pp. 31–35. DOI: 10.1117/12.2600759.

Yuan, Alan Yilun et al. (2020). "Hybrid deep learning network for vascular segmentation in photoacoustic imaging". In: *Biomedical Optics Express*. Vol. 11, no. 11, pp. 6445–6457. DOI: 10.1364/BOE.409246.

Yue, Tong et al. (2022). "Double speed-of-sound photoacoustic image reconstruction at 10 frames-per-second with automatic segmentation". In Proceedings of: *Optics in Health Care and Biomedical Optics XII*. Vol. 12320. International Society for Optics and Photonics, pp. 164–171. DOI: 10.1117/12.2651263.

Zhang, Fan et al. (2023a). "Photoacoustic digital brain and deep-learning-assisted image reconstruction". In: *Photoacoustics*, p. 100517. DOI: 10.1016/j.pacs.2023.100517.

Zhang, Huijuan et al. (2020). "A new deep learning network for mitigating limited-view and under-sampling artifacts in ring-shaped photoacoustic tomography". In: *Computerized Medical Imaging and Graphics*. Vol. 84, p. 101720. DOI: 10.1016/j.compmedimag.2020.101720.

Zhang, Huijuan et al. (2021a). "Deep-E: a fully-dense neural network for improving the elevation resolution in linear-array-based photoacoustic tomography". In: *IEEE Transactions on Medical Imaging*. Vol. 41, no. 5, pp. 1279–1288. DOI: 10.1109/TMI.2021.3137060.

Zhang, Jiadong et al. (2021b). "Limited-view photoacoustic imaging reconstruction with dual domain inputs based on mutual inforamtion". In Proceedings of: *IEEE International Symposium on Biomedical Imaging*. IEEE, pp. 1522–1526. DOI: 10.1109/ISBI48211.2021.9433949.

Zhang, Jiayao, Chen, Bin, Zhou, Meng, Lan, Hengrong, and Gao, Fei (2018a). "Photoacoustic image classification and segmentation of breast cancer: a feasibility study". In: *IEEE Access*. Vol. 7, pp. 5457–5466. DOI: 10.1109/ACCESS.2018.2888910.

Zhang, Jiayao et al. (2018b). "Pathology study for blood vessel of ocular fundus images by photoacoustic tomography". In Proceedings of: *IEEE International Ultrasonics Symposium*. IEEE, pp. 1–4. DOI: 10.1109/ULTSYM.2018.8579931.

Zhang, Jingke, He, Qiong, Wang, Congzhi, Liao, Hongen, and Luo, Jianwen (2021c). "A general framework for inverse problem solving using self-supervised deep learning: validations in ultrasound and photoacoustic image reconstruction". In Proceedings of: *IEEE International Ultrasonics Symposium*. IEEE, pp. 1–4. DOI: 10.1109/IUS52206.2021.9593902.

Zhang, Juze, Zhang, Jingyan, Ge, Peng, and Gao, Fei (2022a). "PAFormer: photoacoustic reconstruction via transformer with mask mechanism". In Proceedings of: *IEEE International Ultrasonics Symposium*. IEEE, pp. 1–4. DOI: 10.1109/IUS54386.2022.9957348.

Zhang, Xiaoman et al. (2021d). "Photoacoustic blood pressure recognition based on deep learning". In Proceedings of: *Optics in Health Care and Biomedical Optics XI*. Vol. 11900. International Society for Optics and Photonics, pp. 424–433. DOI: 10.1117/12.2602784.

Zhang, Xueting et al. (2022b). "Sparse-sampling photoacoustic computed tomography: deep learning vs. compressed sensing". In: *Biomedical Signal Processing and Control*. Vol. 71, p. 103233. DOI: 10.1016/j.bspc.2021.103233.

Zhang, Yuxuan et al. (2021e). "DatasetGAN: efficient labeled data factory with minimal human effort". In Proceedings of: *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10145–10155. URL: https://openaccess.thecvf.com/content/CVPR2021/papers/Zhang_DatasetGAN_Efficient_Labeled_Data_Factory_With_Minimal_Human_Effort_CVPR_2021_paper.pdf.

Zhang, Zhengyuan, Jin, Haoran, Zheng, Zesheng, Luo, Yunqi, and Zheng, Yuanjin (2021f). "Photoacoustic microscopy imaging from acoustic resolution to optical resolution enhancement with deep learning". In Proceedings of: *IEEE International Symposium on Circuits and Systems*. IEEE, pp. 1–5. DOI: 10.1109/ISCAS51556.2021.9401797.

Zhang, Zhengyuan, Jin, Haoran, Zheng, Zesheng, and Zheng, Yuanjin (2022c). "Learning-based algorithm for real imaging system enhancement: acoustic resolution to optical resolution

photoacoustic microscopy". In Proceedings of: *IEEE International Symposium on Circuits and Systems*. IEEE, pp. 2458–2462. DOI: 10.1109/ISCAS48785.2022.9937914.

Zhang, Zhengyuan, Jin, Haoran, Zheng, Zesheng, and Zheng, Yuanjin (2022d). "Super acoustic resolution photoacoustic microscopy imaging enhancement". In Proceedings of: *IEEE Biomedical Circuits and Systems Conference*. IEEE, pp. 208–212. DOI: 10.1109/BioCAS54905.2022.9948686.

Zhang, Zhengyuan et al. (2022e). "Deep and domain transfer learning aided photoacoustic microscopy: acoustic resolution to optical resolution". In: *IEEE Transactions on Medical Imaging*. Vol. 41, no. 12, pp. 3636–3648. DOI: 10.1109/TMI.2022.3192072.

Zhang, Zhengyuan et al. (2023b). "Adaptive enhancement of acoustic resolution photoacoustic microscopy imaging via deep CNN prior". In: *Photoacoustics*. Vol. 30, p. 100484. DOI: 10.1016/j.pacs.2023.100484.

Zhao, Huangxuan et al. (2020). "A new deep learning method for image deblurring in optical microscopic systems". In: *Journal of Biophotonics*. Vol. 13, no. 3, e201960147. DOI: 10.1002/jbio.201960147.

Zhao, Huangxuan et al. (2021). "Deep learning enables superior photoacoustic imaging at ultralow laser dosages". In: *Advanced Science*. Vol. 8, no. 3, p. 2003097. DOI: 10.1002/advs.202003097.

Zhao, Huangxuan et al. (2022). "Deep learning-based optical-resolution photoacoustic microscopy for in vivo 3D microvasculature imaging and segmentation". In: *Advanced Intelligent Systems*. Vol. 4, no. 9, p. 2200004. DOI: 10.1002/aisy.202200004.

Zhao, Huangxuan et al. (n.d.). "Hm-3dce-Net for superior 3d photoacoustic imaging enhancement and segmentation". In: *Available at SSRN 3948474* ().

Zhao, Tianrui, Shi, Mengjie, Ourselin, Sébastien, Vercauteren, Tom, and Xia, Wenfeng (2023). "Deep learning boosts the imaging speed of photoacoustic endomicroscopy". In Proceedings of: *Photons Plus Ultrasound: Imaging and Sensing*. Vol. 12379. International Society for Optics and Photonics, pp. 102–106. DOI: 10.1117/12.2649088.

Zheng, Sun, Jiejie, Du, Yue, Yao, Qi, Meng, and Huifeng, Sun (2022a). "A deep learning method for motion artifact correction in intravascular photoacoustic image sequence". In: *IEEE Transactions on Medical Imaging*. Vol. 42, no. 1, pp. 66–78. DOI: 10.1109/TMI.2022.3202910.

Zheng, Sun, Meng, Qi, and Wang, Xin-Yu (2022b). "Quantitative endoscopic photoacoustic tomography using a convolutional neural network". In: *Applied Optics*. Vol. 61, no. 10, pp. 2574–2581. DOI: 10.1364/AO.441250.

Zheng, Wenhan et al. (2022c). "Deep-E enhanced photoacoustic tomography using three-dimensional reconstruction for high-quality vascular imaging". In: *Sensors*. Vol. 22, no. 20, p. 7725. DOI: 10.3390/s22207725.

Zhou, Jiasheng et al. (2021a). "Photoacoustic microscopy with sparse data by convolutional neural networks". In: *Photoacoustics*. Vol. 22, p. 100242. DOI: 10.1016/j.pacs.2021.100242.

Zhou, Jie et al. (2020). "Graph neural networks: a review of methods and applications". In: *AI open*. Vol. 1, pp. 57–81. DOI: 10.1016/j.aiopen.2021.01.001.

Zhou, Xue et al. (2019). "Analysis of photoacoustic signals of hyperosteogeny and osteoporosis". In Proceedings of: *Photons Plus Ultrasound: Imaging and Sensing*. Vol. 10878. International Society for Optics and Photonics, pp. 421–427. DOI: 10.1117/12.2506671.

Zhou, Yifeng, Sun, Naidi, and Hu, Song (2022). "Deep learning-powered Bessel-beam multi-parametric photoacoustic microscopy". In: *IEEE Transactions on Medical Imaging*. Vol. 41, no. 12, pp. 3544–3551. DOI: 10.1109/TMI.2022.3188739.

Zhou, Yifeng, Zhong, Fenghe, and Hu, Song (2021b). "Temporal and spectral unmixing of photoacoustic signals by deep learning". In: *Optics Letters*. Vol. 46, no. 11, pp. 2690–2693. DOI: 10.1364/OL.426678.

Zhu, Guangming et al. (2022a). "Scene graph generation: a comprehensive survey". In: *arXiv preprint arXiv:2201.00443*.

Zhu, Jun-Yan, Park, Taesung, Isola, Phillip, and Efros, Alexei A (2017a). "Unpaired image-to-image translation using cycle-consistent adversarial networks". In Proceedings of: *IEEE International Conference on Computer Vision*, pp. 2223–2232. DOI: 10.1109/ICCV.2017.244.

Zhu, Jun-Yan et al. (2017b). "Toward multimodal image-to-image translation". In: *Advances in Neural Information Processing Systems*. Vol. 30. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/819f46e52c25763a55cc642422644317-Paper.pdf.

Zhu, Shizhan, Urtasun, Raquel, Fidler, Sanja, Lin, Dahua, and Change Loy, Chen (2017c). "Be your own Prada: fashion synthesis with structural coherence". In Proceedings of: *IEEE International Conference on Computer Vision*, pp. 1680–1688. URL: https://openaccess.thecvf.com/content_ICCV_2017/papers/Zhu_Be_Your_Own_ICCV_2017_paper.pdf.

Zhu, Song-Chun, Mumford, David, et al. (2007). "A stochastic grammar of images". In: *Foundations and Trends® in Computer Graphics and Vision*. Vol. 2, no. 4, pp. 259–362. DOI: 10.1561/060000001.

Zhu, Tianlei, Chen, Junqi, Zhu, Renzhe, and Gupta, Gaurav (2023). "StyleGAN3: generative networks for improving the equivariance of translation and rotation". In: *arXiv preprint arXiv:2307.03898*.

Zhu, Xiaoyi et al. (2022b). "Real-time whole-brain imaging of hemodynamics and oxygenation at micro-vessel resolution with ultrafast wide-field photoacoustic microscopy". In: *Light: Science & Applications*. Vol. 11, no. 1, p. 138. DOI: 10.1038/s41377-022-00836-2.

Zou, Yun, Amidi, Eghbal, Luo, Hongbo, and Zhu, Quing (2022). "Ultrasound-enhanced Unet model for quantitative photoacoustic tomography of ovarian lesions". In: *Photoacoustics*. Vol. 28, p. 100420. DOI: 10.1016/j.pacs.2022.100420.

Zou, Yun, Lin, Yixiao, Kou, Sitai, and Zhu, Quing (2023). "Machine learning method for limited view 3D PAT reconstruction with curved array". In Proceedings of: *Photons Plus Ultrasound: Imaging and Sensing*. Vol. 12379. International Society for Optics and Photonics, pp. 91–94. DOI: 10.1117/12.2649593.