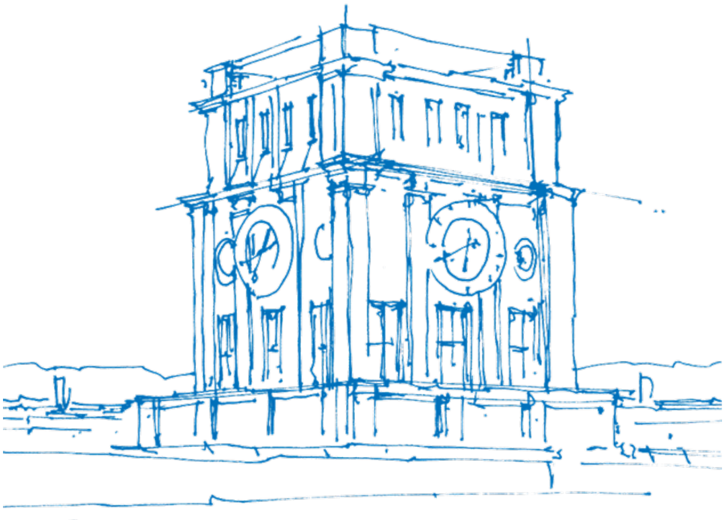
TUM School of Computation, Information and Technology
Technical University of Munich

# Accurate and real-time imaging with multispectral optoacoustic tomography by means of deep learning

Christoph R. Dehner



TUM School of Computation, Information and Technology
Technische Universität München

# Accurate and real-time imaging with multispectral optoacoustic tomography by means of deep learning

Christoph R. Dehner

Vollständiger Abdruck der von der TUM School of Computation, Information and Technology der Technischen Universität München zur Erlangung des akademischen Grades eines

**Doktors der Naturwissenschaften (Dr. rer. nat.)**

genehmigten Dissertation.

**Vorsitz:**

Prof. Dr. Daniel Rückert

**Prüfende der Dissertation:**

1. Prof. Dr. Vasilis Ntziachristos
2. Prof. Dr. Daniel Cremers
3. Prof. Dr. Björn Menze

Die Dissertation wurde am 11.10.2023 bei der Technischen Universität München eingereicht und durch die TUM School of Computation, Information and Technology am 18.09.2024 angenommen.

# Abstract

Multispectral optoacoustic tomography (MSOT) enables non-invasive detection of optical contrast in living tissue with high spatial resolution and several centimeters of penetration depth, and can thus provide novel clinical insights for various multifarious diseases. However, degradations in image quality during in vivo imaging limit the clinical applicability of MSOT and impede the deployment of the imaging modality beyond proof-of-concept studies. Optoacoustic image contrast is considerably reduced by electrical noise generated during signal acquisition due to thermal effects and electromagnetic interference. In addition, the state-of-the-art image reconstruction method for optoacoustic tomography (iterative model-based reconstruction) is not available for in vivo imaging applications because it is too time-consuming to support image feedback in real-time. However, real-time imaging is essential for the clinical use of MSOT to enable precise visio-tactile coordination during imaging, provide in situ diagnoses, and detect dynamic pathophysiological changes associated with disease progression.

In this work, we investigate how the image quality and clinical applicability of MSOT can be improved using discriminative deep learning. We show that discriminative deep neural network models facilitate advanced data processing for MSOT by leveraging their ability to capture complex data transforms in a data-driven manner and efficiently apply these transforms to new data. First, we introduce a deep learning framework to remove electrical noise from optoacoustic images, enabling the detection of optoacoustic contrast with high spatial and temporal resolution several centimeters deep in tissue. Second, we develop a deep learning approach to accelerate state-of-the-art model-based optoacoustic image reconstruction, making it available for real-time use.

Finally, we highlight ongoing and planned research aimed at integrating the deep learning approaches from this work into clinical trials, and outline directions for further research on deep learning methods to improve the imaging capability and clinical applicability of MSOT.

# Zusammenfassung

Die multispektrale optoakustische Tomographie (MSOT) ermöglicht in einzigartiger Weise die nichtinvasive Detektion von optischem Kontrast in lebendem Gewebe mit hoher räumlicher Auflösung und mehreren Zentimetern Eindringtiefe. MSOT kann so neuartige klinische Erkenntnisse für vielfältige Krankheitsbilder liefern. Verschlechterungen der Bildqualität während der In-vivo-Bildgebung begrenzen jedoch die klinische Anwendbarkeit von MSOT und behindern den Einsatz der Bildgebungsmodalität über Proof-of-Concept-Studien hinaus. Optoakustischer Bildkontrast wird erheblich durch elektrisches Rauschen verringert, das bei der Signalerfassung durch thermische Effekte und elektromagnetische Interferenzen entsteht. Darüber hinaus steht die modernste Bildrekonstruktionsmethode in der optoakustische Tomographie (iterative modellbasierte Rekonstruktion) nicht für die In-vivo-Bildgebung zur Verfügung, da sie für die Echtzeitanwendung zu zeitaufwändig ist. Echtzeit-Bildgebung ist jedoch für den klinischen Einsatz von MSOT unerlässlich, um eine präzise visuell-taktile Koordination während der Bildgebung zu ermöglichen, In-situ-Diagnosen bereitzustellen und dynamische pathophysiologische Veränderungen erkennen zu können.

In dieser Arbeit untersuchen wir, wie die Bildqualität und klinische Anwendbarkeit von MSOT mithilfe von diskriminativem Deep Learning verbessert werden können. Wir zeigen, dass diskriminierende tiefe neuronale Netzwerkmodelle fortschrittliche Methoden zur Datenverarbeitung für MSOT bereitstellen, indem sie komplexe Transformationen datengetrieben erfassen und diese Transformationen effizient auf neue Daten anwenden. Zunächst stellen wir ein Deep-Learning-Framework vor, um elektrisches Rauschen aus optoakustischen Bildern zu entfernen und so die Erkennung von optoakustischem Kontrast mit hoher räumlicher und zeitlicher Auflösung mehrere Zentimeter tief im Gewebe zu ermöglichen. Zweitens entwickeln wir einen Deep-Learning-Ansatz, um die hochmoderne modellbasierte optoakustische Bildrekonstruktion zu beschleunigen und für den Echtzeiteinsatz verfügbar zu machen.

Abschließend erläutern wir laufende und geplante Forschungsarbeiten, die darauf abzielen, die Deep-Learning-Ansätze aus dieser Arbeit in klinische Studien zu integrieren, und skizzieren Richtungen für weitereführende Forschung nach Deep-Learning-Methoden zur Verbesserung der Bildgebungsfähigkeit und der klinischen Anwendbarkeit von MSOT.

# Contents

# 1 Discriminative deep learning for multispectral optoacoustic tomography – an overview

## 1.1 Introduction

### 1.1.1 Clinical imaging with multispectral optoacoustic tomography

Multispectral optoacoustic tomography (MSOT) is an emerging clinical imaging modality due to its unique capability to non-invasively detect optical contrast with high spatial resolution and centimeter-scale penetration depth in living tissue [60]. Optoacoustic tomography uses as contrast mechanism the optoacoustic effect, which describes the generation of acoustic pressure waves (ultrasonic waves) by tissue chromophores after transient light absorption. Multispectral imaging with optoacoustic tomography (i.e., repeated scanning of a tissue location with different optical excitation wavelengths) enables to access the multispectral contrast of endogenous tissue chromophores such as oxygenated and deoxygenated hemoglobin, lipids, water, and collagen. MSOT affords high-resolution imaging of deep tissue because acoustic pressure waves scatter significantly less than light in tissue [89]. The resolution that can be achieved with optoacoustic tomography in deep tissue is therefore significantly higher than with a purely optical imaging approach such as for example diffuse optical tomography [35], whose imaging depth is severely limited by optical tissue scattering.

MSOT imaging is well suited for use in clinical applications because it employs non-ionizing radiation, is portable and inexpensive, and can be easily combined with ultrasonography, which is already established in the clinical imaging routine. Several proof-of-concept studies have demonstrated that MSOT can quantify functional tissue parameters related to oxygenation, inflammation, vascularization, and tissue fibrosis, and provide unmatched clinical information for multifarious diseases such as breast cancer [17, 47], oral cancer [84], inflammatory bowel disease [43], Duchenne muscular dystrophy [67], peripheral neuropathy [40], and thyroid disorders [69].

In practice, however, degraded image quality during in vivo imaging limits the clinical applicability of MSOT and impedes its use beyond proof-of-concept studies. Optoacoustic image contrast is considerably decreased by signal corruptions due to electrical noise [80]. Electrical noise arises from thermal effects (thermal noise) and electromagnetic interferences (parasitic noise) and severely limits the imaging sensitivity with optoacoustic tomography. Distortions by electrical noise are particularly detrimental in deep tissue, where the signal-to-noise ratio of optoacoustic signals is additionally challenged by light fluence attenuation. Moreover, multispectral imaging combines scans acquired at different wavelengths, which exacerbates the effects of image distortion. As a result of the limited sensitivity of MSOT, reliable quantification in clinical studies is currently only possible at reduced spatial and temporal resolution through averaging over larger tissue regions and multiple scans.

Optoacoustic image quality is furthermore limited, since in many applications only real-time capable image processing and reconstruction algorithms can be used. Real-time image feedback is essential in optoacoustic imaging, especially in handheld mode, to facilitate visio-tactile coordination, identify and localize relevant tissue structures based on anatomical landmarks in their vicinity, and determine the optimal scanning position for the target region. Real-time optoacoustic imaging is also required to enable in situ guidance and diagnosis during intra-operative and endoscopic imaging, and to detect and monitor dynamic physiological processes in the imaged tissue. However, state-of-the-art image processing algorithms for optoacoustic tomography like variational reconstruction methods are too computationally demanding to enable real-time processing.

In summary, MSOT imaging is currently only available with sub-optimal image quality and after considerable computation time. These shortcomings make the imaging modality inadequate for extensive use in clinical applications. Therefore, in order to improve the imaging capabilities and the clinical applicability of MSOT, novel signal and image processing techniques have to be developed.

### 1.1.2 Thesis objectives

In recent research, discriminative deep learning models have achieved state-of-the-art performance for various general image enhancement and reconstruction tasks [4]. Discriminative deep learning models aim to find appropriate data transformations for regression or classification tasks in a data-driven manner. In doing so, they use the ability of deep neural networks to capture complex data transformations during training and efficiently apply these transformations to new data.

The modeling capabilities and computational efficiency of discriminative deep learning models have already been used in other medical imaging modalities such as magnetic resonance imaging, X-ray computed tomography, and ultrasound imaging, improving image quality, usability, and safety. In magnetic resonance imaging, deep-learning-based image reconstruction of under-sampled data was shown to outperform conventional compressed sensing approaches and facilitate accurate dynamic imaging [74, 97, 31]. In X-ray computed tomography imaging, discriminative deep learning was applied to reconstruct high quality images from low-dose or sparse view measurements and thus reduce the overall amount of ionizing radiation that a patient is exposed to [39, 9]. In ultrasound imaging, deep neural network models could improve the image quality and contrast of ultrasound beamforming [75, 41, 56, 57] and accelerate ultrasound localization microscopy for super-resolution microvascular imaging [19, 83, 82].

In this work, we investigate whether discriminative deep learning can improve the image quality and clinical applicability of MSOT. We show that the modeling capabilities and the computational efficiency of discriminative deep learning models offer suitable data processing methodologies to improve the sensitivity and versatility of MSOT during clinical use. A key challenge in developing such enhancement methods for MSOT is to ensure that they can be applied to any in vivo data with the same high precision. We identify the generation of training data as a key aspect for the use of deep learning for MSOT and present strategies to account for the complexity and highly variable composition of biological tissue when compiling training data.

Specifically, we define the following two objectives for this thesis:

1. Development of a deep learning methodology to remove electrical noise from optoacoustic images and facilitate accurate and reliable (multispectral) optoacoustic imaging of several centimeters deep tissue at high spatial and temporal resolution.
2. Development of a deep learning methodology to speed up variational optoacoustic image reconstruction and enable (multispectral) optoacoustic imaging in real-time with state-of-the-art image quality.

### 1.1.3 Thesis outline

This thesis is structured into four chapters. In chapter 1, we summarize the motivation for the conducted research and explain the most important methodological concepts used in the remainder of the thesis. In section 1.2, we provide an overview about the forward imaging process of MSOT and the most commonly applied inversion techniques. In section 1.3, we introduce the general methodology of discriminative deep learning. In section 1.4, we give an overview about how discriminative deep learning has been applied in prior works to solve inverse problems in imaging. In section 1.5, we finally discuss application strategies of discriminative deep learning for multispectral optoacoustic tomography imaging. For in-depth explanations about individual aspects of the covered topics, we refer the reader to the references provided in the respective paragraphs.

In chapters 2 and 3, we address the two main objectives of this work. In chapter 2, we introduce a deep learning approach to remove electrical noise from recorded optoacoustic signals and reveal high morphological and spectral optoacoustic contrast in several centimeter deep tissue. In chapter 3, we presents a deep-learning-based image reconstruction framework to obtain state-of-the-art optoacoustic images fast enough to enable real-time imaging.

Finally, in chapter 4 we summarize the results of this thesis and discuss possible directions for further research.

### 1.1.4 Publication record

This work is a cumulative dissertation based on the following two publications:

1. **Christoph Dehner**, Ivan Olefir, Kaushik Basak Chowdhury, Dominik Jüstel, Vasilis Ntziachristos, *Deep-learning-based electrical noise removal enables high spectral optoacoustic contrast in deep tissue*, IEEE Transactions on Medical Imaging, 41(11):3182–3193, 2022.

2. **Christoph Dehner**, Guillaume Zahnd, Vasilis Ntziachristos, Dominik Jüstel, *A deep neural network for real-time optoacoustic image reconstruction with adjustable speed of sound*, Nature Machine Intelligence, 2023.

In addition, the research conducted for this dissertation project resulted in a patent application and contributed to four other thematically-related publications:

3. Dominik Jüstel, **Christoph Dehner**, Stefan Morscher, Guillaume Zahnd, Antonia Longo, *Methods and system for optoacoustic and/or ultrasonic imaging, reconstructing optoacoustic and/or ultrasonic images and training an artificial neural network provided therefor*, European Patent Application no. 22177153.8.

4. Kaushik Basak Chowdhury, Maximilian Bader, **Christoph Dehner**, Dominik Jüstel, Vasilis Ntziachristos, *Individual transducer impulse response characterization method to improve image quality of array-based handheld optoacoustic tomography*, Optics Letters, 46(1):1–4, 2021.

5. Jan Kukačka, Stephan Metz, **Christoph Dehner**, Alexander Muckenhuber, Korbinian Paul-Yuan, Angelos Karlas, Eva Maria Fallenberg, Ernst Rummeny, Dominik Jüstel, Vasilis Ntziachristos, *Image processing improvements afford second-generation handheld optoacoustic imaging of breast cancer patients*, Photoacoustics, 26:100343, 2022.

6. Dominik Jüstel, Hedwig Irl, Florian Hinterwimmer, **Christoph Dehner**, Walter Simson, Nassir Navab, Gerhard Schneider, Vasilis Ntziachristos, *Spotlight on nerves: Portable multispectral optoacoustic imaging of peripheral nerve vascularization and morphology*, Advanced Science, 10(19):2301322, 2023.

7. Markus Seeger, **Christoph Dehner**, Dominik Jüstel, Vasilis Ntziachristos, *Label-free concurrent 5-modal microscopy (Co5M) resolves unknown spatio-temporal processes in wound healing*, Communications Biolology, 4(1):1040, 2021.

## 1.2 Methodology of multispectral optoacoustic tomography

MSOT can non-invasively detect optical contrast at acoustic resolution by measuring acoustic pressure waves that are emitted by tissue chromophores after transient illumination. Figure 1.1 illustrates the overall imaging process of MSOT. Acoustic pressure sinograms are recorded for different excitation wavelengths (one sinogram per excitation wavelength). During acoustic inversion, the underlying initial pressure distributions of the recorded sinograms are reconstructed. Finally, during optical inversion, the initial pressure images obtained with different excitation wavelengths are combined to infer the concentrations of different chromophores in the scanned tissue.

In the following, we summarize the key physical and mathematical concepts for MSOT, as well as the most common inversion methods currently in use. The purpose of these explanations is to provide the necessary methodological background for an investigation of meaningful signal and image processing improvements for MSOT. In section 1.2.1, we explain the optoacoustic forward imaging process. In sections 1.2.2 and 1.2.3, we formally define the acoustic and optical inverse problems related to MSOT, respectively, and outline the most commonly applied inversion approaches.

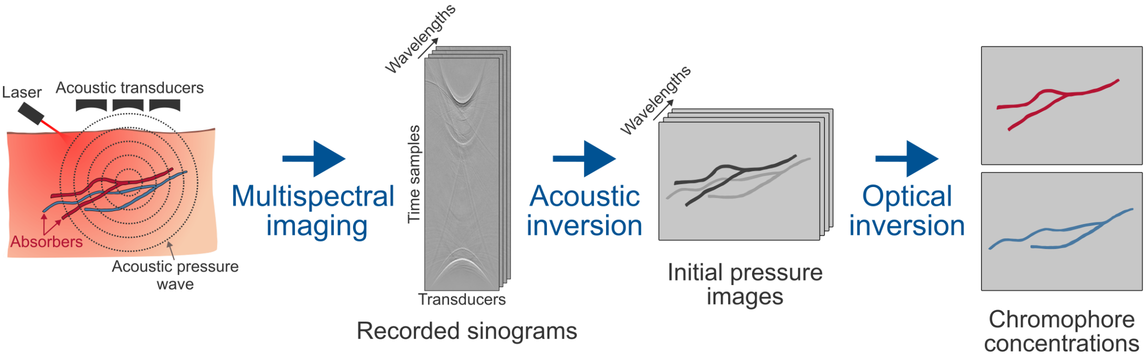


**Figure 1.1** High-level overview of the imaging and reconstruction pipeline for multispectral optoacoustic tomography (MSOT). During multispectral optoacoustic imaging, acoustic pressure sinograms are recorded for different excitation wavelengths (one sinogram per excitation wavelength). During acoustic inversion, the underlying initial pressure distributions of the recorded sinograms are reconstructed. During optical inversion, the initial pressure images obtained with different excitation wavelengths are combined to infer the concentrations of different chromophores in the scanned tissue.

### 1.2.1 Modeling of the imaging physics

The underlying principle of optoacoustic imaging is the generation of pressure waves by means of optical excitation due to a physical process called the optoacoustic effect. Figure 1.2 schematically visualizes the optoacoustic imaging principle. The object of interest is illuminated with a short laser pulse of typically several nanoseconds duration. The deposed optical energy is then (in part) absorbed by chromophores and converted into heat, which induces a local increase in pressure due to thermoelastic expansion. Assuming thermal and stress confinement (i.e., heat conduction and expansion of the absorbing volume are negligible during optical excitation), the increase in pressure $p_0$ at a given location $\boldsymbol{r}$ in the imaged medium may be written as

$$p_0(\boldsymbol{r}) = \Gamma H_a(\boldsymbol{r}) = \Gamma \mu_a(\boldsymbol{r}) \Phi(\boldsymbol{r}), \tag{1.1}$$

where $\Gamma$ is the Grüneisen coefficient describing the conversion efficiency of heat to pressure, $H_a$ is the absorbed energy density, $\mu_a$ is the optical absorption coefficient, and $\Phi$ is the local light fluence. On closer inspection, the optoacoustic efficiency $\Gamma$ is not constant but varies for different tissue types; however, these variations are typically neglected or considered part of the absorption coefficient distribution [92, 13].

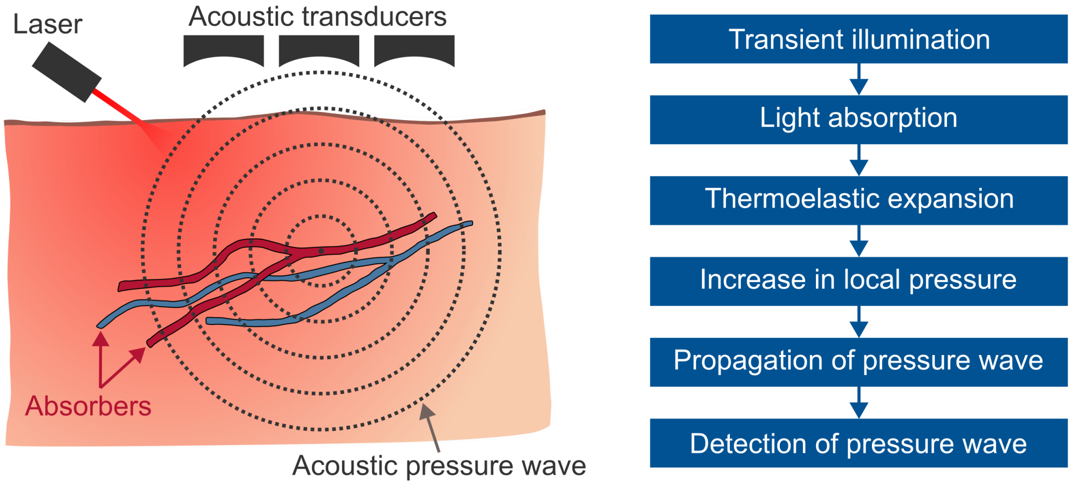


**Figure 1.2** Schematic visualization of the optoacoustic imaging principle. Light from transient illumination is absorbed by tissue chromophores and converted into heat. The local pressure increases due to thermoelastic expansion and an acoustic pressure wave propagates through the tissue; this wave can be recorded with acoustic transducers.

As a result of an increase in local pressure due to the optoacoustic effect, acoustic pressure waves propagate through the imaged object and can be recorded with acoustic transducers (acoustic pressure detectors). Thereby, the pressure $p$ at location $\boldsymbol{r}$ and time $t$ is governed by the acoustic wave equation, which reads as follows under the assumption of an acoustically homogeneous medium:

$$\frac{\partial^2 p(\boldsymbol{r},t)}{\partial t^2} - c^2 \nabla^2 p(\boldsymbol{r},t) = \Gamma \frac{\partial H(\boldsymbol{r},t)}{\partial t}, \tag{1.2}$$

where $c$ is the speed of sound, $\Gamma$ the Grüneisen coefficient, and $H(\boldsymbol{r},t)$ is the deposited energy per unit volume and per unit time. If illumination is modeled as an infinitely short pulse, the source term on the right side of equation 1.2 can be additionally reformulated to $H(\boldsymbol{r},t) = H_a(\boldsymbol{r})\delta(t)$ using the Dirac delta distribution $\delta(t)$. This simplification is valid for most optoacoustic imaging applications because the illumination pulse length is typically below the sampling rate of the acoustic transducers, allowing to derive an analytical solution the for pressure $p(\boldsymbol{r},t)$ from equation 1.2 using the Green's function approach [45, 72]:

$$p(\boldsymbol{r},t) = \frac{1}{4\pi c}\frac{\partial}{\partial t}\int_{\|\boldsymbol{r}-\boldsymbol{r'}\|_2 = ct} \frac{p_0(\boldsymbol{r'})}{\|\boldsymbol{r}-\boldsymbol{r'}\|_2} d\boldsymbol{r'}. \tag{1.3}$$

The integral in equation 1.3 can be interpreted as a superposition of elementary pressure waves on a sphere with radius $\|\boldsymbol{r}-\boldsymbol{r'}\|_2 = ct$.

For individual and radially symmetric absorbers, the resulting acoustic pressure wave can be computed analytically, yielding the characteristic N-shaped wave profile [15]. Specifically, let us consider an absorber at the origin with the radial initial pressure profile $p_0^\Delta(R)$ and diameter $\Delta$ (i.e., $p_0^\Delta(R) = 0$ for $|R| > \Delta$). Then, according equation 11 from [15], the generated pressure $p_{r_d}^\Delta(t)$ at location $\boldsymbol{r_d}$ for $t > 0$ and $\|\boldsymbol{r_d}\|_2 \gg \Delta$ is given as

$$p_{r_d}^\Delta(t) = \frac{\|\boldsymbol{r_d}\|_2 - tc}{2\Delta}\, p_0^\Delta(\|\boldsymbol{r_d}\|_2 - tc). \tag{1.4}$$

The ability to analytically derive the generated acoustic pressure wave for a radially symmetric absorber enables to numerically simulate the acoustic part of the optoacoustic forwards imaging process in an efficient manner as follows: First, the initial pressure $p_0(\mathbf{r})$ is discretized with radially symmetric basis functions; and second, the pressure wave $p(\boldsymbol{r}, t)$ is calculated as a superposition of the pressure waves from the basis functions according to the equation 1.4, while also taking into account their travel times with respect to each other. This simulation approach is used in model-based inversion methods for the acoustic part of the optoacoustic imaging process, which is explained in detail in the following section (section 1.2.2).

### 1.2.2 Acoustic inverse problem

The acoustic inverse problem related to optoacoustic tomography comprises reconstruction of the spatial initial pressure distribution $p_0(\boldsymbol{r}) \equiv p(\boldsymbol{r}, t = 0)$, given acoustic pressure measurements $p(\boldsymbol{r_d}, t)$ at the detection locations $\boldsymbol{r_d}$ and times $t > 0$. Reconstruction of the initial pressure is a linear inverse problem and is in practice ill-posed because of limited angle acquisition, finite bandwidth of acoustic transducers, and measurement noise. In the following, we review the two most commonly applied inversion approaches for optoacoustic tomography: Backprojection and model-based reconstruction. Acoustic wave propagation in tissue is generally a 3D phenomenon, however, the problem setting can be simplified to 2D if pressure measurements are restricted to a specific plane by using focused transducers, or if the imaged medium is homogeneous in the third dimension [72]. Figure 1.3 compares backprojection and model-based reconstructions for a scan of a human breast to illustrate the advantages and disadvantages of the two methods.

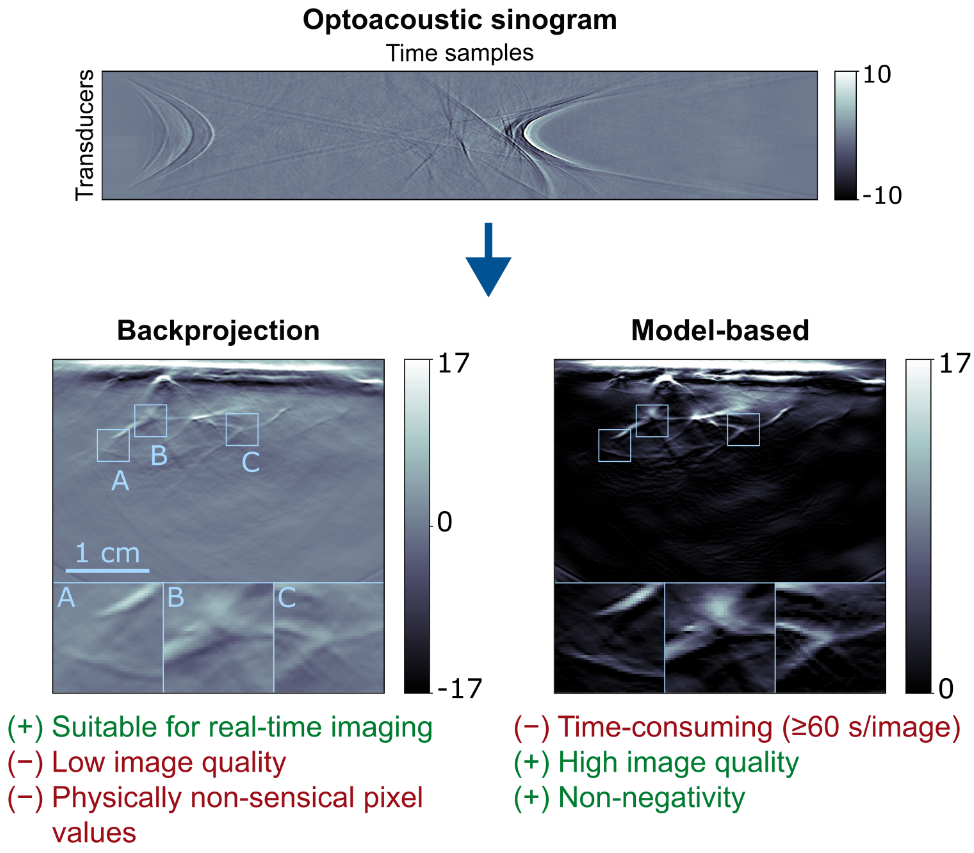


**Figure 1.3** Comparison of backprojection and model-based reconstruction. Top: Optoacoustic sinogram from a scan of a human breast at 800 nm. Bottom: Initial pressure images from backprojection (left) and model-based reconstruction (right).

### Backprojection

Backprojection formulae provide closed-form solutions for the reconstruction of the initial pressure distribution (i.e. the inversion of equation 1.3) and can be derived for spherical, planar, and cylindrical detection geometries under the assumption of ideal imaging conditions such as point transducers, infinite-bandwidth signal acquisition, and full enclosure of the object of interest in case of a spherical detection geometry [91, 72]. The universal backprojection algorithm (equation 20 in [91]) is arguably one of the best known and most commonly applied backprojection formula. However, its numerical implementation is particularly sensitive to noise in the input signals because the algorithm involves calculating time derivatives of these signals (before the delay operation).

An alternative backprojection algorithm can be derived using the explicit inversion formulae for the spherical mean Radon transformation by Kunyansky [48] together with the solution of the wave equation based on the Poisson–Kirchhoff formulas (equations 19.6 and 19.7 in [46]). The resulting formula allows for a more robust numerical implementation because it involves a time integral of the input signals (the integral reverses derivative operation from Equation 19.6 in [46]). This integration acts as a filter, smoothing out high-frequency distortions in the input signals before calculating derivatives (see equation 8 in [48] for the 2D case). However, a drawback of this backprojection formula in comparison to the universal backprojection formula is that the obtained images may be slightly more blurred.

Backprojection algorithms are commonly applied because they support reconstruction in real-time (at least 24 frames per second are necessary for full-video rendering). However, the method can reconstruct the initial pressure only sub-optimally because the presumed ideal imaging conditions cannot be realized in practice and because the algorithm is unable to mitigate the ill-posedness of the inverse problem. Figure 1.3 (bottom left side) visualizes the image quality achieved with backprojection using an example scan of a human breast. As a result of the limitations listed above, backprojection images can suffer from reduced contrast and spatial resolution, as well as negative pixel values that obstruct a physically-meaningful interpretation as initial pressure.

### Model-based reconstruction

Model-based reconstruction (also called "variational reconstruction") relies on discretizing the forward imaging process based on equations 1.3 and 1.4. As wave propagation is linear, the discretized forward imaging operator can be written in matrix form as

$$\boldsymbol{s} = \boldsymbol{M}\boldsymbol{p_0}, \tag{1.5}$$

where $\boldsymbol{s}$ is a column vector containing the acoustic pressure measurements for a set of locations and time points, $\boldsymbol{M}$ the forward model matrix, and $\boldsymbol{p_0}$ a column vector containing the initial pressure values from the imaging grid.

For the inversion of equation 1.5, the following constrained least-square minimization problem is solved:

$$\boldsymbol{p_0} = \arg\min_{\boldsymbol{p}\geq 0} \left( \|\boldsymbol{M}\boldsymbol{p} - \boldsymbol{s}\|_2^2 + \alpha R(\boldsymbol{p}) \right), \tag{1.6}$$

where $R(\cdot)$ is the regularization functional, $\alpha \geq 0$ denotes the parameter to control the importance of the regularization, and the inequality sign $\boldsymbol{p} \geq 0$ refers to entrywise non-negativity. The minimization problem from equation 1.6 can be solved, for example, using bound-constrained sparse reconstruction by separable approximation [88, 8].

The main advantages of variational reconstruction are the ability to constrain the reconstructed values to be non-negative and thus enable a physically-meaningful interpretation as initial pressure, to mitigate the ill-posedness of the inverse problem via regularization, and to integrate into the inversion any linear effect from the imaging process as for example the total impulse response of the scanner or acoustic heterogeneities of the imaged object [71, 16, 11, 12]. Typical choices for the regularization are Tikhonov, total variation, or Shearlet-based regularization, of which the latter is particularly beneficial to decrease reconstruction artifacts that arise due to limited-angle acquisition [24, 49].

Figure 1.3 (bottom right side) shows the image quality obtained with model-based reconstruction using an example scan of a human breast. Model-based reconstruction obtains considerably better images in terms of spatial resolution, contrast, and interpretability than backprojection reconstruction. However, a critical disadvantage of model-based reconstruction is that the approach is in practice too time-consuming for real-time imaging (i.e., the reconstruction of one image can take up to one minute).

### 1.2.3 Optical inverse problem

The optical inverse problem related to multispectral optoacoustic tomography comprises reconstruction of chromophore concentrations $c_k(\boldsymbol{r})$ from initial pressure images obtained with different excitation wavelengths. Biological tissue contains only a small number of endogenous chromophores with known and distinct absorption spectra. Figure 1.4 plots the absorption spectra of important endogenous chromophores in biological tissue, namely oxygenated and deoxygenated hemoglobin, lipid, water, collagen, and melanin.

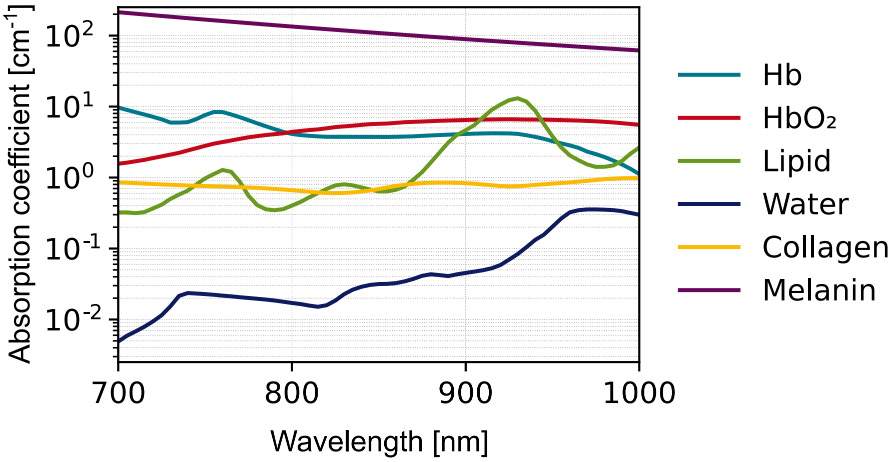


**Figure 1.4** Absorption spectra of endogenous chromophores in biological tissue [65]. Hb and HbO$_2$ denote deoxygenated and oxygenated hemoglobin, respectively.

Different chromophores can be unmixed using multispectral imaging because their concentrations determine the wavelength-dependent optical absorption in the imaged medium:

$$\mu_a^{(\lambda)}(\boldsymbol{r}) = \sum_{k=1}^{K} c_k(\boldsymbol{r}) s_k(\lambda), \tag{1.7}$$

where $\lambda$ denotes the excitation wavelength, $K$ is the number of chromophores in the imaged object, and $s_k(\lambda)$ is the known absorption spectra of the respective chromophores. The complete relationship between the chromophore concentrations and the initial pressure distribution for a specific wavelength $p_0^{(\lambda)}(\boldsymbol{r})$ is obtained by substituting equation 1.7 into equation 1.1:

$$p_0^{(\lambda)}(\boldsymbol{r}) = \Gamma \Phi^{(\lambda)}(\boldsymbol{r}) \sum_{k=1}^{K} c_k(\boldsymbol{r}) s_k(\lambda). \tag{1.8}$$

Retrieval of chromophore concentrations based on equation 1.8 is a non-linear and ill-posed inverse problem because the light fluence $\Phi^{(\lambda)}(\boldsymbol{r})$ depends in a non-linear and wavelength-dependent manner on the optical properties (and thus also the chromophore concentrations) of the entire illuminated volume. Wavelength-dependent light fluence attenuation causes distortions of the perceived absorption spectra of chromophores in deeper tissue by light absorption of chromophores in superficial tissue – an effect called spectral coloring – and severely limits the ability to detect and quantify different chromophores with MSOT [13, 79]. Correcting these spectral coloring distortions is still an open research question, as accurate information about optical tissue properties is not available during in vivo imaging. Previous research studies demonstrated the possibility to mitigate spectral coloring effects for a pre-clinical MSOT system using a simulation-derived eigenspectral model for light fluence in tissue, called "eMSOT" [81, 61, 62]. However, the feasibility of this methodology for high-resolution clinical imaging remains questionable and has yet to be demonstrated, as the proposed methodology relies on an overly-simplified tissue model, utilizes a pre-clinical imaging setup with full-angle illumination and imaging depths of only about one centimeter, considers only oxygenated and deoxygenated hemoglobin as tissue chromophores, and estimates the light fluence on a coarse grid only.

Because the optical inverse problem given in equation 1.8 cannot be precisely solved, approximate unmixing techniques for MSOT are commonly applied to disentangle different tissue chromophores and estimate their concentrations. In the following, we describe two of the most widely used approximate unmixing methods: Linear unmixing and blind spectral unmixing. We thereby denote with $N$ the number of spatial locations for which spectra are simultaneously unmixed, with $L$ the number of measurements with different wavelengths for each location, and with $K$ the number of unmixing components.

### Linear unmixing

Linear unmixing assumes that the light fluence is constant for all imaging locations and wavelengths. With this simplifying assumption, the initial pressure measurements with different excitation wavelengths are linearly related to the chromophore concentrations:

$$\boldsymbol{d} = \boldsymbol{S}\boldsymbol{c}, \tag{1.9}$$

where $\boldsymbol{d}$ is a column vector of size $L$ with the initial pressure measurements for the considered set of excitation wavelengths, $\boldsymbol{S}$ is a matrix of size $L \times K$ with as columns the reference absorption spectra of $K$ considered unmixing chromophores, and $\boldsymbol{c}$ is a column vector of size $K$ with the concentrations of the considered chromophores. To recover the concentrations of $K$ chromophores, measurements with $L \geq K$ different wavelengths are required. Equation 1.9 can be inverted using, for example, linear matrix inversion or least squares regression. In practice, least-squares regression is usually a preferred inversion approach because it can ensure a unique unmixing solution with a physically-meaningful interpretation as chromophore concentrations by introducing a regularization and constraining the found coefficients to be non-negative:

$$\boldsymbol{c} = \arg\min_{\tilde{\boldsymbol{c}} \geq 0} \left( \|\boldsymbol{d} - \boldsymbol{S}\tilde{\boldsymbol{c}}\|_2^2 + \alpha R(\tilde{\boldsymbol{c}}) \right), \tag{1.10}$$

where $R(\cdot)$ denotes a regularization functional, $\alpha \geq 0$ is a parameter to control the importance of the regularization, and the inequality sign $\tilde{\boldsymbol{c}} \geq 0$ refers to entry-wise non-negativity.

### Blind spectral unmixing

Another unmixing method for MSOT is blind spectral unmixing via non-negative matrix factorization [53]. Blind spectral unmixing finds both the spectra and unmixing coefficients in a data-driven way and thus can extract variants of the reference spectra that consider effects from spectral coloring. As additional advantage, blind spectral unmixing via non-ngeative matrix factorization enables to incorporate prior information into the unmixing process via regularization. For example, $L^1$-based regularization of the unmixing coefficients may be used to promote sparse decompositions and account for the fact that spectral contrast in biological tissue is typically dominated by a small number of chromophores.

$$(\boldsymbol{S}, \boldsymbol{C}) = \arg \min_{(\tilde{\boldsymbol{S}}, \tilde{\boldsymbol{C}}) \geq 0} \left( \|\boldsymbol{D} - \tilde{\boldsymbol{S}}\tilde{\boldsymbol{C}}\|_F^2 + \alpha_1 R_1(\tilde{\boldsymbol{S}}) + \alpha_2 R_2(\tilde{\boldsymbol{C}}) \right), \tag{1.11}$$

where $\boldsymbol{C}$ is a matrix of size $K \times N$ with as columns the reconstructed chromophore concentrations of all considered scanning locations, $\boldsymbol{S}$ is a matrix of size $L \times K$ with as columns the reconstructed absorption spectra, $\boldsymbol{D}$ is a matrix of size $L \times N$ with as columns the initial pressure measurements of all data points at the wavelengths $\lambda_1, \ldots, \lambda_L$. $R_1(\cdot)$ and $R_2(\cdot)$ are regularization functionals, $\alpha_1 \geq 0$ and $\alpha_2 \geq 0$ parameters to control the importance of the regularization, the inequality sign $(\tilde{\boldsymbol{S}}, \tilde{\boldsymbol{C}}) \geq 0$ refers to entry-wise non-negativity, and $\| \cdot \|_F$ denotes the Frobenius norm.

## 1.3 Methodology of discriminative deep learning

Discriminative deep learning has emerged as a powerful signal and image processing technique in recent research because it enables to capture complex data transformations in a data-driven way and efficiently apply these transformations to new data. In the current and following sections, we provide an overview about discriminative deep learning and its use in imaging, as a basis for developing and investigating advanced signal and image processing techniques for MSOT. First, we explain in the further course of this section the general mathematical background of discriminative deep learning. Subsequently, we review in section 1.4 the methodologies of existing discriminative-deep-learning-based approaches for solving inverse problems related to image enhancement and reconstruction.

### 1.3.1 Discriminative modeling

Discriminative modeling aims to approximate a mapping $\mathbf{f} : \mathcal{Y} \rightarrow \mathcal{X}$ with a parameterized function $\mathbf{g}_\theta$,

$$\mathbf{g}_\theta(\mathbf{y}) \approx \mathbf{x} = \mathbf{f}(\mathbf{y}), \tag{1.12}$$

using a dataset of $N$ independent, identically distributed observations $\{\mathbf{y}_i \in \mathcal{Y} \mid 1 \leq i \leq N\}$ and corresponding outputs $\{\mathbf{x}_i \in \mathcal{X} \mid 1 \leq i \leq N\}$. To tune the learnable parameters $\theta$ of the function $\mathbf{g}_\theta$, its average prediction error on the given dataset is minimized:

$$\theta = \arg\min_{\tilde{\theta}} \frac{1}{N} \sum_{i=1}^{N} e(\mathbf{g}_{\tilde{\theta}}(\mathbf{y}_i), \mathbf{x}_i), \tag{1.13}$$

where $e : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}_{\geq 0}$ denotes a suitable error function that may be chosen specifically for a given problem. Equation 1.13 can also be viewed as a maximum likelihood estimation of $\theta$, given a suitable probabilistic interpretation of the target values $\mathbf{x}$. For example, using the mean squared error function, equation 1.13 yields the the maximum likelihood estimate of $\theta$ under the assumption that the conditional probability for the target values is given by a Gaussian distribution, $p(\mathbf{x}|\mathbf{y}, \theta) = \mathcal{N}\left(\mathbf{x} \mid \mathbf{g}_\theta(\mathbf{y}), \beta^{-1}\mathbf{I}\right)$, where $\mathbf{I}$ denotes an identity matrix and $\beta \in \mathbb{R}$ the shared noise precision [5].

### 1.3.2 Deep neural networks

Deep neural networks are among the most used and successful parametric models in recent times [4]. A deep neural network implements the mapping between its input and output domains via a sequence of parameterized linear transformations and non-linear activation functions, which is in its simplest form given as

$$\mathbf{g}_\theta(\mathbf{y}) = \psi \circ \mathbf{h}^{(L)} \circ \cdots \circ \mathbf{h}^{(1)}(\mathbf{y}), \tag{1.14}$$

$$\mathbf{h}^{(l)}(\mathbf{y}) = \mathbf{z}^{(l)} \circ \sigma\left(\mathbf{y}\right), \tag{1.15}$$

where $\mathbf{h}^{(l)}$ refers to one layer of the network, comprising one parameterized linear transformation $\mathbf{z}^{(l)}$ followed by one non-linear and element-wise applied activation function $\sigma$, $L$ denotes the number of layers of the network, and $\psi$ refers to an optional and application-specific outermost activation function. The notion of a *deep* neural network (as opposed to a *shallow* neural network) refers to the network being composed of multiple layers. Typical choices for the non-linear activation $\sigma$ are the Sigmoid function, the hyperbolic tangent function, or a rectified linear unit function.

Another important design choice for a deep neural network is the structure of its linear transformations $\mathbf{z}^{(l)}$. A so-called *fully-connected* deep neural network is obtained if weighted sums are employed as linear transformations, i.e.

$$\mathbf{z}^{(l)}(\mathbf{y}) = \mathbf{W}^{(l)}\mathbf{y} + \mathbf{b}^{(l)}, \tag{1.16}$$

where both the input $\mathbf{y}$ and the output $\mathbf{z}^{(l)}(\mathbf{y})$ are vector-shaped, and the weight matrix $\mathbf{W}^{(l)}$ and the bias vector $\mathbf{b}^{(l)}$ are the trainable parameters.

For imaging data, *convolutional* neural networks are typically used because their design enables to model translationally-invariant data transformations and decouple the required number of trainable parameters from the size of the input and output images. A convolutional neural network is obtained if discrete convolutions are employed as the linear transformations, i.e.

$$\mathbf{z}_{i_d}^{(l)}(\mathbf{y}) = \sum_{i_c=1}^{c} \mathbf{y} * \mathbf{w}_{i_c,i_d}^{(l)} + b_{i_d}^{(l)} \quad \text{for} \quad i_d \in \{1, 2, \ldots, d\}, \tag{1.17}$$

where the input and the output of the linear transformation $\mathbf{z}^{(l)}$ are multichannel images (i.e., $\mathbf{y} \in \mathbb{R}^{n \times n \times c}$, $\mathbf{z}^{(l)}(\mathbf{y}) \in \mathbb{R}^{m \times m \times d}$, $n, m, c, d \in \mathbb{N}$), the variable $\mathbf{z}_{i_d}^{(l)}(\mathbf{y})$ refers to one channel of the output image, the set of trainable parameters is comprised by the bias terms $\{b_{i_d}^{(l)} \in \mathbb{R}\}_{i_d}$ and the filters $\{\mathbf{w}_{i_c,i_d}^{(l)} \in \mathbb{R}^{n_w \times n_w}\}_{i_c,i_d}$ ($n_w \in \mathbb{N}$ is the kernel width), and $*$ denotes a discrete convolution. The above description of discrete convolutions as linear transformations is given for square images and filters but can be straightforwardly extended to non-square image. Also, the above explanations focus on the basic structure of a convolutional neural network. Deep convolutional neural networks used in real-world applications usually include additional components to improve their modeling and learning capabilities, such as skip connections, batch normalization, and pooling layers [37, 70, 34].

Given a suitable training dataset, the parameters of a discriminative deep neural network model (i.e., the weights and biases from the linear transformations $\mathbf{z}^{(l)}$ of the network) are tuned by minimizing the average distance between network outputs and ground truth targets. A formal description of the minimization problem underlying the training process is given in equation 1.13. Typical choices for the metric used to calculate the distances between network outputs and ground truth targets are the mean squared error and the mean absolute error. To solve the minimization problem from equation 1.13 and obtain a suitable parameter configuration for a deep neural network model, gradient-based methods such as stochastic gradient descent are applied, for which the gradients for the update are computed via backpropagation.

Stochastic gradient descent together with gradient backpropagation are the established methodology to train deep neural network model, yielding suitable parameter configurations in many applications. However, no mathematical guarantees can be derived that the parameter configurations found by stochastic gradient descent are optimal. Since the mappings implemented by deep neural networks are non-linear, the minimization problem from equation 1.13 is non-convex and gradient-based minimization methods can get stuck in a local minimum of the objective function [5, 52, 32]. Therefore, deep neural network models rely on extensive empirical validation.

## 1.4 Discriminative deep learning for inverse problems in imaging

### 1.4.1 Inverse problems in imaging

Discriminative deep learning has been successfully applied for various general image enhancement tasks like Gaussian denoising, deblurring, inpainting, superresolution, decompression, and dehazing, as well as for image reconstruction in magnetic resonance and X-ray computed tomography imaging [7, 94, 96, 95, 90, 58, 68, 74, 1, 31, 97, 9, 39]. All of these problems represent ill-posed inverse problems, in which an unknown and sought-after image $\mathbf{x}$ is observed via noisy measurements $\mathbf{y}$. The associated forward imaging processes can be generally expressed as

$$\mathbf{y} = \mathcal{M}(\mathbf{x}) = \mathcal{A}(\mathbf{x}) + \epsilon, \tag{1.18}$$

where the image $\mathbf{x}$ and measurements $\mathbf{y}$ are given as vectors, $\mathcal{M}(\cdot)$ denotes the complete forward imaging process, $\mathcal{A}(\cdot)$ represents an arbitrary forward measurement operator, and $\epsilon$ is an arbitrarily-distributed noise vector (in practice, measurement noise is often modeled as white Gaussian noise). In table 1.1, we list the forward measurement operator for some common inverse problems in imaging [64]. For linear imaging processes, the forward measurement operator corresponds to a matrix vector product $\mathcal{A}(\mathbf{x}) = \mathbf{A}\mathbf{x}$. Equation 1.18 relies on an additive noise model, but can easily be adapted for other types of noise, such as multiplicative noise.

**Table 1.1** Examples of inverse problems in imaging.

| Application | Forward operator | Description |
|---|---|---|
| (Gaussian) denoising | $\mathbf{A} = \mathbf{I}$ | $\mathbf{I}$ is the identity matrix. |
| Deblurring | $\mathcal{A}(\mathbf{x}) = \mathbf{h} * \mathbf{x}$ | $\mathbf{h}$ is a known blur kernel, $*$ denotes a discrete convolution. |
| Impainting | $\mathbf{A} = \mathbf{S}$ | $\mathbf{S}$ is a diagonal matrix with $S_{i,i} = 1$ for the pixels that are sampled and $S_{i,i} = 0$ for the pixels that are not sampled. |
| Superresolution | $\mathbf{A} = \mathbf{SB}$ | $\mathbf{S}$ is an undersampling matrix (identity matrix with missing rows) and $\mathbf{B}$ is a blurring operator (convolution with a blur kernel). |
| Decompression | $\mathcal{A}(\mathbf{x}) = \mathcal{E}(\mathbf{x})$ | $\mathcal{E}(\cdot)$ is a lossy image compression algorithm such as JPEG encoding. |
| (Undersampled) MRI reconstruction | $\mathbf{A} = \mathbf{SF}$ | $\mathbf{S}$ is an undersampling matrix (identity matrix with missing rows) and $\mathbf{F}$ is the discrete Fourier transformation (assuming Cartesian sampling). |
| X-ray computed tomography | $\mathbf{A} = \mathbf{R}$ | $\mathbf{R}$ is the discrete Radon transformation. |

Variational methods offer a general framework to solve inverse problems in imaging. Given a model of the image degradation process and suitable image priors (such as sparse representation [22, 18, 36] or low rank models [27, 63]), the desired image is obtained by solving the minimization problem

$$\boldsymbol{x} = \arg\min_{\tilde{\boldsymbol{x}}} \left( \|\mathcal{A}(\tilde{\boldsymbol{x}}) - \boldsymbol{y}\|_2^2 + \alpha R(\tilde{\boldsymbol{x}}) \right), \tag{1.19}$$

where $R(\cdot)$ is the regularization functional and $\alpha \geq 0$ denotes the parameter to control the importance of the regularization [96]. Variational inversion methods can achieve appreciable image quality, but have the following three disadvantages: First, variational inversion is iterative and thus time-consuming. Second, the analytical prior models used by variational inversion methods are typically unable to capture the full complexity of the output image domain, resulting in sub-optimal image quality. Third, variational methods involve hyper-parameters such as the regularization strength that need to be manually tuned and whose optimal values may vary for different input data.

### 1.4.2 Deep-learning-based inversion via direct inference

Discriminative deep learning methods have been proposed to tackle the aforementioned limitations of variational inversion methods. Discriminative deep neural network models enable to accurately capture inverse data transformations in a data-driven way and efficiently apply these transformationss to new data using modern graphics processing units. A key requirement for data-driven inversion techniques is the availability of a large training dataset of corrupted input measurements and corresponding noise-free ground-truth images. Strategies available for obtaining such a suitable training dataset depend on the specific characteristics and practical constraints of an inverse problem. A particularly common approach is to obtain noise-free ground truth images using an alternative (more complex) data processing or recording method that cannot be used permanently in practice (e.g., fully-sampled instead of under-sampled MRI data acquisition, or photographs taken with a high-end camera), and simulate thereof the corresponding corrupted measurements using the forward imaging models given in equation 1.18 and table 1.1.

In the following, we provide an overview about different discriminative deep-learning-based approaches for image reconstruction and enhancement and review their advantages and disadvantages. Existing deep-learning-based inversion methods for image enhancement and reconstruction approaches can be roughly divided into two categories: Direct inference methods and loop unrolling methods [96]. We focus on direct inference approaches in the current section and discuss loop unrolling methods in the following section.

Figure 1.5 schematically visualizes the methodology of direct inversion with a deep neural network. Direct inference approaches aim to invert a forward imaging process (as given in general form in equation 1.18) with a deep neural network and obtain the sought image via one forward inference pass through the trained network [39, 97, 64]:

$$\mathbf{x} = \mathcal{M}^{-1}(\mathbf{y}) \approx \mathbf{g}_\theta(\mathbf{y}), \tag{1.20}$$

where $\mathbf{g}_\theta(\cdot)$ denotes a deep neural network with trainable parameters $\theta$.

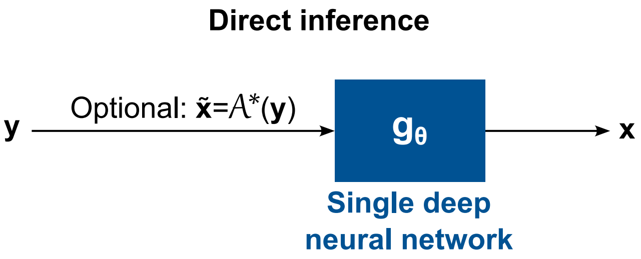


**Figure 1.5** Schematic methodology of deep-learning-based inversion via direct inference. The output image is obtained via one forward pass through a deep neural network.

If the input and output data of an imaging process have different sizes or come from different domains (as for example in the case of super-resolution or tomographic image reconstruction; also see table 1.1 for more details), the inverse data transformation from equation 1.20 includes a domain transformation. To efficiently implement such a transformation of the input measurements $\mathbf{y}$ to the output image domain of $\mathbf{x}$, deep neural networks in direct inversion approaches commonly apply an (approximate) adjoint operator of the corresponding forward process, denoted in the following with $\mathcal{A}^*(\cdot)$:

$$\mathbf{x} = \mathcal{M}^{-1}(\mathbf{y}) \approx \mathbf{g}_\theta(\mathcal{A}^*(\mathbf{y})). \tag{1.21}$$

Applying an (approximate) adjoint operator to the input data enables the deep neural network to operate in the image domain and potentially better exploit spatial invariance properties. Specifically, equation 1.21 can be rewritten to $\mathbf{x} \approx \mathbf{g}_\theta(\mathcal{A}^*(\mathcal{A}(\mathbf{x}))$ by inserting the definition of the noise-free forward operator ($\mathbf{y} = \mathcal{A}(\mathbf{x})$; also see equation 1.18), highlighting that the task of the deep neural network is to invert the normal operator in the given setting, i.e. $\mathbf{g}_\theta \approx (\mathcal{A}^* \circ \mathcal{A})^{-1}$ [1].

An alternative methodology for efficiently learning and expressing inverse data transformations between two different data domains is given by the AUTOMAP framework [97]. The paradigm of the AUTOMAP framework is to obtain the mapping between the measurement-domain input and the image-domain output data in a purely data-driven manner using fully-connected layers. Thus, the method does not require knowledge about or access to the targeted imaging operator. However, the application of fully-connected layers in the current version of the AUTOMAP framework prevents direct use in imaging applications that involve larger images.

With the ability to adjust to complex data characteristics during training, discriminative deep learning models for direct inversion can facilitate more accurate inversion results than traditional methods that rely on rigid analytical models. Moreover, deep-learning-based inversion can obviate the need of manually-tuned hyper-parameters, that may change at test time for different images, such as the regularization strength. Furthermore, direct inversion enables to compute inverse images in real-time and is therefore in most circumstances considerably faster than iterative variation approaches. However, a disadvantage of direct inversion with a deep neural network in comparison to variational inversion techniques is the lack of convergence guarantees. Discriminative deep neural networks are designed as black-box models that do not provide a way to derive optimality guarantees for their outputs, but instead rely on extensive empirical validations.

### 1.4.3 Deep-learning-based inversion via loop unrolling

Deep-learning-based inversion via loop unrolling incorporates discriminative deep learning into a variational formulation of the inverse problem, aiming to combine the advantages of both methods. The methodology is applicable to linear inverse problems and can be derived in three steps:

First, a deep neural network is integrated into the variational inversion formula from equation 1.19 as the regularization term:

$$\boldsymbol{x} = \arg\min_{\tilde{\boldsymbol{x}}} \left( \|\mathcal{A}(\tilde{\boldsymbol{x}}) - \boldsymbol{y}\|_2^2 + \alpha \, \|\tilde{\boldsymbol{x}} - \mathbf{g}_\theta(\tilde{\boldsymbol{x}})\|_2^2 \right), \tag{1.22}$$

where $\mathbf{g}_\theta(\cdot)$ is a deep neural network that obtains the denoised version of a given input image.

Second, using variable splitting, an iterative solution scheme for equation 1.22 is derived:

$$\boldsymbol{z}_n = \mathbf{g}_\theta(\boldsymbol{x}_n), \tag{1.23}$$

$$\boldsymbol{x}_{n+1} = \arg\min_{\tilde{\boldsymbol{x}}} \left( \|\mathcal{A}(\tilde{\boldsymbol{x}}) - \boldsymbol{y}\|_2^2 + \alpha \, \|\boldsymbol{x}_n - \boldsymbol{z}_n\|_2^2 \right), \tag{1.24}$$

where $\boldsymbol{z}_n$ is an intermediate variable that stores the the output of the deep-neural-network-based denoiser $\mathbf{g}_\theta$ (see equation 1.23). The sub-problem in equation 1.24 is called the data consistency step and can be solved using the normal equations:

$$\boldsymbol{x}_{n+1} = (\mathcal{A}^*\mathcal{A} + \alpha\mathcal{I})^{-1} (\mathcal{A}^*(\boldsymbol{y}) + \alpha \boldsymbol{z}_n), \tag{1.25}$$

where $\mathcal{A}^*$ denotes the adjoint operator of $\mathcal{A}$. A solution to equation 1.25 can be computed analytically for simple operators and in general using gradient descent or conjugate gradient minimization.

Third, the iterative scheme given in equations 1.23 and 1.24 is unrolled into an all-encompassing processing pipeline for a fixed number of iterations. Figure 1.6 schematically visualizes the unrolling process. The obtained processing pipeline consists of alternating noise estimation and data consistency steps that correspond to deep neural network forward passes and numerical optimizations, respectively. The deep neural networks used for noise estimation can be either trained separately and in advance for different noise levels [96, 95, 7] or can be optimized together via end-to-end training of the complete processing pipeline [74, 1, 31]. During end-to-end training, the regularization parameter in equation 1.24 can be also tuned in the same way as the parameters of the deep neural network.

Loop unfolding can obtain high quality images that are comparable to the results from direct inference with a deep neural network. The method supports computing inverse images in real-time for easy-to-calculate forward operators such as in the case of MRI reconstruction [74], but is unsuited for real-time inversion of computationally-expensive forward operators because it necessitates to repeatedly evaluate the respective forward and adjoint operators in the data consistency steps of the algorithm. With a means of ensuring data consistency similar to variational methods, loop unrolling in principle allows to derive rigorous convergence guarantees. For real-world imaging applications, however, such guarantees are only of secondary importance, as it is not practical to iterate the unrolled network until convergence [1].

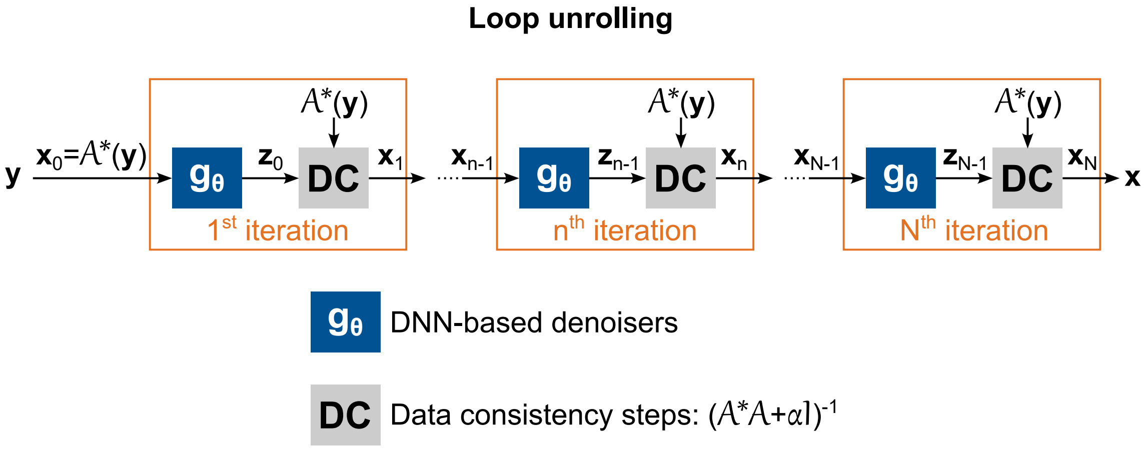


**Figure 1.6** Schematic methodology of deep-learning-based inversion using loop unrolling. The processing pipeline consists of alternating noise estimation and data consistency steps that correspond to deep neural network forward passes and numerical optimizations, respectively.

## 1.5 Application strategies of discriminative deep learning for MSOT imaging

In the following, we interlink the contents from the previous sections and discuss application strategies of discriminative deep learning for MSOT imaging. Thereby, we distinguish between two general directions of deep-learning-based improvements for MSOT: On the one hand, deep learning methods that improve the clinical applicability of MSOT by *accelerating* manual or computationally intensive parts of the *existing* imaging pipeline; and on the other hand, deep learning methods that rely on data-driven modeling to *improve* the image quality of MSOT *beyond* current state-of-the-art methods. We explain the two improvement strategies in detail in sections 1.5.1 and 1.5.2, respectively. Then, in section 1.5.3, we address the challenge of obtaining suitable training data to apply deep learning to MSOT.

### 1.5.1 Acceleration of manual and time-consuming tasks of the MSOT imaging pipeline

Discriminative deep learning can be used to accelerate time-consuming and manual parts of the MSOT imaging pipeline that impede a more widespread deployment of MSOT in clinical imaging. Several proof-of-concept studies have demonstrated the capability of MSOT to provide valuable clinical insights for a variety of diseases [17, 47, 69, 43, 67, 40, 69]. However, the data processing and analysis pipelines used in these studies are in most cases unsuitable for direct integration into clinical imaging routines, as they include manual tasks such as tuning the speed of sound used during image reconstruction, selecting the highest quality and most insightful scans, and identifying the regions of interest in reconstructed images. These tasks are tedious, time consuming, and require domain-specific expertise that is not covered by standard training of clinical staff. Furthermore, clinical application of MSOT imaging is complicated by state-of-the-art iterative model-based image reconstruction being too computationally expensive for real-time imaging. Instead, clinical studies must resort to low-quality images from backprojection reconstruction to find the best scan location and position during imaging and make on-site diagnoses.

Discriminative deep neural network models are well suited to automate and accelerate the reconstruction, processing, and analysis of optoacoustic images during clinical applications. Crucially, deep neural networks allow efficient evaluation of the data transformation learned during model training using modern graphics processing units. In addition, the process to be automated and accelerated provides an inherent method for obtaining reference data for model training.

In chapter 3, we introduce a deep learning approach to speed up MSOT imaging. The approach involves training a deep neural network to express model-based optoacoustic image reconstruction and thus output high quality optoacoustic images in real-time. Deep-learning-based automation and acceleration is also applicable to other parts of the MSOT imaging process. Therefore, the explanations above and from chapter 3 can serve as starting points for the development of further deep learning approaches to improve the versatility and clinical applicability of MSOT. In table 1.2, we compile an overview of possible discriminative deep learning applications to speed up computationally-slow and manual tasks of the optoacoustic imaging pipeline.

### 1.5.2 Deep-learning-based image quality improvements for MSOT

Discriminative deep learning can be also used to improve the image quality of MSOT beyond state-of-the-art analytical methods. Crucially, discriminative deep learning methods can access the topology and statistics of the ground truth training data to distinguish between real signals and noise, and to extrapolate from incompletely recordings. In contrast, analytical image reconstruction and

**Table 1.2** Discriminative deep learning applications to accelerate computationally-slow and manual tasks of the optoacoustic imaging pipeline.

| Task description | Benefits for clinical use | Related references |
|---|---|---|
| Accelerate iterative model-based reconstruction of optoacoustic initial pressure images. | Real-time MSOT imaging in clinical applications with state-of-the-art image quality. | See chapter 3. |
| Automate tuning of the speed of sound value used to reconstruct optoacoustic initial pressure images. | Obtain in-focus optoacoustic images ad hoc, reduce the required domain-specific expertise for clinical application. | [59, 38] |
| Automate evaluating the image quality of MSOT. | Monitor image quality in real-time, select optimal images ad hoc from a sequence of scans, reduce the required domain-specific expertise in clinical applications. | [2] |
| Automate the identifying, segmenting, and annotating anatomical structures and regions of interest in MSOT images. | Streamline and scale up the calculation of functional tissue parameters from MSOT images. | [10, 50] |

enhancement methods can only integrate prior knowledge in a rigid and generic way through regularization. A key challenge in applying deep learning to improve MSOT image quality is to generate appropriate training data. In general, only reference data from simulations are available for model training, as there is no easy way to determine the true chromophore concentration or initial pressure distributions for in vivo tissues. However, simulations often necessitate simplifying assumptions and can lead to the trained model generalizing imprecisely to in vivo data. In section 1.5.3, we discuss in more detail possible approaches to generate adequate training data for MSOT as well as general risks of data-driven modeling.

In prior works, several deep learning methods have been suggested to improve the image quality of MSOT [32, 26]. Table 1.3 provides an overview of previous deep-learning-based approaches aimed at the acoustic inverse problem of MSOT. Deep neural network models have been applied to enhance the bandwidth of recorded signals [30], recover full-view acquisition images from sparse-view measurements [14], reduce noise and sparse-view artifacts as part of deep-learning-based acoustic inversion [42, 51, 85, 21, 78, 28, 29], and remove reflection artifacts [3]. These studies provide valuable insights into the methodology and possible application strategies of deep-learning-based image enhancements for MSOT. However, existing methods are still in the proof-of-concept stage and are therefore unsuitable for direct use in clinical applications. Table 1.3 also summarizes individual shortcomings in all of the approaches listed.

In chapter 2, we introduce a deep-learning-based image enhancement method that falls into the category discussed in this section. The proposed methodology involves denoising recorded optoacoustic sinograms using a deep neural network trained on experimentally acquired ground truth noise and simulated ground truth signals. The presented denosing methodology affords significant improvements in morphological and multispectral optoacoustic image contrast.

**Table 1.3** Deep-learning-based image quality improvement approaches related to the acoustic inverse problem of MSOT.

| **Task description** | **Limitations for clinical use** | **Related references** |
|---|---|---|
| Enhance the bandwidth of recorded signals. | Oversimplistic experimental setup, no comprehensive validation based on in vivo data, only a full-view acquisition scheme is considered. | [30] |
| Recover from sparse-view measurements the initial pressure images corresponding to a full-view acquisition scheme. | Pre-clinical imaging setup, full-view ground-truth data acquisition is unavailable in the clinical context because human body parts like the arm, foot, torso, neck, or throat are too large to be illuminated and enclosed by transducers in a meaningful way. | [14] |
| Reconstruct initial pressure images and jointly correct for noise, finite transducer bandwidth, and sparse-view artifacts. | No comprehensive validation based on in vivo data, only rudimentary image quality when applied to isolated in vivo scans, sub-optimal and in part pre-clinical imaging setups. | [42, 51, 85, 21, 78, 28, 29, 33], also see discussion of these references in chapter 3. |
| Remove acoustic reflection artifacts in initial pressure images. | Overly-simplistic problem setting as only point absorbers are considered, no comprehensive validation based on in vivo data. | [3] |

Deep learning has also been applied to tackle the non-linear optical inverse problem related to MSOT, for which a comprehensive solution generally does not yet exist. Table 1.4 summarizes the scope and limitations of discriminative deep learning methods aimed at the optical inverse problem of MSOT. Deep neural networks have been proposed to estimate light fluence distributions, absorption coefficients, and chromophore concentrations from initial pressure images [6, 25, 62, 54]. These approaches perform reasonably well when using simulated data, but their practical applicability to clinical data is limited (see individual limitations in Table 1.4). Therefore, it is the subject of future work to improve these existing approaches and develop novel deep learning methodologies for the optical inversion of MSOT, with more focus on ensuring that developed solutions are applicable to in vivo data from clinical scanners.

**Table 1.4** Deep-learning-based image quality improvement approaches related to the optical inverse problem of MSOT.

| Task description | Limitations for clinical use | Related references |
|---|---|---|
| Infer the absorption coefficient distribution from a single initial pressure image. | Strongly ill-posed inverse problem since only the initial pressure image corresponding to one wavelength is considered, No comprehensive validation based on experimental data (in particular, no quantitative evaluation). | [54] |
| Infer the the oxygen saturation distribution from a set of initial pressure images obtained with different wavelengths. | No validation based on experimental data, simplified (pre-clinical) imaging setup with full-angle illumination and imaging depths of only about one centimeter. | [6] |
| Infer per pixel the oxygen saturation from a set of initial pressure measurements acquired with different wavelengths. | Strongly ill-posed inverse problem since the spatial context of a considered pixel is disregarded, no comprehensive validation based on experimental data. | [25] |
| Infer eigenfluence parameters for ad hoc spectral inversion within the eMSOT eigenspectra model. | Simplistic model of light fluence in tissue, only two chromophores are considered (oxygenated and deoxygenated hemoglobin), only evaluated for a pre-clinical MSOT imaging system with full-angle illumination and an imaging depth of about one centimeter. | [62] |

### 1.5.3 Selection of training data

The availability of training data is a key prerequisite for applying deep learning to improve the image quality and clinical applicability of MSOT. In the following, we discuss possible approaches for selecting suitable training data for MSOT, as well as risks associated with the selection of the training data.

#### Selection of ground-truth targets

First, we focus on the sub-problem of obtaining ground-truth targets alongside the corresponding input data. Ground truth targets act as the reference during deep neural network training, and their quality directly affects the performance of the trained model. Precise ground truth targets are by design available for deep learning techniques from section 1.5.1 that aim to speed up manual or computationally slow parts of the MSOT images pipeline. However, there is no simple method to obtain accurate ground truth values for the initial pressure, light fluence, or chromophore concentrations from in vivo tissue, as would be required for deep-learning-based image enhancement approaches from section 1.5.2. Therefore, such deep learning approaches to improve MSOT image quality must be based on synthesized training data.

The synthesis of the training data consists of two steps:

1. Distributions for the concentrations of chromophores or for the initial pressureare are defined. These distributions serve as ground truth references when training deep learning models.
2. Signals corresponding to the defined chromophore concentration or initial pressure distributions are simulated using a physics-based forward model of the imaging process. These simulated signals are used as input data when training deep learning models.

The distributions used in the first step of the simulation process can be generated using simple shapes such as circles and rectangles, handcrafted models of tissue-like structures, or reference images generated from other imaging modalities [26]. The optical forward imaging process of MSOT can be simulated using Monte Carlo methods [86, 20] or diffusion approximation [13]. The acoustic part of the optoacoustic forward process can be simulated using k-space pseudo-spectral methods or the forward models described in section 1.2.2.

### Risk of out-of-distribution samples

Generality of training data is another important consideration when employing deep neural network models. Deep neural network models take a data-driven approach to finding an appropriate data transformation for a specific task. Their ability to leverage information from the training data manifold during model optimization represents a key methodological advantage over analytical models. However, this data-driven mode of operation also carries the risk of reduced performance for out-of-distribution samples. Out-of-distribution samples refer to test data with features that are not contained in the training dataset. They are an intrinsic problem when using synthesized data, since forward simulations of the MSOT imaging process require in practice simplifying assumptions and are therefore likely to induce a domain-shift between the synthesized training and in vivo test data. In current research, such differences between synthesized training and experimental test data pose a key challenge for the applicability of deep learning methods for MSOT to in vivo data [42, 51, 85, 21, 78, 28, 29, 33, 6, 25, 54]. A possible approach to reduce the domain gap between synthesized training and in vivo test data is given by deep generative models such as generative adversarial networks. As presented in a recent study [73], deep generative models can be used to synthesize (more) realistic optoacoustic images, thus enabling the training of deep learning models with better applicability to in vivo data.

Furthermore, the widely varying characteristics of MSOT data pose another risk of out-of-distribution sampling. The appearance of MSOT images can change notably depending on the imaging setup, the targeted anatomy, and individual characteristics of the scanned person such as for example the skin color, body type or disease state. Therefore, composing a suitably general training dataset to account for all these variations represents a challenge regardless of whether the data is obtained experimentally or using simulations.

### Practical considerations

Finally, another consideration for selecting MSOT training data is how resource-intensive and convenient it is to obtain a sufficient amount of training data. In order to obtain in a practical manner specifically-trained deep learning models for new scanners or different imaging settings, the chosen methodology for obtaining training data must be reasonably flexible and applicable with manageable effort. From a practical point of view, model training using synthesized data is preferable, as it affords fully-automated training data generation. In contrast, model training using experimental data requires significantly more effort, since a cohort of participants must be recruited and scanned for each new scanner or other imaging setting. The two research papers included hereafter in chapters 2 and 3 have been designed with these considerations in mind and propose efficiently methodologies to generate a large training datasets based on utilizing a diverse collection of publicly available real-world image as initial pressure distributions.

# 2 Deep-learning-based electrical noise removal enables high spectral optoacoustic contrast in deep tissue

This chapter contains the research paper "Deep-learning-based electrical noise removal enables high spectral optoacoustic contrast in deep tissue" by Christoph Dehner, Ivan Olefir, Kaushik Basak Chowdhury, Dominik Jüstel, and Vasilis Ntziachristos. The paper has been published in the peer-reviewed journal IEEE Transactions on Medical Imaging, volume 41, issue 11, 2022 (see appendix A).

## 2.1 Summary

Electrical noise can severely reduce image contrast in optoacoustic tomography. Electrical noise can also limit the spatial and temporal resolution during optoacoustic tomography because it necessitates averaging over large tissue regions and multiple scans for reliable quantification. Corruptions by electrical noise are particularly detrimental for the imaging of deep tissue, where optoacoustic sensitivity is additionally decreased due to light fluence attenuation, and during multispectral imaging, for which scans at different wavelength are combined and thus also the therein contained distortions due to electrical noise accumulate.

Electrical noise in optoacoustic tomography arises from thermal effects (thermal noise) and electromagnetic interferences (parasitic noise caused by the electronics of the imaging system or the environment) and appears as an additive component in the recorded sinograms. Whereas thermal noise can be modeled as white Gaussian noise, parasitic noise entails complex spatio-temporal correlations and thus cannot be efficiently captured by an analytical model. Previously employed signal processing techniques have proven insufficient to remove the effects of electrical noise because they rely on simplified models that fail to capture complex characteristics of signal and noise. Moreover, they often involve time-consuming processing steps that are unsuited for real-time imaging applications.

In this work, we develop a discriminative deep learning approach to separate electrical noise from optoacoustic signals prior to image reconstruction. We reformulate the removal of electrical noise from optoacoustic sinograms as a probabilistic decomposition problem and derive, based on this formulation, a deep neural network model to simultaneously denoise an entire optoacoustic sinogram. Denoising the entire sinogram is a crucial design decision of the presented method, as it allows the deep neural network to capture spatio-temporal correlations within the true optoacoustic signals and within the parasitic noise, thus separating the two more efficiently. We train the deep neural network model using pure electrical noise sinograms acquired experimentally by scanning water at wavelengths with negligible absorption, and noise-free optoacoustic sinograms synthesized from general-feature images using an accurate forward physical model of the target scanner. The proposed deep learning approach obtains a denoised optoacoustic sinogram via one forward pass through a deep neural network and is thus suitable for real-time operation.

We validate the ability of the trained model to accurately remove electrical noise using synthetic signals as well as scans of phantom and the human breast. We demonstrate significant enhancements of morphological and spectral optoacoustic images, reaching 19% higher blood vessel contrast and localized spectral contrast at depths of about two centimeters for images acquired in vivo. In conclusion, the presented denoising framework considerably improves the imaging capabilities of MSOT and can enable detailed studies of endogenous biomarkers in deep tissue, such as breast vasculature or hemoglobin contrast inside a cancerous tumor.

## 2.2 Contribution

Dominik Jüstel came up with the initial idea of training a deep learning model on experimentally acquired electrical noise to improve the image quality of MSOT. Christoph Dehner developed the exact methodology and the computational framework of the presented deep-learning-based denoising approach, implemented the acoustic forward model of the employed MSOT scanner to synthesize noise-free optoacousic training sinograms, carried out the training of all deep learning models, applied the trained models to all considered test data, reconstructed the optoacoustic initial pressure images of the physical phantom and the clinical breast scans, and conducted all evaluations of the denoising approach based on optoacoustic sinograms and single-wavelength initial pressure images. Kaushik Basak Chowdhury experimentally acquired the electrical noise samples used during deep neural network training and provided a characterization of the total impulse response of the used MSOT system. Ivan Olefir conducted the evaluation of the denoising approach based on multispectral breast images using blind spectral unmixing via non-negative matrix factorization. Christoph Dehner wrote the manuscript with inputs from all authors. Dominik Jüstel and Vasilis Ntziachristos supervised the project.

# 3 A deep neural network for real-time optoacoustic image reconstruction with adjustable speed of sound

This chapter contains the research paper "A deep neural network for real-time optoacoustic image reconstruction with adjustable speed of sound" by Christoph Dehner, Guillaume Zahnd, Vasilis Ntziachristos, and Dominik Jüstel. The paper has been published in the peer-reviewed journal Nature Machine Intelligence, 2023 (see appendix B).

## 3.1 Summary

Real-time imaging is imperative for clinical imaging with multispectral optoacoustic tomography (MSOT). In handheld mode, image feedback in real-time is required to avoid hindering visio-tactile coordination, identify and localize relevant tissue structures using anatomical landmarks in their surroundings, and find the optimal transducer pose for the target region. Real-time optoacoustic imaging is also necessary to visualize dynamic pathophysiological changes associated with disease progression and enable in situ diagnoses.

The backprojection algorithm can reconstruct optoacoustic images in real-time but only with reduced quality. Backprojection images suffer from reduced spatial resolution and contrast as well as negative pixel values that invalidate a physical interpretation of the image as initial pressure distribution because the backprojection formula is based on over-simplified modeling assumptions of the imaging process and cannot compensate for the ill-posedness of the underlying inverse problem. Model-based reconstruction, on the other hand, delivers state-of-the-art optoacoustic images by incorporating an accurate physical model of the forward imaging process, introducing regularization to mitigate the ill-posedness of the inverse problem, and constraining the reconstructed initial pressure image to be non-negative. However, the advanced image quality provided by model-based reconstruction remains inaccessible during real-time imaging because the algorithm is iterative and computationally demanding.

Deep learning may afford faster reconstructions for real-time optoacoustic imaging. However, only synthesized data are available for model training, which is why existing approaches offer reduced reconstruction accuracy for in vivo data. In this work, we demonstrate that learning a well-posed reconstruction operator enables accurate generalization from synthesized training data to experimental test data. We present a deep-learning framework, termed DeepMB, to learn the model-based reconstruction operator and infer optoacoustic images with state-of-the-art quality in 31 ms per image. DeepMB takes as inputs an optoacoustic sinogram and a speed of sound value and infers thereof the corresponding optoacoustic image via one forward pass through a deep neural network. Deep neural network training is conducted with optoacoustic signals that are synthesized from real-world images, while using as ground-truth the optoacoustic images generated via model-based reconstruction of the corresponding signals. DeepMB is suitable for straightforward adoption into clinical routing because it supports to dynamic adjustments of the reconstruction speed of sound during imaging and is compatible with the data rates and image sizes of modern multispectral optoacoustic tomography scanners.

We evaluate DeepMB both qualitatively and quantitatively on a diverse dataset of in vivo images and demonstrate that the framework reconstructs optoacoustic images with similar quality to state-of-the-art iterative model-based reconstruction and at speeds enabling live imaging. In addition, we perform ablation studies to assess how the choice of training data and the encoding strategy for the input speed of sound affect the performance of the trained DeepMB model.

## 3.2 Contribution

Christoph Dehner and Dominik Jüstel came up with the initial idea of expressing the model-based optoacoustic reconstruction operator with a deep neural network to obtain state-of-the-art optoacoustic images in real-time. Christoph Dehner and Guillaume Zahnd contributed equally to this work: They developed the exact methodology and the computational framework for the deep-learning-based reconstruction approach, acquired all training and evaluation data, carried out all model trainings, performed all evaluation experiments, and wrote the manuscript with inputs from the two other authors. Dominik Jüstel and Vasilis Ntziachristos supervised the project.

# 4 Conclusion and outlook

Clinical translation is the ultimate goal for MSOT, but is hampered by degraded image quality due to measurement noise and inaccurate real-time image reconstruction methods. The presented work showcases and discusses how discriminative deep learning can mitigate these limitations and enable a more widespread use of MSOT in clinical imaging applications. Discriminative deep learning provides powerful means of deriving new solutions or accelerating existing solutions to inverse imaging problems because they can learn complex data transformations in a data-driven manner and efficiently evaluate these transformations using modern graphics processing units. Deep-learning-based removal of electrical noise from optoacoustic sinograms reveals rich morphological and spectral optoacoustic contrast at high resolution several centimeters deep in tissue. Deep-learning-based optoacoustic image reconstruction obtains initial pressure images with similar quality as state-of-the-art model-based reconstruction in 31 milliseconds per image, thus enabling real-time imaging in clinical applications with the highest image quality available.

In order to benefit from the deep learning methods developed in this work, they need to be integrated into clinical studies. The deep learning methodology to remove electrical noise from chapter 2 has already been taken up in a first study investigating the ability of multispectral optoacoustic tomography to examine the vascular environment and the morphology of peripheral nerves [40]. Thereby, improving image quality by removing electrical noise has turned out to be essential to visualize for the first time intraneural vessels in healthy nerves in vivo and to detect spectral optoacoustic contrast in the connective tissue of peripheral nerves, which can be related to the endogeneus contrast of hemoglobin and collagen. The deep-learning-based reconstruction method from chapter 3 (named "DeepMB") needs to be integrated into a clinical scanner to take advantage of the improved real-time imaging capabilities in clinical trials. This integration has already planned together with a MSOT device manufacturer (iThera Medical GmbH) and is currently being implemented in the further course of the research project "DeepOPUS" (funded by the Bavarian Ministry of Economic Affairs, Energy and Technology), in which context the initial development of DeepMB has already taken place.

The findings from this thesis may also serve as a basis for the development of solution approaches for the optical inverse problem of MSOT. While practical solutions suitable for clinical use exist for the acoustic inverse problem of MSOT - also through the two deep learning approaches presented in this work - the optical inverse problem of MSOT is currently still largely unsolved. However, the availability of a high-quality optoacoustic initial pressure images makes it possible to intensify research on optical inversion techniques for MSOT and to further develop existing methodologies, such as model-based inversion methods [55, 76, 93, 87] or approaches that combine spectral unmixing and statistical analysis techniques [40].

Another promising line of research to further improve imaging capabilities with MSOT are uncertainty estimation approaches. Model-based reconstruction, the deep-learning-based electrical noise removal presented in chapter 2, and DeepMB all obtain their respective outputs as point estimates (specifically, as maximum likelihood or maximum a posteriori estimates), and thus do not provide any means for modeling and quantifying uncertainty. However, uncertainty is an inevitable part of the optoacoustic imaging process because of noise, limited-angle acquisition, finite transducer bandwidth, and light fluence attenuation. Therefore, methodologies to (approximately) model full conditional posterior distributions during the optoacosutic inversion process are expected to further improve the reliability and accuracy of MSOT, especially for quantitative imaging. Previous studies have applied simple methods to estimate uncertainty during optoacoustic inversion process, such as restricting the probabilistic model to Gaussian priors [77] or approximating the posterior distributions using a truncated Taylor series around the maximum a posteriori estimate [66]. It is the subject of future research to propose, develop, and evaluate more sophisticated Bayesian methods to model a full conditional posterior distribution, such as for example variational inference approaches [44, 23].

Finally, the ongoing clinical integration of MSOT will bring additional uses of deep learning in the context of automated image analysis and interpretation. Currently, deep-learning-based segmentation, annotation, quality control, and disease prediction techniques for MSOT are not yet applicable on a larger scale due to small cohorts and too pronounced differences in the scanning setups and image qualities of MSOT systems. However, the ongoing technical development and standardization as well as a more widespread use of MSOT will inevitably pave the way for an increased need for such downstream analysis tasks. Overall, we are convinced that advanced deep-learning-based data processing methods will enable to utilize the unique imaging capabilities of MSOT in clinical applications.

# Acknowledgement

I would like to thank my supervisor Prof. Vasilis Ntziachristos for giving me the possibility to work on this project and for his support of my research.

Thanks to Dominik Jüstel for his guidance, trust, and support throughout this project and for always being available for questions.

Thanks to Guillaume Zahnd for the excellent cooperation in the DeepMB project and for his helpfulness, patience, and cheerfulness during all our interactions.

Thanks to my colleagues and friends Antonia Longo, Bego Rojas, Francesca Gasparin, Ivan Olefir, Jan Kukačka, Kaushik B. Chowdhury, Lukas Platz, Manuel Gehmeyr, Maximilian Bader, Philipp Haim, Sarah Franceschin, Sarkis Ter Martirosyan and Suhanyaa Nitkunanantharajah for helpful discussions about different topics related to my work and for their emotional support throughout this project.

Thanks to Andreas Hillmair, Bettina Lehman, Corinna Lieleg, Doris Bengel, Ines Baumgartner, Lidia Seidl, Manfred Türke, Ute Chambers, and Susanne Stern for their help in administrative matters and to Robert Wilson for his support and guidance related to scientific writing.

Finally, I thank my family and my friends for all the little and big things they did to support me.

# Bibliography

[1] H. K. Aggarwal, M. P. Mani, and M. Jacob. Modl: Model-based deep learning architecture for inverse problems. *IEEE Transactions on Medical Imaging*, 38(2):394–405, 2019.

[2] N. Akhlaghi, T. J. Pfefer, K. A. Wear, B. S. Garra, and W. C. Vogt. Multidomain computational modeling of photoacoustic imaging: verification, validation, and image quality prediction. *Journal of Biomedical Optics*, 24(12):121910, 2019.

[3] D. Allman, A. Reiter, and M. A. L. Bell. Photoacoustic source detection and reflection artifact removal enabled by deep learning. *IEEE Transactions on Medical Imaging*, 37(6):1464–1477, 2018.

[4] L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M. A. Fadhel, M. Al-Amidie, and L. Farhan. Review of deep learning: concepts, cnn architectures, challenges, applications, future directions. *Journal of Big Data*, 8(1):53, 2021.

[5] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, 2006.

[6] C. Cai, K. Deng, C. Ma, and J. Luo. End-to-end deep neural network for optical inversion in quantitative photoacoustic imaging. *Optics Letters*, 43(12):2752–2755, 2018.

[7] J. R. Chang, C.-L. Li, B. Póczos, B. Vijaya Kumar, and A. C. Sankaranarayanan. One network to solve them all — solving linear inverse problems using deep projection models. In *IEEE International Conference on Computer Vision (ICCV)*, pages 5889–5898, 2017.

[8] R. Chartrand and B. Wohlberg. Total-variation regularization with bound constraints. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 766–769, 2010.

[9] H. Chen, Y. Zhang, M. K. Kalra, F. Lin, Y. Chen, P. Liao, J. Zhou, and G. Wang. Low-dose ct with a residual encoder-decoder convolutional neural network. *IEEE Transactions on Medical Imaging*, 36(12):2524–2535, 2017.

[10] N.-K. Chlis, A. Karlas, N.-A. Fasoula, M. Kallmayer, H.-H. Eckstein, F. J. Theis, V. Ntziachristos, and C. Marr. A sparse deep learning approach for automatic segmentation of human vasculature in multispectral optoacoustic tomography. *Photoacoustics*, 20:100203, 2020.

[11] K. B. Chowdhury, J. Prakash, A. Karlas, D. Justel, and V. Ntziachristos. A synthetic total impulse response characterization method for correction of hand-held optoacoustic images. *IEEE Transactions Medical Imaging*, 39(10):3218–3230, 2020.

[12] K. B. Chowdhury, M. Bader, C. Dehner, D. Justel, and V. Ntziachristos. Individual transducer impulse response characterization method to improve image quality of array-based handheld optoacoustic tomography. *Optics Letters*, 46(1):1–4, 2021.

[13] B. Cox, J. G. Laufer, S. R. Arridge, and P. C. Beard. Quantitative spectroscopic photoacoustic imaging: a review. *Journal of Biomedical Optics*, 17(6):061202, 2012.

[14] N. Davoudi, X. L. Deán-Ben, and D. Razansky. Deep learning optoacoustic tomography with sparse data. *Nature Machine Intelligence*, 1(10):453–460, 2019.

[15] G. J. Diebold, T. Sun, and M. I. Khan. Photoacoustic monopole radiation in one, two, and three dimensions. *Physical Review Letters*, 67:3384–3387, 1991.

[16] L. Ding, X. L. Dean-Ben, and D. Razansky. Real-time model-based inversion in cross-sectional optoacoustic tomography. *IEEE Transactions on Medical Imaging*, 35(8):1883–1891, 2016.

[17] G. Diot, S. Metz, A. Noske, E. Liapis, B. Schroeder, S. V. Ovsepian, R. Meier, E. Rummeny, and V. Ntziachristos. Multispectral optoacoustic tomography (msot) of human breast cancer. *Clinical Cancer Research*, 23(22):6912–6922, 2017.

[18] W. Dong, L. Zhang, G. Shi, and X. Li. Nonlocally centralized sparse representation for image restoration. *IEEE Transactions on Image Processing*, 22(4):1620–1630, 2013.

[19] C. Errico, J. Pierre, S. Pezet, Y. Desailly, Z. Lenkei, O. Couture, and M. Tanter. Ultrafast ultrasound localization microscopy for deep super-resolution vascular imaging. *Nature*, 527(7579): 499–502, 2015.

[20] Q. Fang and D. A. Boas. Monte carlo simulation of photon migration in 3d turbid media accelerated by graphics processing units. *Optics Express*, 17(22):20178–20190, 2009.

[21] J. Feng, J. Deng, Z. Li, Z. Sun, H. Dou, and K. Jia. End-to-end res-unet based reconstruction algorithm for photoacoustic imaging. *Biomedical Optics Express*, 11(9):5321–5340, 2020.

[22] M. Figueiredo and R. Nowak. An em algorithm for wavelet-based image restoration. *IEEE Transactions on Image Processing*, 12(8):906–916, 2003.

[23] P. Frank, R. Leike, and T. A. Enßlin. Geometric variational inference. *Entropy*, 23(7), 2021.

[24] J. Frikel and E. T. Quinto. Artifacts in incomplete data tomography with applications to photoacoustic tomography and sonar. *SIAM Journal on Applied Mathematics*, 75(2):703–725, 2015.

[25] J. Gröhl, T. Kirchner, T. J. Adler, L. Hacker, N. Holzwarth, A. Hernández-Aguilera, M. A. Herrera, E. Santos, S. E. Bohndiek, and L. Maier-Hein. Learned spectral decoloring enables photoacoustic oximetry. *Scientific Reports*, 11(1):6565, 2021.

[26] J. Gröhl, M. Schellenberg, K. Dreher, and L. Maier-Hein. Deep learning for biomedical photoacoustic imaging: A review. *Photoacoustics*, 22:100241, 2021.

[27] S. Gu, Q. Xie, D. Meng, W. Zuo, X. Feng, and L. Zhang. Weighted nuclear norm minimization and its applications to low level vision. *International Journal of Computer Vision*, 121(2):183–208, 2017.

[28] S. Guan, A. A. Khan, S. Sikdar, and P. V. Chitnis. Limited-view and sparse photoacoustic tomography for neuroimaging with deep learning. *Scientific Reports*, 10(1):8510, 2020.

[29] M. Guo, H. Lan, C. Yang, J. Liu, and F. Gao. As-net: Fast photoacoustic reconstruction with multi-feature fusion from sparse data. *IEEE Transactions on Computational Imaging*, 8:215–223, 2022.

[30] S. Gutta, V. S. Kadimesetty, S. K. Kalva, M. Pramanik, S. Ganapathy, and P. K. Yalavarthy. Deep neural network-based bandwidth enhancement of photoacoustic data. *Journal of Biomedical Optics*, 22(11):116001, 2017.

[31] K. Hammernik, T. Klatzer, E. Kobler, M. P. Recht, D. K. Sodickson, T. Pock, and F. Knoll. Learning a variational network for reconstruction of accelerated mri data. *Magnetic Resonance in Medicine*, 79(6):3055–3071, 2018.

[32] A. Hauptmann and B. Cox. Deep learning in photoacoustic tomography: current approaches and future directions. *Journal of Biomedical Optics*, 25(11):112903, 2020.

[33] A. Hauptmann, F. Lucka, M. Betcke, N. Huynh, J. Adler, B. Cox, P. Beard, S. Ourselin, and S. Arridge. Model-based learning for accelerated, limited-view 3-d photoacoustic tomography. *IEEE Transactions on Medical Imaging*, 37(6):1382–1393, 2018.

[34] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[35] Y. Hoshi and Y. Yamada. Overview of diffuse optical tomography and its clinical applications. *Journal of Biomedical Optics*, 21(9):091312, 2016.

[36] Y. Hu and M. Jacob. Higher degree total variation (hdtv) regularization for image recovery. *IEEE Transactions on Image Processing*, 21(5):2559–2571, 2012.

[37] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pages 448–456, 2015.

[38] S. Jeon, W. Choi, B. Park, and C. Kim. A deep learning-based model that reduces speed of sound aberrations for improved in vivo photoacoustic imaging. *IEEE Transactions on Image Processing*, 30:8773–8784, 2021.

[39] K. H. Jin, M. T. McCann, E. Froustey, and M. Unser. Deep convolutional neural network for inverse problems in imaging. *IEEE Transactions on Image Processing*, 26(9):4509–4522, 2017.

[40] D. Jüstel, H. Irl, F. Hinterwimmer, C. Dehner, W. Simson, N. Navab, G. Schneider, and V. Ntziachristos. Spotlight on nerves: Portable multispectral optoacoustic imaging of peripheral nerve vascularization and morphology. *Advanced Science*, 10(19):2301322, 2023.

[41] S. Khan, J. Huh, and J. C. Ye. Adaptive and compressive beamforming using deep learning for medical ultrasound. *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, 67(8):1558–1572, 2020.

[42] M. Kim, G. S. Jeng, I. Pelivanov, and M. O'Donnell. Deep-learning image reconstruction for real-time photoacoustic system. *IEEE Transactions on Medical Imaging*, 39(11):3379–3390, 2020.

[43] F. Knieling, C. Neufert, A. Hartmann, J. Claussen, A. Urich, C. Egger, M. Vetter, S. Fischer, L. Pfeifer, A. Hagel, C. Kielisch, R. S. Gortz, D. Wildner, M. Engel, J. Rother, W. Uter, J. Siebler, R. Atreya, W. Rascher, D. Strobel, M. F. Neurath, and M. J. Waldner. Multispectral optoacoustic tomography for assessment of crohn's disease activity. *New England Journal of Medicine*, 376 (13):1292–1294, 2017.

[44] J. Knollmüller and T. A. Enßlin. Metric gaussian variational inference. *arXiv:1901.11033*, 2019.

[45] R. A. Kruger, P. Liu, Y. R. Fang, and C. R. Appledorn. Photoacoustic ultrasound (paus)—reconstruction tomography. *Medical Physics*, 22(10):1605–1609, 1995.

[46] P. Kuchment and L. Kunyansky. *Mathematics of Photoacoustic and Thermoacoustic Tomography*, pages 817–865. Springer New York, 2011.

[47] J. Kukačka, S. Metz, C. Dehner, A. Muckenhuber, K. Paul-Yuan, A. Karlas, E. M. Fallenberg, E. Rummeny, D. Jüstel, and V. Ntziachristos. Image processing improvements afford second-generation handheld optoacoustic imaging of breast cancer patients. *Photoacoustics*, 26:100343, 2022.

[48] L. A. Kunyansky. Explicit inversion formulae for the spherical mean radon transform. *Inverse Problems*, 23(1):373–383, 2007.

[49] G. Kutyniok, W.-Q. Lim, and R. Reisenhofer. Shearlab 3d: Faithful digital shearlet transforms based on compactly supported shearlets. *ACM Transactions on Mathematical Software*, 42(1), 2016.

[50] B. Lafci, E. Merčep, S. Morscher, X. L. Deán-Ben, and D. Razansky. Deep learning for automatic segmentation of hybrid optoacoustic ultrasound (opus) images. *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, 68(3):688–696, 2021.

[51] H. Lan, D. Jiang, C. Yang, F. Gao, and F. Gao. Y-net: Hybrid deep learning image reconstruction for photoacoustic tomography in vivo. *Photoacoustics*, 20:100197, 2020.

[52] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[53] D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems*, volume 13, 2000.

[54] J. Li, C. Wang, T. Chen, T. Lu, S. Li, B. Sun, F. Gao, and V. Ntziachristos. Deep learning-based quantitative optoacoustic tomography of deep tissues in the absence of labeled experimental data. *Optica*, 9(1):32–41, 2022.

[55] S. Li, B. Montcel, Z. Yuan, W. Liu, and D. Vray. Multigrid-based reconstruction algorithm for quantitative photoacoustic tomography. *Biomedical Optics Express*, 6(7):2424–2434, 2015.

[56] A. C. Luchies and B. C. Byram. Deep neural networks for ultrasound beamforming. *IEEE Transactions on Medical Imaging*, 37(9):2010–2021, 2018.

[57] B. Luijten, R. Cohen, F. J. de Bruijn, H. A. W. Schmeitz, M. Mischi, Y. C. Eldar, and R. J. G. van Sloun. Adaptive ultrasound beamforming using deep learning. *IEEE Transactions on Medical Imaging*, 39(12):3967–3978, 2020.

[58] X. Mao, C. Shen, and Y.-B. Yang. Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

[59] T. P. Matthews, J. Poudel, L. Li, L. V. Wang, and M. A. Anastasio. Parameterized joint reconstruction of the initial pressure and sound speed distributions for photoacoustic computed tomography. *SIAM Journal on Imaging Sciences*, 11(2):1560–1588, 2018.

[60] V. Ntziachristos and D. Razansky. Molecular imaging by means of multispectral optoacoustic tomography (msot). *Chemical Reviews*, 110(5):2783–2794, 2010.

[61] I. Olefir, S. Tzoumas, H. Yang, and V. Ntziachristos. A bayesian approach to eigenspectra optoacoustic tomography. *IEEE Transactions on Medical Imaging*, 37(9):2070–2079, 2018.

[62] I. Olefir, S. Tzoumas, C. Restivo, P. Mohajerani, L. Xing, and V. Ntziachristos. Deep learning-based spectral unmixing for optoacoustic imaging of tissue oxygen saturation. *IEEE Transactions on Medical Imaging*, 39(11):3643–3654, 2020.

[63] G. Ongie, S. Biswas, and M. Jacob. Convex recovery of continuous domain piecewise constant images from nonuniform fourier samples. *IEEE Transactions on Signal Processing*, 66(1):236–250, 2018.

[64] G. Ongie, A. Jalal, C. A. Metzler, R. G. Baraniuk, A. G. Dimakis, and R. Willett. Deep learning techniques for inverse problems in imaging. *IEEE Journal on Selected Areas in Information Theory*, 1(1):39–56, 2020.

[65] S. Prahl. Assorted spectra. `https://omlc.org/spectra/`. Accessed: 04.01.2023.

[66] A. Pulkkinen, B. T. Cox, S. R. Arridge, J. P. Kaipio, and T. Tarvainen. Estimation and uncertainty quantification of optical properties directly from the photoacoustic time series. In *Photons Plus Ultrasound: Imaging and Sensing*, 100643N, 2017.

[67] A. P. Regensburger, L. M. Fonteyne, J. Jungert, A. L. Wagner, T. Gerhalter, A. M. Nagel, R. Heiss, F. Flenkenthaler, M. Qurashi, M. F. Neurath, N. Klymiuk, E. Kemter, T. Frohlich, M. Uder, J. Woelfle, W. Rascher, R. Trollmann, E. Wolf, M. J. Waldner, and F. Knieling. Detection of collagens by multispectral optoacoustic tomography as an imaging biomarker for duchenne muscular dystrophy. *Nature Medicine*, 25(12):1905–1915, 2019.

[68] W. Ren, S. Liu, H. Zhang, J. Pan, X. Cao, and M.-H. Yang. Single image dehazing via multiscale convolutional neural networks. In *Computer Vision – ECCV 2016*, pages 154–169, 2016.

[69] W. Roll, N. A. Markwardt, M. Masthoff, A. Helfen, J. Claussen, M. Eisenblatter, A. Hasenbach, S. Hermann, A. Karlas, M. Wildgruber, V. Ntziachristos, and M. Schafers. Multispectral optoacoustic tomography of benign and malignant thyroid disorders: A pilot study. *Journal of Nuclear Medicine*, 60(10):1461–1466, 2019.

[70] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241, 2015.

[71] A. Rosenthal, V. Ntziachristos, and D. Razansky. Model-based optoacoustic inversion with arbitrary-shape detectors. *Medical Physics*, 38(7):4285–4295, 2011.

[72] A. Rosenthal, V. Ntziachristos, and D. Razansky. Acoustic inversion in optoacoustic tomography: A review. *Current Medical Imaging Reviews*, 9(4):318–336, 2013.

[73] M. Schellenberg, J. Gröhl, K. K. Dreher, J.-H. Nölke, N. Holzwarth, M. D. Tizabi, A. Seitel, and L. Maier-Hein. Photoacoustic image synthesis with generative adversarial networks. *Photoacoustics*, 28:100402, 2022.

[74] J. Schlemper, J. Caballero, J. V. Hajnal, A. N. Price, and D. Rueckert. A deep cascade of convolutional neural networks for dynamic mr image reconstruction. *IEEE Transactions on Medical Imaging*, 37(2):491–503, 2018.

[75] W. Simson, M. Paschali, N. Navab, and G. Zahnd. Deep learning beamforming for sub-sampled ultrasound data. In *IEEE International Ultrasonics Symposium (IUS)*, pages 1–4, 2018.

[76] Y. Sun, E. S. Sobel, and H. Jiang. Quantitative three-dimensional photoacoustic tomography of the finger joints: an in vivo study. *Journal of Biomedical Optics*, 14(6):064002, 2009.

[77] J. Tick, A. Pulkkinen, and T. Tarvainen. Image reconstruction with uncertainty quantification in photoacoustic tomography. *The Journal of the Acoustical Society of America*, 139(4):1951, 2016.

[78] T. Tong, W. Huang, K. Wang, Z. He, L. Yin, X. Yang, S. Zhang, and J. Tian. Domain transform network for photoacoustic tomography from limited-view and sparsely sampled data. *Photoacoustics*, 19:100190, 2020.

[79] S. Tzoumas, N. Deliolanis, S. Morscher, and V. Ntziachristos. Unmixing molecular agents from absorbing tissue in multispectral optoacoustic tomography. *IEEE Transactions Medical Imaging*, 33(1):48–60, 2014.

[80] S. Tzoumas, A. Rosenthal, C. Lutzweiler, D. Razansky, and V. Ntziachristos. Spatiospectral denoising framework for multispectral optoacoustic imaging based on sparse signal representation. *Medical Physics*, 41(11):113301, 2014.

[81] S. Tzoumas, A. Nunes, I. Olefir, S. Stangl, P. Symvoulidis, S. Glasl, C. Bayer, G. Multhoff, and V. Ntziachristos. Eigenspectra optoacoustic tomography achieves quantitative blood oxygenation imaging deep in tissues. *Nature Communications*, 7(1):12121, 2016.

[82] R. J. G. van Sloun, R. Cohen, and Y. C. Eldar. Deep learning in ultrasound imaging. *Proceedings of the IEEE*, 108(1):11–29, 2020.

[83] R. J. G. van Sloun, O. Solomon, M. Bruce, Z. Z. Khaing, H. Wijkstra, Y. C. Eldar, and M. Mischi. Super-resolution ultrasound localization microscopy through deep learning. *IEEE Transactions on Medical Imaging*, 40(3):829–839, 2021.

[84] J. Vonk, J. Kukačka, P. Steinkamp, J. de Wit, F. Voskuil, W. Hooghiemstra, M. Bader, D. Jüstel, V. Ntziachristos, G. van Dam, and M. Witjes. Multispectral optoacoustic tomography for in vivo detection of lymph node metastases in oral cancer patients using an egfr-targeted contrast agent and intrinsic tissue contrast: A proof-of-concept study. *Photoacoustics*, 26:100362, 2022. ISSN 2213–5979.

[85] D. Waibel, J. Gröhl, F. Isensee, T. Kirchner, K. Maier-Hein, and L. Maier-Hein. Reconstruction of initial pressure from limited view photoacoustic images using deep learning. In *Photons Plus Ultrasound: Imaging and Sensing*, 104942S, 2018.

[86] L. Wang, S. L. Jacques, and L. Zheng. Mcml—monte carlo modeling of light transport in multi-layered tissues. *Computer Methods and Programs in Biomedicine*, 47(2):131–146, 1995.

[87] Y. Wang, J. He, J. Li, T. Lu, Y. Li, W. Ma, L. Zhang, Z. Zhou, H. Zhao, and F. Gao. Toward whole-body quantitative photoacoustic tomography of small-animals with multi-angle light-sheet illuminations. *Biomedical Optics Express*, 8(8):3778–3795, 2017.

[88] S. J. Wright, R. D. Nowak, and M. A. T. Figueiredo. Sparse reconstruction by separable approximation. *IEEE Transactions on Signal Processing*, 57(7):2479–2493, 2009.

[89] J. Xia and L. V. Wang. Small-animal whole-body photoacoustic tomography: A review. *IEEE Transactions on Biomedical Engineering*, 61(5):1380–1389, 2014.

[90] J. Xie, L. Xu, and E. Chen. Image denoising and inpainting with deep neural networks. In *Advances in Neural Information Processing Systems*, volume 25, 2012.

[91] M. Xu and L. V. Wang. Universal back-projection algorithm for photoacoustic computed tomography. *Physical Review E*, 71(1 Pt 2):016706, 2005.

[92] M. Xu and L. V. Wang. Photoacoustic imaging in biomedicine. *Review of Scientific Instruments*, 77(4):041101, 2006.

[93] Z. Yuan, Q. Wang, and H. Jiang. Reconstruction of optical absorption coefficient maps of heterogeneous media by photoacoustic tomography coupled with diffusion equation based regularized newton method. *Optics Express*, 15(26):18076–18081, 2007.

[94] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE Transactions on Image Processing*, 26(7): 3142–3155, 2017.

[95] K. Zhang, W. Zuo, S. Gu, and L. Zhang. Learning deep cnn denoiser prior for image restoration. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2808–2817, 2017.

[96] L. Zhang and W. Zuo. Image restoration: From sparse and low-rank priors to deep priors [lecture notes]. *IEEE Signal Processing Magazine*, 34(5):172–179, 2017.

[97] B. Zhu, J. Z. Liu, S. F. Cauley, B. R. Rosen, and M. S. Rosen. Image reconstruction by domain-transform manifold learning. *Nature*, 555(7697):487–492, 2018.

# Appendix A: Deep-learning-based electrical noise removal enables high spectral optoacoustic contrast in deep tissue

Research paper by Christoph Dehner, Ivan Olefir, Kaushik Basak Chowdhury, Dominik Jüstel, and Vasilis Ntziachristos, as published in the journal IEEE Transactions on Medical Imaging, volume 41, issue 11, 2022.

# Deep-Learning-Based Electrical Noise Removal Enables High Spectral Optoacoustic Contrast in Deep Tissue

Christoph Dehner, Ivan Olefir, Kaushik Basak Chowdhury, Dominik Jüstel, and Vasilis Ntziachristos

***Abstract*—Image contrast in multispectral optoacoustic tomography (MSOT) can be severely reduced by electrical noise and interference in the acquired optoacoustic signals. Previously employed signal processing techniques have proven insufficient to remove the effects of electrical noise because they typically rely on simplified models and fail to capture complex characteristics of signal and noise. Moreover, they often involve time-consuming processing steps that are unsuited for real-time imaging applications. In this work, we develop and demonstrate a discriminative deep learning approach to separate electrical noise from optoacoustic signals prior to image reconstruction. The proposed deep learning algorithm is based on two key features. First, it learns spatiotemporal correlations in both noise and signal by using the entire optoacoustic sinogram as input. Second, it employs training on a large dataset of experimentally acquired pure noise and synthetic optoacoustic signals. We validated the ability of the trained model to accurately remove electrical noise on synthetic data and on optoacoustic images of a phantom and the human breast. We demonstrate significant enhancements of morphological and spectral optoacoustic images reaching 19% higher blood vessel contrast and localized spectral contrast at depths of more than 2 cm for images acquired in vivo. We discuss how the proposed denoising framework is applicable to clinical multispectral optoacoustic tomography and suitable for real-time operation.**

***Index Terms*—Breast cancer, denoising, dynamic MSOT, photoacoustic tomography, signal decomposition, sinogram.**

## I. Introduction

ELECTRICAL noise is a key source of signal corruption in optoacoustic imaging and arises due to thermal effects (thermal noise) and from electromagnetic interference (parasitic noise); the latter possibly generated by the optoacoustic system itself or the environment [1]. While thermal noise can be modeled as white Gaussian noise [2], parasitic noise entails complex spatiotemporal correlations, and thus cannot be efficiently captured by an analytical model [1]. Both thermal and parasitic noise cause artifacts in reconstructed optoacoustic images that severely decrease morphological and spectral contrast. Whereas shielding hardware can suppress some parasitic noise, this solution is device-specific and often incomplete [1]. Therefore, additional computational denoising techniques, which are applicable across platforms, are needed to remove both parasitic and thermal noise.

Noise in optoacoustic images hinders the detection and identification of fine structures in tissue, particularly as the signal-to-noise ratio (SNR) in the data decreases with increasing depth [3]. Besides the reduction of image contrast, noise also challenges the quantification and spectral un-mixing of optoacoustic images acquired at multiple wavelengths [4], [5]. In particular, corrupted spectral information decreases the spatial resolution of multispectral optoacoustic tomography (MSOT) because it necessitates averaging the spectra obtained from large tissue regions for reliable quantification [6]–[8]. Efficient noise reduction algorithms are therefore critical for improving the performance of spectral optoacoustics.

Frequency filtering using band-pass filters cannot adequately separate thermal and parasitic noise from optoacoustic signals because the frequency content of optoacoustic signals and noise overlap significantly. For this reason, data averaging and regularization methods have been more commonly considered to minimize the effects of electrical noise from optoacoustic tomographic images [3], [7]–[12]. While data averaging effectively reduces zero-mean electrical noise, combining multiple

Manuscript received 29 April 2022; accepted 30 May 2022. Date of publication 3 June 2022; date of current version 27 October 2022. This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 862811 (RSENSE), from the European Research Council (ERC) under Grant Agreement No. 694968 (PREMSOT), and by the Deutsche Forschungsgemeinschaft (DFG), Sonderforschungsbereich-824 (SFB-824), subproject A1. *(Corresponding authors: Vasilis Ntziachristos; Dominik Jüstel.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Local Ethics Committee of the Technical University of Munich under Application No. 27/18 S.

Christoph Dehner, Ivan Olefir, Kaushik Basak Chowdhury, and Vasilis Ntziachristos are with the Institute of Biological and Medical Imaging, Helmholtz Zentrum Munchen, 85764 Neuherberg, Germany, and also with the School of Medicine, Chair of Biological Imaging, Technical University of Munich, 80333 München, Germany (e-mail: christoph.dehner@helmholtz-muenchen.de; ivan.olefir@tum.de; kaushikbchowdhury@gmail.com; bioimaging.translatum@tum.de).

Dominik Jüstel is with the Institute of Biological and Medical Imaging and the Institute of Computational Biology, Helmholtz Zentrum Munchen, 85764 Neuherberg, Germany, and also with the School of Medicine, Chair of Biological Imaging, Technical University of Munich, 80333 München, Germany (e-mail: dominik.juestel@helmholtz-muenchen.de).

Digital Object Identifier 10.1109/TMI.2022.3180115

acquisitions reduces imaging rates and increases vulnerability to motion artifacts, particularly in clinical applications or when using a handheld system. Regularization of model-based reconstructions may decrease the effects of electrical noise, but this reduction is either limited to the noise characteristics captured by the regularization functional or realized at the cost of data fidelity, thereby corrupting the meaningfulness of the reconstructed image. Furthermore, iterative model-based reconstruction is computationally intensive, and therefore unsuitable for applications that require real-time feedback [4], [13], [14]. Another approach to reduce noise is based on sparse (typically Wavelet-based) representations for optoacoustic signals [1], [15], [16]. Noise is assumed to distort the sparsity properties of optoacoustic signals, which enables its removal, e.g., via thresholding techniques. However, the denoising performance of such methods is limited by their reliance on oversimplified models of noise and optoacoustic signals.

Recently, discriminative deep neural network models have achieved state-of-the-art performance on general image denoising tasks, like Gaussian denoising, deblocking, super-resolution, inpainting, and dehazing [17]–[20]. These deep neural network models capture the required data transformations for denoising in a data-driven way by utilizing large ground truth training datasets. In this way, discriminative deep neural networks are capable of more accurate, robust, and fast denoising than traditional methods that rely on rigid analytical models because they can adjust to complex data characteristics during training and are evaluated in real-time on modern GPUs. Similar deep-learning-based approaches have also been applied to remove reconstruction [21], [22] and reflection artifacts [23], [24] from optoacoustic images, and to enhance contrast of images acquired with low energy illumination elements such as LED-based systems [25]–[27].

In this work, we examine whether discriminative deep learning can separate thermal and parasitic noise from optoacoustic signals. We show that the modeling capabilities and the computational efficiency of a deep neural network facilitates denoising in optoacoustic tomography that is both precise enough to remove noise with complex characteristics and fast enough for real-time imaging applications. We design a deep neural network model to simultaneously denoise the entire sinogram of an optoacoustic scan, i.e. the complete dataset acquired from all transducers. Entire sinogram denoising enables the network to capture spatiotemporal correlations within both parasitic noise and true optoacoustic responses, and thus more efficiently separate the two. Exploiting the independence of electrical noise and optoacoustic signals, we train the network on a large ground truth dataset of experimentally acquired pure noise and synthetic optoacoustic sinograms. We validate that the model removes thermal and parasitic noise from both synthetic sinograms and optoacoustic images of a phantom. We lastly apply the trained model to clinical MSOT images of breast tissue and show significant enhancements in morphological and spectral contrast. The improved contrast allows for tissue components to be more accurately localized and quantified and yields more meaningful correlations with the spectra of known absorbers in tissue, thereby increasing access to endogenous biomarkers in deep tissue, such as breast vasculature or hemoglobin contrast inside a cancerous tumor.

## II. Methods

In the following, we formalize our methodology for removing electrical noise from optoacoustic sinograms. First, we reformulate the denoising problem as a decomposition task. Based on this formulation, we derive a discriminative deep learning framework for denoising optoacoustic sinograms. At the end of the chapter, we explain the experiments that we use to validate this approach.

### A. Denoising via Decomposition

In this section, we formalize the rationale for reformulating denoising of optoacoustic sinograms as a decomposition task and conclude that 1) electrical noise in optoacoustic tomography is an additive component that is independent from the optoacoustic signals, and 2) for denoising, an acquired optoacoustic sinogram $s$ can be decomposed into a component $s_{\mathrm{OA}}$ containing the true optoacoustic sinogram and an electrical noise component $s_{\mathrm{noise}}$: $s = s_{\mathrm{OA}} + s_{\mathrm{noise}}$.

An optoacoustic scan at a fixed excitation wavelength consists of measured optoacoustic pressure signals $s_d\,[t]$, indexed by time samples $t \in [1, 2, \ldots, N_{\mathrm{samples}}]$ and transducer locations $d \in [1, 2, \ldots, N_{\mathrm{transducers}}]$. The collection of signals recorded at all the transducers compose the sinogram $s\,[d, t] := s_d[t]$ of the scan. We model the measured optoacoustic sinograms probabilistically as samples $s$ of a random field $S$, i.e. a collection of random variables that model all recorded signals of a sinogram $s_d\,[t]$, $t \in [1, 2, \ldots, N_{\mathrm{samples}}]$, $d \in [1, 2, \ldots, N_{\mathrm{transducers}}]$. In the remainder of this paper, we will denote random fields with capital letters, and specific samples of random fields with lower case letters. The main assumption that leads to the formulation of denoising as a decomposition problem is that $S$ is the sum of two independent random fields $S_{\mathrm{OA}}$ and $S_{\mathrm{noise}}$, which describe the signal content due to optoacoustic responses and electrical noise, respectively. This assumption is justified by the fact that electrical noise in optoacoustic tomography typically originates from common system thermal noise and electromagnetic interference that is not influenced by the optoacoustic signal [1]. The probability distributions underlying $S_{\mathrm{OA}}$ and $S_{\mathrm{noise}}$ are denoted by $P_{\mathrm{OA}}$ and $P_{\mathrm{noise}}$:

$$\begin{aligned} &S = S_{\mathrm{OA}} + S_{\mathrm{noise}} \\ &\text{with } S_{\mathrm{OA}} \sim P_{\mathrm{OA}} \text{ and } S_{\mathrm{noise}} \sim P_{\mathrm{noise}} \text{ independent.} \end{aligned} \tag{1}$$

Optoacoustic scans at different wavelengths are modeled as independent realizations of $S$ because the noise is caused by the electronics of the imaging system that are not affected by the wavelength switching of the laser. In summary, isolating the noise-free optoacoustic sinogram $s_{\mathrm{OA}}$ given $s$ is equivalent to decomposing $s = s_{\mathrm{OA}} + s_{\mathrm{noise}}$ into its two components $s_{\mathrm{OA}}$ and $s_{\mathrm{noise}}$.

### B. Deep-Learning-Based Denoising Framework

Next, we step-by-step derive a deep-learning-based denoising framework that can decompose optoacoustic sinograms into their signal and noise components. In short, we train a deep neural network to infer the electrical noise component from a noisy input sinogram using experimentally acquired pure noise sinograms and synthetic optoacoustic sinograms.

To solve the decomposition problem introduced in subsection A, we need access to the distributions $P_{\text{OA}}$ and $P_{\text{noise}}$. However, both random fields $S_{\text{OA}}$ and $S_{\text{noise}}$ are non-homogeneous and anisotropic with intricate spatial and temporal correlations due to the physics underlying the signals and the fact that electrical noise in optoacoustic systems often contains complex parasitic noise (Fig. 2c) [1]. As a result, accurate explicit models for the complex distributions of optoacoustic signals and electrical noise $P_{\text{OA}}$ and $P_{\text{noise}}$ are difficult to obtain. We therefore present a data-driven approach that allows us to rely on the empirical distributions of $P_{\text{OA}}$ and $P_{\text{noise}}$ via sampling of $S_{\text{OA}}$ and $S_{\text{noise}}$. We first explain sample acquisition and then elaborate on our methodology for solving the decomposition task.

Because of the independence of $S_{\text{OA}}$ and $S_{\text{noise}}$ and the wavelength independence of $S_{\text{noise}}$, electrical noise can be directly measured in the absence of any absorbers that would emit optoacoustic responses. We thus obtained samples of the electrical noise distribution $S_{\text{noise}}$ of the test system by immersing the scanner in a water tank and measuring from 700 to 790 nm, where light absorption in water is negligible (i.e., the optical absorption coefficients of water at 700 – 790 nm are between 0.006 $\text{cm}^{-1}$ and 0.026 $\text{cm}^{-1}$, whereas for example the absorption coefficients of oxygenated hemoglobin and fat at 800 nm and 930 nm are 4.4 $\text{cm}^{-1}$ and 13 $\text{cm}^{-1}$, respectively [28]–[30]). Note that in theory the effects of water absorption on the acquired noise samples may be further decreased by interrupting the optical path between the laser and the imaging probe during the noise acquisitions, or by acquiring noise samples only at the wavelength with the lowest water absorption coefficient from the range 700 – 790 nm. However, in practice such adjustments were not essential, as the obtained electrical noise samples did not correlate with the respectively used wavelength from the range 700 – 790 nm (which confirms that water absorption is indeed negligible in the whole range), and as the presented denoising framework facilitated accurate denoising with the current noise acquisition setup.

Acquiring samples of $S_{\text{OA}}$ in an experimental setup is a laborious and time-consuming process that requires averaging multiple scans of the same location to remove electrical noise. Additionally, patient or operator movement impedes accurate averaging. Therefore, instead of experimentally acquiring noise-free optoacoustic sinograms, we generated samples of $P_{\text{OA}}$ via simulation by applying an accurate acoustic forward model of the scanner [31], [32] to publicly available images from the PASCAL VOC2012 dataset [33], a diverse collection of over 17 000 images covering a large range of features. Utilizing these images as underlying initial pressure distributions in the simulations enables us to account for a broad range of potential features in optoacoustic sinograms and should yield a good approximation of the empirical distribution of $P_{\text{OA}}$. In addition, the general scope of the training data ensures that the denoising model is universally and with uniform performance applicable to arbitrary scans acquired by the considered system.

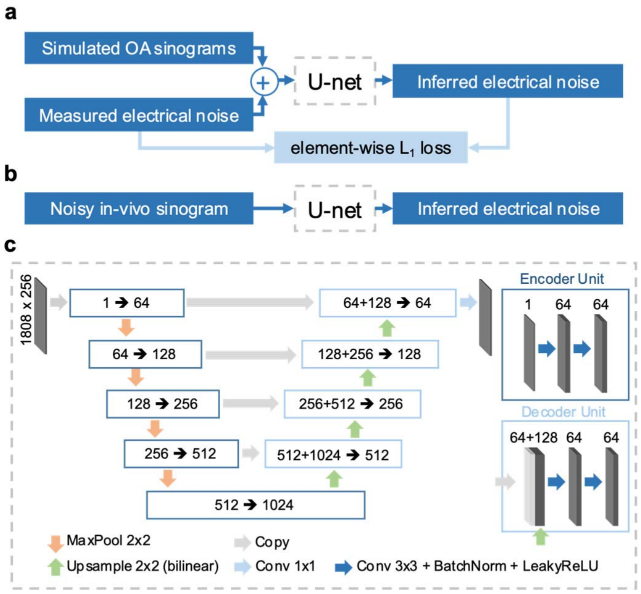


Fig. 1. Discriminative deep learning framework for denoising optoacoustic sinograms. a) Training setup of the method. The network was trained iteratively using simulated samples from the optoacoustic sinogram distribution of the test system and experimentally acquired samples of the electrical noise distribution of the system. b) Evaluation setup of the method. The trained neural network can infer the electrical noise from a noisy input sinogram. Subtracting the inferred noise from the input sinogram yields the denoised sinogram. c) U-net architecture of the deep neural network.

Using the samples from $S_{\text{OA}}$ and $S_{\text{noise}}$, we utilized a U-net-like deep neural network [34] to solve the decomposition task at hand. Fig. 1 depicts the deep-learning-based approach. Fig. 1a shows the training setup. We iteratively trained the network on randomly selected pairs of samples $s_{\text{OA}}$ and $s_{\text{noise}}$ from $S_{\text{OA}}$ and $S_{\text{noise}}$ to infer the noise component ($s_{\text{noise}}$) from a noisy input sinogram $s = s_{\text{OA}} + s_{\text{noise}}$. In this way, the network is optimized to adopt the complex characteristics of $P_{\text{OA}}$ and $P_{\text{noise}}$. Fig. 1b depicts the evaluation setup. Once trained, we can use the neural network to infer electrical noise from noisy input scans.

## C. Experiments

As a test system for the denoising algorithm, we used a custom prototype of an MSOT Acuity Echo handheld scanner (iThera Medical GmbH, Munich, Germany). The system is equipped with a tunable laser that illuminates tissue with laser pulses of ∼8 ns duration with an energy of 16 mJ and a repetition rate of 25 Hz. The ultrasound detector of the system consists of 256 piezoelectric transducers with a central frequency of 4 MHz, which are equidistantly placed on a circular arc with a radius of 6 cm and an angular coverage of 145°. Ultrasound signals are recorded with a sampling frequency of 40 MHz. Inside the imaging probe, heavy water with a speed of sound of approximately 1397 m/s is used as coupling medium. Fig. 2 provides an overview of the imaging system and its output. Fig. 2a shows the

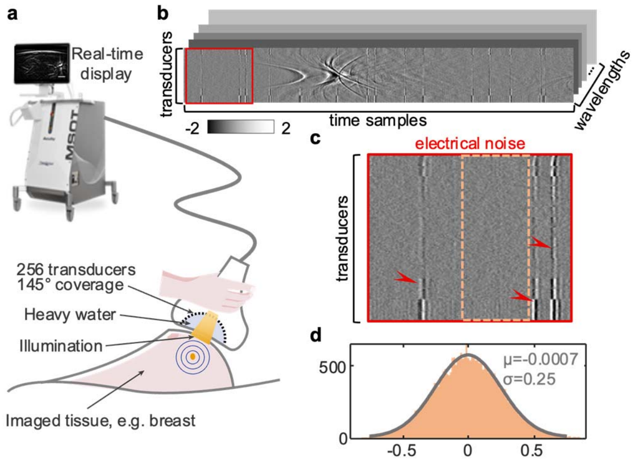


Fig. 2. Overview of the handheld MSOT system and its output used to evaluate the proposed denoising method. a) Illustration of the scanning procedure using the handheld imaging probe of the test system. b) Data layout of a measured multispectral stack of sinograms. The depicted sinogram shows the recorded signals during a representative scan of a human breast lesion at 960 nm. c) Magnification of the marked signals in b, which were recorded prior to responses from tissue and thus are predominately comprised of electrical noise. d) Histogram and fitted Gaussian distribution ($R^2 = 99.5\%$) for parts of the electrical noise with visually low amounts of parasitic noise (signals marked in c with the dashed rectangle) illustrating the characterization of the thermal noise of the system.

scanning procedure using the handheld imaging probe of the system. Fig. 2b illustrates the data layout of a multispectral stack acquired by the imaging system. A multispectral stack consists of 28 sinograms recorded at wavelengths from 700 – 970 nm in steps of 10 nm. Fig. 2c shows electrical noise from a representative optoacoustic in vivo scan. We observed that electrical noise in the system consists of two additive components: normally distributed thermal noise with a mean of zero and a standard deviation in the range of 0.2 – 0.3 and complex parasitic noise that is presumably caused by the switching-mode power supply of the system (examples marked with red arrows in Fig. 2c). Fig. 2d illustrates the characterization of thermal noise of the system based on parts of experimentally acquired electrical noise samples with visually low amounts of electrical noise and confirms that the thermal noise can be modeled as white Gaussian noise. The presented characterization of the thermal noise of the system nevertheless remains an approximation as the parasitic and thermal components of electrical noise cannot be completely separated.

We first evaluated the ability of the proposed deep learning framework to remove the combination of Gaussian thermal and complex parasitic electrical noise observed in the test system. We trained and evaluated a deep neural network on experimentally acquired samples of the electrical noise distribution $S_{\text{noise}}$ of the system and simulated samples of the optoacoustic signal distribution $S_{\text{OA}}$ (denoted as Dataset-EN). Next, we applied the trained denoising network to measurements of a physical phantom (denoted as Dataset-Ph). The arrangement of the phantom is shown in the inlay in Fig. 4a. Two plastic tubes with inner diameters of 3.0 mm and 0.86 mm and outer diameters of 3.2 mm and 1.52 mm were filled with ink and imaged cross-sectionally to simulate absorbers of different sizes and at different depths. These tubes were immersed into two layers of agar of slightly different densities mixed with Intralipid (6 ml 20% emulsion / 100 ml water) to mimic light scattering and small variations of the speed of sound distribution in biological tissue. Additionally, a copper plate was integrated into the arrangement as a reference that can be seen in both optoacoustic and ultrasound images.

To evaluate the denoising performance of the framework on in vivo scans, we subsequently applied the trained deep neural network to 81 multispectral optoacoustic scans of human breast cancer lesions (denoted as Dataset-BC). These scans were acquired in a study that was approved by the local ethics committee of the Technical University of Munich (27/18 S). All participants gave written informed consent upon recruitment.

Lastly, we evaluated the ability of the trained network to adapt to changes of the hardware configuration and of environmental conditions such as humidity and temperature of the used system, which might alter the amounts of thermal or parasitic electrical noise. For that, we applied the trained model to optoacoustic signals that were corrupted by a combination of measured electrical noise sinograms scaled with a $\text{factor}_{\text{EN}} \in \{0, 0.5, 1, \ldots, 3\}$, and white Gaussian noise with standard deviation $\sigma_{\text{GN}} \in \{0, 0.2, 0.4, \ldots, 2\}$ (denoted as Dataset-EN+GN). A summary of the four datasets is given in Table I.

### *D. Data Pre-Processing and Network Training*

We band-pass filtered all recorded signals from 500 kHz – 10 MHz to remove signals outside the bandwidth of the transducers and reduce low frequency responses that would otherwise dominate the contrast in reconstructed optoacoustic images. Additionally, all signals were slightly cropped in the time domain to remove filtering artifacts at the signal boundaries and to make the number of signal samples divisible by 16, as required by the chosen neural network architecture, leading to 1808 time samples for each of the 256 detectors per scan.

A detailed illustration of the proposed neural network is given in Fig. 1c. We adopted the U-Net neural network architecture [34] with a depth of 5 layers and a width of 64 channels, and designed the network to infer only the electrical noise $s_{\text{noise}}$ from a noisy input sinogram $s$ to minimize the necessary expressiveness of the network [18]. The network was trained for 300 epochs using the L1 norm of the difference of inferred and ground truth noise as loss functional, and the ADAM optimizer [35] with batch size = 1 and momentum parameters $\beta_1 = 0.5$ and $\beta_2 = 0.999$. The learning rate was set to 0.0001 and was linearly decreased to zero in the last 50 epochs. To accelerate the learning process, we scaled all input data of the neural network by a constant factor of 0.004 to achieve a signal range of $[-1; 1]$. After having passed the neural network, all signals were rescaled to the original range. After training, the final model was selected based on the minimal loss on a validation split of the dataset. One training

TABLE I
STRUCTURE AND SIZE OF THE FOUR DATASETS USED IN THIS PAPER

| Name | Description | Train split | Val. split | Test split |
|---|---|---|---|---|
| Dataset-EN | Simulated noise-free optoacoustic sinograms | 3000 sinograms | 590 sinograms | 629 sinograms |
|  | Measurements of pure electrical noise sinograms | 2110 sinograms | 590 sinograms | 629 sinograms |
| Dataset-Ph | Scans of a phantom | - | - | 28 sinograms in 1 multispectral stack |
| Dataset-BC | In vivo scans of human breast lesions | - | - | 2268 sinograms in 81 multispectral stacks |
| Dataset-EN+GN | Simulated noise-free optoacoustic sinograms | - | - | 629 sinograms |
|  | Measurements of pure electrical noise sinograms + simulated white Gaussian noise | - | - | (7 x 11) x 629 sinograms of measured electrical noise sinograms that were scaled with factor$_{\text{EN}}$ ∈ $\{0, 0.5, 1, \ldots, 3\}$ and additionally augmented by additive white Gaussian noise with a standard deviation $\sigma_{\text{GN}} \in \{0, 0.2, 0.4, \ldots, 2\}$ |

process took approximately three days on an NVIDIA GeForce RTX2070 GPU.

## E. Signal-to-Noise Ratio

For Dataset-EN and Dataset-GN, we used the signal-to-noise ratio (SNR), i.e. the ratio of signal power and noise power, to quantify the noise levels in the signals before and after denoising. We calculated the power $P(s)$ of a whole sinogram $s[d, t], d \in [1, 2, \ldots, N_{\text{transducers}}], t \in [1, 2, \ldots, N_{\text{samples}}]$ as

$$P(s) := \frac{1}{N_{\text{transducers}} N_{\text{samples}}} \sum\nolimits_{d=1}^{N_{\text{transdicers}}} \sum\nolimits_{t=1}^{N_{\text{samples}}} s[d, t]^2. \tag{2}$$

Based on equation 2, the SNR of a sinogram $s$ with ground truth noise $s_{\text{noise}}$ and inferred noise $s'_{\text{noise}}$ is defined as

$$\text{SNR} := 10 \log_{10} \left( \frac{P\left(s - s_{\text{noise}}\right)}{P\left(s_{\text{noise}} - s'_{\text{noise}}\right)} \right) dB. \tag{3}$$

Setting $s'_{\text{noise}} = 0$ yields the SNR before denoising; setting $s'_{\text{noise}}$ to the output of the trained network yields the SNR after denoising.

Since computing the SNR requires direct access to the ground truth noise $s_{\text{noise}}$ of a scan, the metric cannot be directly transferred to the in vivo scans of Dataset-BC. We therefore defined an alternative metric, $\text{SNR}_{\text{mean}}$, that enabled us to approximate the ground truth electrical noise for the in vivo scans from Dataset-BC by considering the per-pixel mean sinogram amplitudes across $N_{\text{scans}}$ different scans of the dataset, $s^{(1)}, s^{(2)}, \ldots, s^{(N_{\text{scans}})}$, $\langle|s|\rangle := \frac{1}{N_{\text{scans}}} \sum_{n=1}^{N_{\text{scans}}} |s^{(n)}|$.

$$\text{SNR}_{\text{mean}} := 10 \log_{10} \left( \frac{P_{\langle|s|\rangle - \langle|s_{\text{noise}}|\rangle}}{P_{\langle|s_{\text{noise}}|\rangle - \langle|\, s'_{\text{noise}}|\rangle}} \right) dB. \tag{4}$$

We approximated the ground truth electrical noise $\langle|s_{\text{noise}}|\rangle$ from the first 100 averaged time samples of all scans in Dataset-BC $\langle|s[\cdot, t']|\rangle, t' \in [1, 2, \ldots, 100]$:

$$\langle|s_{\text{noise}}[d, t]|\rangle \approx \frac{1}{100} \sum_{t'=1}^{100} \langle|s[d, t']|\rangle \quad \text{for } t \in [1, 2, \ldots, N_{\text{samples}}]. \tag{5}$$

Note that equation 5 yields a meaningful approximation of $\langle|s_{\text{noise}}[d, t]|\rangle$ for two reasons: First, the 100 signals recorded at the beginning of a scan do not contain optoacoustic responses but mostly electrical noise because they originate from the coupling medium inside the imaging probe. Second, we observed from the electrical noise sinograms in Dataset-EN that for the used test system, $\langle|s_{\text{noise}}[d, t]|\rangle$ is constant over time. Thus, an estimation of electrical noise based on a subset of time samples of the sinogram (i.e. the first 100 time samples in equation 4) is applicable to all time steps $t \in [1, 2, \ldots, N_{\text{samples}}]$. Furthermore, sinograms from Dataset-BC were cropped to the first 1732 recorded signal samples before evaluating the $\text{SNR}_{\text{mean}}$, as subsequent signals originate from outside the designated field of view of the scans and contain strong reflections and filtering artifacts.

$\text{SNR}_{\text{mean}}$ can approximate the SNR of in vivo images, for which the true amount of electrical noise is unknown. However, the metric may overestimate the SNR after denoising because the per-pixel mean amplitudes of the predicted noise $\langle|s'_{\text{noise}}|\rangle$ is subtracted in the denominator of equation 4, disregarding the possibility that the predicted and the ground truth noise have different signs: $\langle|s_{\text{noise}}|\rangle - \langle|s'_{\text{noise}}|\rangle = \langle|s_{\text{noise}} - s'_{\text{noise}}|\rangle$ only if $\forall 1 \ldots n \ldots N_{\text{scans}}\colon \text{sgn}\left(s_{\text{noise}}^{(n)'}\right) = \text{sgn}\left(s_n^{(n)}\right)$. An empirical comparison of $\text{SNR}_{\text{mean}}$ and SNR on Dataset-EN (for which ground truth noise samples are also available) showed that $\text{SNR}_{\text{mean}}$ correctly estimated the average true SNR before denoising ($\text{SNR}_{\text{mean}} = 9.6$ dB, avg. SNR $= 9.3$ dB) and overestimated the SNR after denoising by approximately 6 dB ($\text{SNR}_{\text{mean}} = 26.5$ dB, avg. SNR $= 20.3$ dB). The offset of the $\text{SNR}_{\text{mean}}$ after denoising is however smaller than the reported $\text{SNR}_{\text{mean}}$ improvements of

20 – 22 dB of the presented denoising method for Dataset-BC (see section III).

### *F. Evaluation on Reconstructed Images*

To evaluate the effects of the denoising method visually and quantitatively on optoacoustic images, we reconstructed the initial pressure $p_0$ of all breast scans in Dataset-BC, both with and without denoising the recorded sinograms with the trained neural network, using a model-based reconstruction algorithm [31], [32]. We added two regularization terms to address the two main causes of the ill-posedness of the inverse problem: simple Tikhonov regularization to mitigate limited view noise and Laplacian-based regularization to mitigate sub-resolution noise.

$$p_0 := \arg\min_{\mathrm{p}\geq 0} \|Mp - s\|_2^2 + \lambda_1 \|p\|_2^2 + \lambda_2 \|\Delta p\|_2^2. \quad (6)$$

The reconstructed optoacoustic images are of the size 400 × 400 pixels and correspond to FOVs of 3.99 cm × 3.99 cm. We denote the obtained datasets of reconstructed MSOT breast images as $D_{\text{original}}$ and $D_{\text{denoised}}$. Additional reconstructions were also obtained using backprojection reconstruction [36], [37].

We quantified the effects of the denoising method in the reconstructed images by calculating the contrast resolution and the contrast-to-noise ratio of blood vessels. For that, we manually segmented blood vessels in the images, as well as background ROIs from the surroundings of all segmented vessels. The segmentations were based on scans at 870 nm, where blood contrast is at a maximum. The background areas were chosen so as not to overlap with regions below and above strong absorbers, which are affected by limited view artifacts. We chose vessels of different sizes and at different depths to obtain general estimates for the blood contrast in the dataset. Fig. 3f shows the segmented regions for a representative scan. Based on these segmentations, the contrast resolution of a scan with mean intensities $I_{\text{vessels}}$ and $I_{\text{background}}$ in its respective vessel and background ROIs is defined as

$$\text{CR} := \frac{I_{\text{vessels}} - I_{\text{background}}}{I_{\text{vessels}} + I_{\text{background}}}, \quad (7)$$

and the contrast-to-noise ratio of a scan with standard deviation $\sigma_{\text{background}}$ in its background ROIs is defined as

$$\text{CNR} := 10 \log_{10} \left( \frac{I_{\text{vessels}} - I_{\text{background}}}{\sigma_{\text{background}}} \right) dB. \quad (8)$$

Negative values in backprojection images were set to zero before calculating the contrast resolution.

To evaluate the effects of denoising on the spectral contrast of MSOT, we applied blind spectral unmixing via non-negative matrix factorization (NMF) [38] to the multispectral optoacoustic images from $D_{\text{original}}$ and $D_{\text{denoised}}$ and compared the variety and biological interpretability of obtained spectral decompositions. For each of the two datasets (consisting each of 400 × 400 × 81 = 12960000 acquired spectra), we obtained a spectral decomposition into 10 non-negative spectral components $H$ (size 10 × 28) and corresponding non-negative unmixing coefficients $W$ (size 12 960 000 × 10).

$$(W, H) := \arg\min_{(W,H)\geq 0} \frac{1}{2} \|S - WH\|_F^2 + \lambda_1 \left( \|W\|_1 + \|H\|_1 \right) + \frac{1}{2} \lambda_F \left( \|W\|_F^2 + \|H\|_F^2 \right), \quad (9)$$

where $S$ (size 12 960 000 x 28) denotes all spectra of the dataset, $\|M\|_F := \left( \sum_{i,j} m_{i,j}^2 \right)^{\frac{1}{2}}$ denotes the Frobenius norm, $\|M\|_1 := \sum_{i,j} |m_{i,j}|$ denotes the entrywise $L^1$-norm of a matrix $M = \left( m_{i,j} \right)_{i,j}$, and $M \geq 0$ refers to an entrywise inequality. The entrywise $L^1$-regularization was chosen to promote a maximally sparse decomposition of the spectra, guided by the fact that the spectral contrast of biological tissue is dominated by a small number of abundant chromophores. The number of spectral components chosen was purposefully greater than the number of different chromophores in tissue to also extract variants of the chromophores' absorption spectra that are perceived because of spectral coloring (i.e., distortions of the perceived absorption spectra of chromophores in deeper tissue by light absorption of chromophores in superficial tissue layers). The specific number of components and regularization parameters $\lambda_1 = 50.1$ and $\lambda_F = 50.1$ were selected via parameter space exploration and meaningfulness of the resulting spectral components. Furthermore, the residual norm

$$r_{\text{NMF}} := \|S - WH\|_F^2 / \|S\|_F^2 \quad (10)$$

was evaluated for each of the two NMF runs to quantify the accuracy of the obtained unmixing solution.

## III. Results

The proposed deep learning framework for denoising optoacoustic sinograms significantly improves the SNR of both simulated and in vivo data in real-time. The average inference time of the employed deep neural network was 9 milliseconds. Based on the resulting increased quality of the optoacoustic signal data, the denoising method enables improved optoacoustic image contrast and spectral unmixing performance. In the following, we report the detailed findings in the signal, image, and spectral domains.

### *A. Denoising Performance in the Signal Domain*

Optoacoustic signals and electrical noise are both complex broadband signals whose characteristics overlap significantly. The presented data-driven approach can disentangle these overlaps by accessing and separating the data manifolds of signal and noise in sinograms. We observed significant reductions in noise levels, both visually and quantitatively, upon application of the denoising method to sinograms that were corrupted by a combination of Gaussian thermal and complex parasitic electrical noise (Dataset-EN, Dataset-BC, and Dataset-EN+GN with factor$_{\text{EN}} > 0$), as well as to sinograms that were corrupted only by Gaussian noise (Dataset-EN+GN with factor$_{\text{EN}} = 0$). Fig. 3 summarizes these results. We begin by visually inspecting the effects of the denoising method for a representative scan of a breast lesion in Fig. 3a-f. Fig. 3a

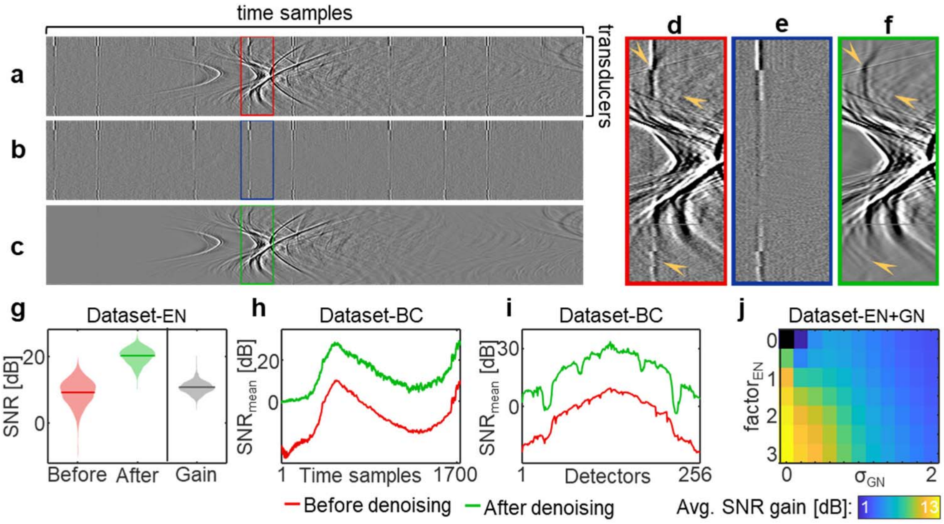


Fig. 3. Evaluation of the proposed denoising approach in the signal domain. a) Noisy sinogram from a representative scan of a human breast lesion. b) Electrical noise component inferred by the neural network. c) Denoised sinogram obtained by subtracting b from a. d-f) Magnifications of the marked areas in a-c. g-j) Quantitative evaluation of the denoising performance. g) Comparison of the SNR distributions in simulated optoacoustic sinograms that are distorted by electrical noise before and after denoising. The mean gain is 10.9 *dB*. h-i) Evaluation of in vivo scans of human breast lesions. h) Mean SNR ($\text{SNR}_{\text{mean}}$) of individual time samples. The average increase is 20.8 *dB*. i) Individual $\text{SNR}_{\text{mean}}$ of all detectors. The average increase is 22.4 *dB*. j) Average SNR gains ("SNR after denoising – SNR before denoising") of the trained model for optoacoustic signals that were corrupted by a combination of measured electrical noise sinograms scaled with $\text{factor}_{\text{EN}} \in \{0, 0.5, 1, \dots, 3\}$, and white Gaussian noise with standard deviations $\sigma_{\text{GN}} \in \{0, 0.2, 0.4, \dots, 2\}$. One of the detectors (no. 61) was defective and excluded from the plots in g-j.

shows the noisy sinogram before denoising. Because of the radial nature of wave propagation and the circular shape of the used imaging probe, optoacoustic responses appear as bow-shaped structures in the sinogram. Also note that optoacoustic signals from a spherical target at the center of the circular transducer array appear as bow-shaped structures because of refraction at the interface between the coupling medium inside the imaging probe (heavy water with a speed of sound of 1397 m/s) and the imaged tissue (speed of sound typically in the range 1450 – 1550 m/s). The sinogram is distorted by additive electrical noise that is composed of zero-mean Gaussian noise and complex noise artefacts with strong spatiotemporal correlations (also shown in Fig. 2c). Fig. 3b depicts the electrical noise component inferred by the trained neural network, which demonstrates the network's ability to extract electrical noise. Finally, Fig. 3c shows the denoised sinogram, which was obtained by subtracting the inferred noise from the recorded sinogram. Fig. 3d-f depict enlargements of identical temporal sections of the images in Fig. 3a-c, highlighting the fine features of the optoacoustic signals that are exposed upon removal of electrical noise (yellow arrows).

Fig. 3g-j provide an in-depth quantitative analysis of the network's denoising performance. These results confirmed the ability of the network to consistently remove electrical noise with high accuracy from both synthetic and in vivo optoacoustic sinograms. Fig. 3g compares the distributions of SNRs within the test split of Dataset-EN before and after denoising. Application of the denoising method to the sinograms in Dataset-EN resulted in an average improvement in SNR of 10.9 dB, with improvements for individual sinograms ranging from 4.6 dB to 20.0 dB. After the neural network was trained and tested on Dataset-EN, we applied it to denoise scans of breast lesions (Dataset-BC) to demonstrate its applicability to in vivo data. Fig. 3h shows a plot of the mean $\text{SNR}_{\text{mean}}$ of these individual time samples from Dataset-BC before and after application of the network, which indicates a time-independent increase in $\text{SNR}_{\text{mean}}$ of approximately 20.8 dB after denoising. The uniformity of the increase in $\text{SNR}_{\text{mean}}$ demonstrates that the trained neural network can extract electrical noise both from strong optoacoustic responses in superficial tissue (time samples 400 – 700), as well as from signals deeper in tissue, which have lower amplitudes due to light fluence attenuation (time samples 1100 – 1400).

Furthermore, we demonstrated the ability of the method to compensate for the variations in parasitic electrical noise within the transducer array of the test system (see Fig. 2c for details) to confirm the applicability of the trained network to in vivo scans. For that, we calculated the $\text{SNR}_{\text{mean}}$ for Dataset-BC individually for all transducer elements, rather than for the whole sinograms, before and after denoising. As shown in Fig. 3i, applying the trained network to the breast scans from Dataset-BC improves the $\text{SNR}_{\text{mean}}$ at all transducers by an average of 22.4 dB in a near uniform manner. Note that the transducers at the boundaries of the detector probe have lower SNRs than the central transducers due to the probe layout partially shielding the outermost transducers from arriving acoustic waves. The lower $\text{SNR}_{\text{mean}}$ values at transducers 30-43, 79-87, 167-175, 213-227 result from acoustic noise waves propagating along the transducer array, which corrupts the ground truth noise estimation used to calculate the $\text{SNR}_{\text{mean}}$ (see Equation 5). These noise waves depend on the imaged tissue, and therefore cannot be removed by the neural network.

Thus far, we have demonstrated the ability of the denoising method to accurately isolate the electrical noise of the test system. Next, we investigated whether the trained model can also adopt electrical noise with altered levels of parasitic and thermal noise components, e.g. caused by changes in hardware configurations or environmental conditions such as humidity and temperature. We applied the trained model to optoacoustic signals that were corrupted with augmented electrical noise sinograms containing, in comparison to the measured electrical noise used during training, up to three times the amount of parasitic noise and approximately up to four times the amount of thermal noise (see Dataset-EN+GN in Table I). Fig. 3j summarizes the average denoising performance of the trained network, which afforded improvements in SNR for all tested combinations of parasitic and thermal noise components. These results demonstrate that the trained neural network can also generalize to previously unseen amounts of electrical noise and thus facilitate robust denoising in real-world imaging applications, where the amounts of electrical noise may change over time.

### B. Denoising Enables High Contrast in Optoacoustic Images

Thus far, we have demonstrated the ability of the denoising network to isolate and remove electrical noise in optoacoustic sinograms. In this section, we analyze the effects of denoising on reconstructed optoacoustic images. First, we ensure that the denoising network successfully removes noise artifacts without distorting any true optoacoustic image structures using optoacoustic images of a phantom. Subsequently, we evaluate the improved image contrast due to denoising in a clinical dataset of scans of human breast lesions, to show the potential for improved diagnostic capability of optoacoustic tomography.

We utilized a model-based inversion algorithm to reconstruct optoacoustic images (i.e. initial spatial pressure distributions) from all scans in the datasets Dataset-Ph and Dataset-BC, both with and without denoising the recorded sinograms with the trained neural network. Fig. 4a-e illustrate the qualitative improvements to selected images upon application of the neural network. Fig. 4a shows an optoacoustic image of a phantom at 700 nm, reconstructed from a noisy sinogram. Zero-mean Gaussian noise in the recorded sinogram reduces the overall contrast in the image, whereas parasitic noise leads to ring artifacts that obscure potentially relevant image features. The arrangement of the phantom is shown in the inlay of Fig. 4a. Fig. 4b depicts the same optoacoustic image reconstructed from a denoised sinogram, demonstrating that the neural network can significantly reduce both the background noise and the ring artifacts. Fig. 4c plots the difference between Figs. 4a and b, which yields artifacts and background noise but no real structures, emphasizing the ability of the network to accurately identify and isolate noise in optoacoustic images. Fig. 4d and e show the optoacoustic images of a malignant breast tumor at 870 nm. The denoised image in Fig. 4e appears significantly richer in contrast than the original image in Fig. 4d and contains structures that are not visible prior to the denoising. To highlight the clinical relevance of the improved contrast, note that in Fig. 4e, the optoacoustic contrast inside the tumor core regions (outlined in blue) is separated from the noise that dominates these regions in Fig. 4d. Furthermore, we reconstructed the scans from Dataset-BC with the backprojection algorithm and confirmed that the presented denoising method also achieves visible reductions in the background noise and the ring artifacts for these images.

Next, we evaluated the contrast resolution (CR) of blood vessels in optoacoustic images reconstructed via model-based inversion from Dataset-BC to quantify the enhancement capabilities of the trained neural network in the image domain. Blood vessels and background ROIs were first manually segmented, as depicted in Fig. 4f, and used to calculate the blood contrast resolution. The distributions of contrast resolution in Dataset-BC before and after denoising are compared in Fig. 4g, which shows an average improvement of 0.083 with a range of 0.003 to 0.55 for individual images. As shown in Fig. 4h, the average improvement in blood contrast resolution is consistent across all wavelengths, demonstrating the network's ability to remove noise, independent of the varying strength of individual absorbers across the accessible spectrum. The denoising capabilities of the presented denoising method were also confirmed when evaluating the contrast-to-noise ratio of blood vessels in the images from model-based inversion, and the contrast resolution and contrast-to-noise ratio of blood vessels in images from backprojection reconstruction, with average improvements of 1.7 dB, 0.043, and 2.6 dB, respectively.

### C. Deep-Learning-Based Denoising Enables High Spectral Contrast

A further central finding of this work is the ability of the presented denoising approach to significantly improve spectral contrast in MSOT, i.e. the differentiation of chromophores based on their absorption spectra. We found that upon application of the denoising method to the MSOT scans from Dataset-BC, the dominant absorbers in breast tissue – hemoglobin, lipids, and water – are more accurately identified and localized. To analyze the spectral contrast, we applied blind spectral unmixing via non-negative matrix factorization (NMF) to the original and denoised breast images and decomposed each of the two datasets into 10 spectra and corresponding unmixing coefficients. Note that unlike linear unmixing based on the reference absorption spectra of chromophores in tissue, NMF finds both the spectra and unmixing coefficients in a data driven way and thus extracts variants of the reference spectra that consider effects from spectral coloring.

Fig. 5 compares the spectral contrast of the original and denoised MSOT breast images from Dataset-BC. In Fig. 5a-c, we show the NMF spectra obtained from the original (Fig. 5a) and from the denoised (Fig. 5b) data next to the reference absorption spectra of the most prominent chromophores in tissue (oxygenated and deoxygenated hemoglobin, water, and lipids, see Fig. 5c). The spectra derived from the original data show a significant number of sharp peaks (spectra no. 3, 5, 7, 8, 9, 10) attributable to ring artifacts from parasitic noise, rather than specific absorbers in tissue. In contrast, the spectra

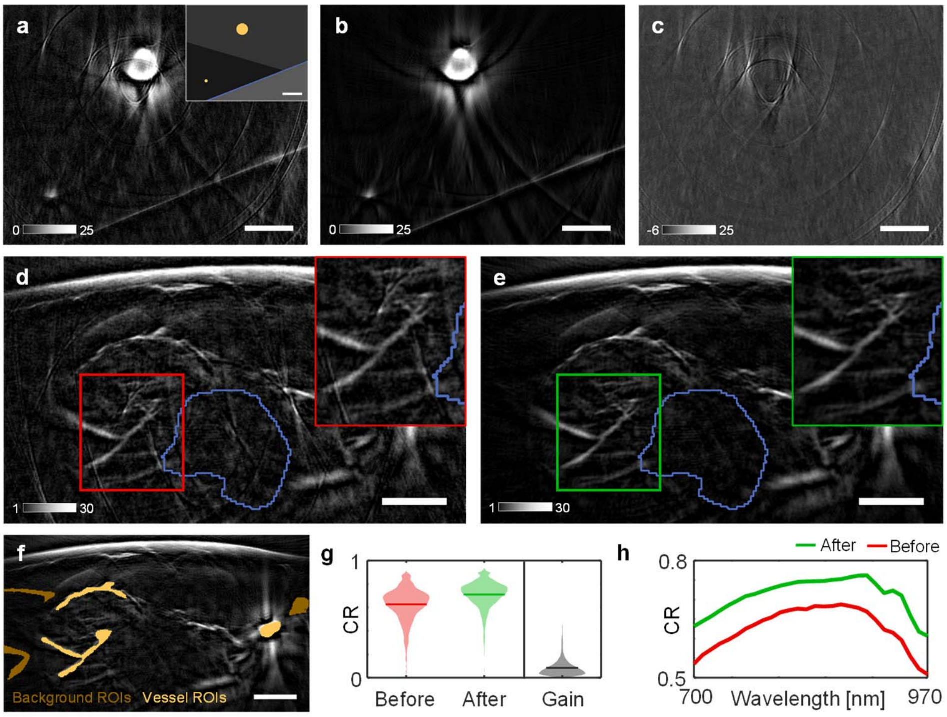


Fig. 4. Demonstration of improved image quality in denoised scans of a phantom and of human breast lesions. a) Optoacoustic image of a phantom before denoising. The overlayed image shows the arrangements of the phantom: tubes filled with ink (yellow), copper sheet (blue), and agar layers with slightly different speed of sound distributions (grey). b) Corresponding optoacoustic image after denoising. c) Difference between a and b. d-e) Optoacoustic image of a malignant breast tumor (d) before and (e) after denoising. The location of the hypoechoic tumor core, obtained from ultrasound images, is outlined in blue. f) Examples for the vessel and background ROIs that are used to compute the contrast resolution. g-h) Quantification of the contrast resolution (CR) of blood vessels in scans of breast lesions before and after denoising. The average increase is 0.083. The minimal gain is 0.003. The depicted optoacoustic images of the phantom and the breast lesion are obtained at 700 nm and 870 nm, respectively. All scale bars correspond to 5 mm.

obtained from the denoised images are broader, smoother, and are more easily related to the reference absorption spectra of hemoglobin (spectra no. 1, 2, 6, 7, 8, 10), fat (spectra no. 4, 9), and water (spectra no. 3, 5). The increased number of meaningful spectra found by NMF demonstrates superior spectral contrast of the denoised images compared to the original images. The improved unmixing accuracy upon application of the denoising method was also confirmed quantitatively by evaluating the residual norms of the NMF runs (see equation 10). The 10 NMF components obtained from the original MSOT breast scans could only represent 83.2% of the data ($r_{\text{NMF}} = 16.8\%$), whereas the 10 NMF components obtained from the denoised MSOT breast scans could represent 91.8% of the data ($r_{\text{NMF}} = 8.2\%$).

In Fig. 5d-g, we visualize the obtained spectral decompositions before and after denoising of a representative multispectral stack to visually confirm the ability of the network to enable better spectral contrast. We color-encode and blend the unmixing coefficients of three NMF spectra at a time, which correlate with the reference absorption spectra of hemoglobin (Fig. 5d,e), lipids, and water (Fig. 5f,g), covering approximately the same spectral regions for the original and the denoised data. To improve the dynamic range of the rendered images, we display the square roots of all coefficients in the visualizations. Whereas the visualizations derived from the original data are dominated by overlapping coefficients of different spectra (appearing as white in the color-encoding) and by ring noise artifacts (example marked with white arrows in Fig. 5f), the visualizations derived from denoised data show a reduction of noise artifacts and express significantly richer spectral contrast. In addition, while the tumor core (outlined in white) in Fig. 5d,f contains a lot of noise, this noise is removed by the denoising method in Fig. 5e,g, revealing hemoglobin contrast inside the tumor (white arrows in Fig. 5e). In summary, improved spectral contrast is observable in two ways upon application of the denoising method to the scans from Dataset-BC: First, blind spectral unmixing retrieves a more versatile set of spectral components and second, the denoising method enables a more meaningful decomposition of the acquired images into the found spectra.

## IV. Conclusion

Optoacoustic signals are relatively weak and thus susceptible to corruption by electrical noise during the imaging process, which impedes morphological and spectral contrast. In this work, we presented a discriminative deep-learning-based denoising method for optoacoustic sinograms, which employs a deep neural network trained on samples of experimentally acquired electrical noise and simulated ground truth optoacoustic signals. We demonstrated that the trained deep neural network could accurately remove electrical noise from

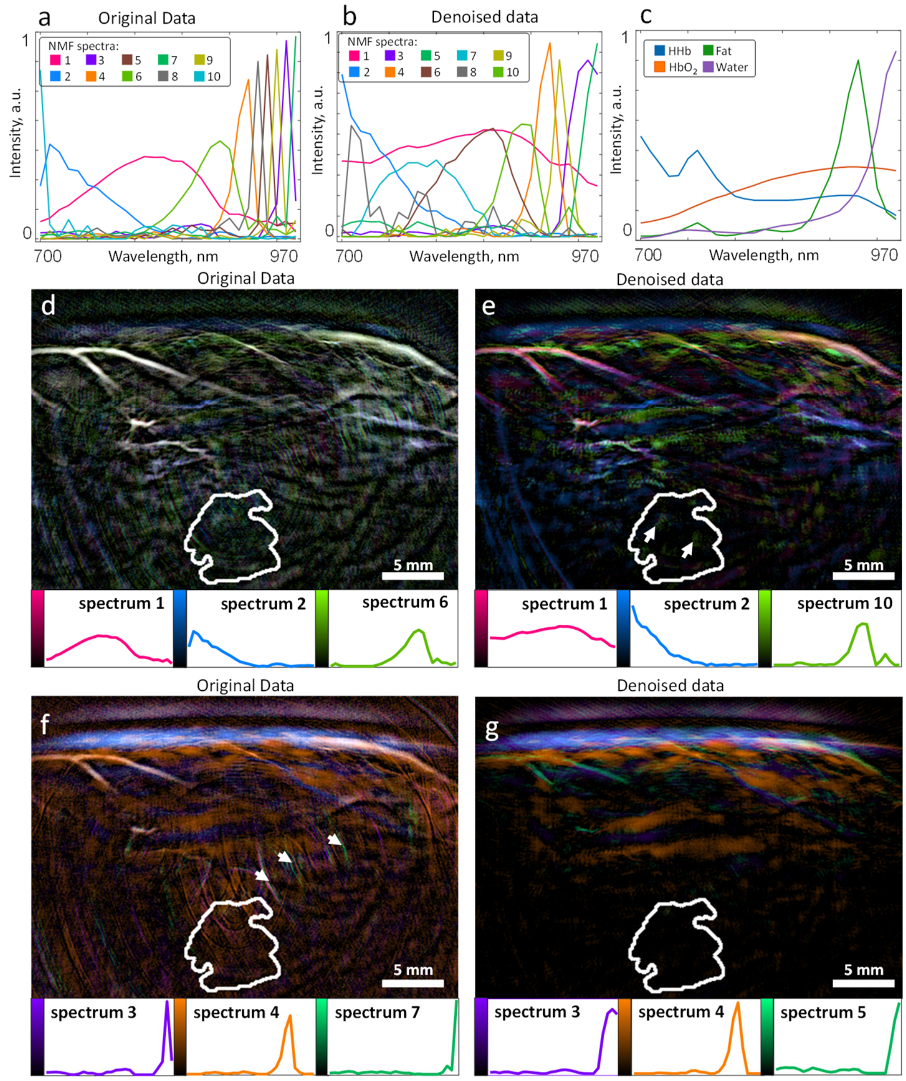


Fig. 5. Effects of denoising on the spectral content of optoacoustic images. a-b) NMF spectra that were obtained from the original (a) and the denoised (b) MSOT images of human breast lesions from Dataset-BC. c) Reference absorption spectra of the most prominent chromophores in breast tissue. d-g) Visualizations of the NMF decomposition of a representative MSOT image showing a malignant breast tumor at approximately 2 cm depth before (d, f) and after (e, g) denoising. The images color-encode the contributions of three spectra that respectively correlate with the absorption spectra of hemoglobin (d, e), fat, and water (f, g). The position of the tumor, obtained from ultrasound images, is demarcated by the white outlines.

in vivo scans of a MSOT system. The proposed signal processing technique offers a fast and accurate approach to improve the SNR of recorded optoacoustic sinograms, increase morphological image contrast, and enable rich spectral contrast at high resolution in handheld MSOT imaging.

The presented deep-learning-based denoising framework is effective because it can access the topology and the statistics of pure electrical noise and optoacoustic signal datasets. This structural information contained in large datasets has recently been made accessible by advances in computational power and methodology and is the driving force behind the increasing success of deep learning methods in medical imaging [39], [40]. We generated such a large and high-quality dataset by complementing the experimentally acquired pure noise data with simulated optoacoustic sinograms. The simulated data was obtained by applying a mathematical model of the imaging system to a general-feature image database, thereby incorporating prior knowledge about the imaging system without sacrificing general applicability of the method to any data acquired by the system. This is an example of the integration of a physical model into data-driven methods, which remains a major challenge in machine learning [27], [41], [42]. The proposed method allows a trade-off between model accuracy and generality. For example, one could potentially enhance denoising performance by selecting an optoacoustic signal dataset that more specifically reflects typical tissue responses. However, the method achieves accurate denoising and good generalization beyond the training data without any such specialization. To further evaluate the utility of the presented denoising approach, future research may also focus on strategies to improve the interpretability of the employed deep neural network model [43].

Furthermore, the trained deep neural network model provides a means of fast denoising. Clinical optoacoustic imaging systems typically provide real-time feedback to the device operator on a built-in monitor. Due to the restricted processing times, these online images are usually much lower in quality than those produced offline, which can lead to longer imaging sessions and incorrect selection of regions of interest. We demonstrated that the method can denoise a full optoacoustic sinogram of the MSOT system in approximately 9 milliseconds, which is fast enough for real-time feedback during device operation. Improving instantaneous image quality enhances the dynamic imaging capabilities of MSOT [44] while decreasing examination times.

In addition to better image quality, the denoising method also enhances the fidelity of the obtained spectral information. MSOT enables molecular contrast by extending the high-resolution optical contrast of optoacoustic imaging to the spectral dimension [45]. However, previous clinical MSOT studies extracted spectral information mostly by averaging over larger areas in MSOT images [6]–[8], thereby sacrificing the superior resolution of optoacoustic imaging. In this work, we demonstrated that denoising overcomes the necessity to average over large tissue regions and enables spectral contrast down to the system resolution, which is $\sim$200 $\mu$m in the test system. High-resolution spectral contrast was highlighted by localizing hemoglobin contrast inside a 2 cm deep breast tumor. Spectral contrast is of the utmost interest for clinical applications of MSOT, since it, for example, enables detailed studies of local blood oxygenation and tissue metabolism.

Finally, the presented denoising framework is also applicable to other (optoacoustic) imaging systems. For example, optoacoustic mesoscopy [46] and microscopy systems [47] are beset with similar electrical noise, making the approach of acquiring pure noise measurements and simulating signals with a numerical model applicable to these systems without any major changes. Other noise sources, like speckle noise in ultrasound imaging [48] or optical coherence tomography [49] and shot noise in coherent diffraction imaging [50] can be modeled as independent multiplicative noise and can thus be approached by adapting the proposed method accordingly. More generally, the presented methodology can in any context disentangle two random fields that are mixed in a known way and whose distributions can be accessed by sampling. In particular, the denoising approach can also be applied to remove signal-dependent noise if samples are obtained from the conditional probability distribution of the noise $P(S_{\text{noise}}|S_{\text{OA}} = s_{\text{OA}})$. Then, the denoising network could be trained with signal and noise samples that are generated through the following two-step process: 1: Get sample $s_{\text{OA}}$ from $S_{\text{OA}} \sim P_{\text{OA}}$. 2: Get sample $s_{\text{noise}}$ from $(S_{\text{noise}}|S_{\text{OA}} = s_{\text{OA}}) \sim P(S_{\text{noise}}|S_{\text{OA}} = s_{\text{OA}})$.

In summary, the deep learning framework that we propose in this work is an efficient and flexible method for denoising optoacoustic tomography data. By significantly improving the data quality of the considered MSOT system, we move one step closer to the full potential of handheld MSOT imaging, which is dynamic high-resolution molecular contrast deep in tissue.

## Source Code

The source code for the presented denoising framework is available at
https://github.com/juestellab/msot-sinogram-denoising.

## Acknowledgment

The authors would like to thank Hong Yang for providing and scanning the phantom, Angelos Karlas and Korbinian Paul-Yuan for designing the study to acquire the breast scans, Stefan Metz for acquiring the breast scans, Jan Kukačka for his help with the analysis of the MSOT breast images, and Robert Wilson for his attentive reading and improvements of the manuscript.

## References

[1] S. Tzoumas, A. Rosenthal, C. Lutzweiler, D. Razansky, and V. Ntziachristos, "Spatiospectral denoising framework for multispectral optoacoustic imaging based on sparse signal representation," *Med. Phys.*, vol. 41, no. 11, 2014, Art. no. 113301.

[2] J. R. Barry, E. A. Lee, and D. G. Messerschmitt, *Digital Communication*. New York, NY, USA: Springer, 2004.

[3] B. T. Cox, J. G. Laufer, P. C. Beard, and S. R. Arridge, "Quantitative spectroscopic photoacoustic imaging: A review," *J. Biomed. Opt.*, vol. 17, no. 6, 2012, Art. no. 061202.

[4] A. Taruttis and V. Ntziachristos, "Advances in real-time multispectral optoacoustic imaging and its applications," *Nature Photon.*, vol. 9, no. 4, pp. 219–227, Apr. 2015.

[5] P. Beard, "Biomedical photoacoustic imaging," *Interface Focus*, vol. 1, no. 4, pp. 602–631, Aug. 2011.

[6] F. Knieling *et al.*, "Multispectral optoacoustic tomography for assessment of Crohn's disease activity," *New England J. Med.*, vol. 376, no. 13, pp. 1292–1294, Mar. 2017.

[7] A. P. Regensburger *et al.*, "Detection of collagens by multispectral optoacoustic tomography as an imaging biomarker for Duchenne muscular dystrophy," *Nature Med.*, vol. 25, no. 12, pp. 1905–1915, Dec. 2019.

[8] W. Roll *et al.*, "Multispectral optoacoustic tomography of benign and malignant thyroid disorders: A pilot study," *J. Nucl. Med.*, vol. 60, no. 10, pp. 1461–1466, Oct. 2019.

[9] K. Wang, R. Su, A. A. Oraevsky, and M. A. Anastasio, "Investigation of iterative image reconstruction in three-dimensional optoacoustic tomography," *Phys. Med. Biol.*, vol. 57, no. 17, pp. 5399–5423, Aug. 2012.

[10] A. Buehler, A. Rosenthal, T. Jetzfellner, A. Dima, D. Razansky, and V. Ntziachristos, "Model-based optoacoustic inversions with incomplete projection data," *Med. Phys.*, vol. 38, no. 3, pp. 1694–1704, 2011.

[11] H. Yang *et al.*, "Soft ultrasound priors in optoacoustic reconstruction: Improving clinical vascular imaging," *Photoacoustics*, vol. 19, Sep. 2020, Art. no. 100172.

[12] J. Kukačka *et al.*, "Image processing improvements afford second-generation handheld optoacoustic imaging of breast cancer patients," *Photoacoustics*, vol. 26, Jun. 2022, Art. no. 100343.

[13] S. Sethuraman *et al.*, "Intravascular photoacoustic imaging using an IVUS imaging catheter," *IEEE Trans. Ultrason., Ferroelectr., Freq. Control*, vol. 54, no. 5, pp. 978–986, May 2007.

[14] J.-M. Yang *et al.*, "Photoacoustic endoscopy," *Opt. Lett.*, vol. 34, no. 10, pp. 1591–1593, 2009.

[15] L. Zeng, D. Xing, H. Gu, D. Yang, S. Yang, and L. Xiang, "High antinoise photoacoustic tomography based on a modified filtered backprojection algorithm with combination wavelet," *Med. Phys.*, vol. 34, no. 2, pp. 556–563, Jan. 2007.

[16] S. H. Holan and J. A. Viator, "Automated wavelet denoising of photoacoustic signals for circulating melanoma cell detection and burn image reconstruction," *Phys. Med. Biol.*, vol. 53, no. 12, pp. N227–N236, Jun. 2008.

[17] J. Xie, L. Xu, and E. Chen, "Image denoising and inpainting with deep neural networks," in *Advances in Neural Information Processing Systems*, vol. 25, F. Pereira, C. J. C. Burges, L. Bottou, Eds. Red Hook, NY, USA: Curran Associates, 2012, pp. 341–349.

[18] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3142–3155, Jul. 2017.

[19] X.-J. Mao, C. Shen, and Y.-B. Yang, "Image restoration using very deep convolutional encoder–decoder networks with symmetric skip connections," in *Proc. 30th Int. Conf. Neural Inf. Process. Syst.*, Barcelona, Spain, 2016, pp. 2810–2818.

[20] W. Ren, S. Liu, H. Zhang, J. Pan, X. Cao, and M.-H. Yang, "Single image dehazing via multi-scale convolutional neural networks," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 154–169.

[21] T. Tong *et al.*, "Domain transform network for photoacoustic tomography from limited-view and sparsely sampled data," *Photoacoustics*, vol. 19, Sep. 2020, Art. no. 100190.

[22] N. Davoudi, X. L. Deán-Ben, and D. Razansky, "Deep learning optoacoustic tomography with sparse data," *Nature Mach. Intell.*, vol. 1, no. 10, pp. 453–460, Oct. 2019.

[23] D. Allman, A. Reiter, and M. A. L. Bell, "Photoacoustic source detection and reflection artifact removal enabled by deep learning," *IEEE Trans. Med. Imag.*, vol. 37, no. 6, pp. 1464–1477, Jun. 2018.

[24] H. Shan, G. Wang, and Y. Yang, "Accelerated correction of reflection artifacts by deep neural networks in photo-acoustic tomography," *Appl. Sci.*, vol. 9, no. 13, p. 2615, Jun. 2019.

[25] E. M. A. Anas, H. K. Zhang, J. Kang, and E. Boctor, "Enabling fast and high quality LED photoacoustic imaging: A recurrent neural networks based approach," *Biomed. Opt. Exp.*, vol. 9, no. 8, pp. 3852–3866, 2018.

[26] A. Hariri, K. Alipour, Y. Mantri, J. P. Schulze, and J. V. Jokerst, "Deep learning improves contrast in low-fluence photoacoustic imaging," *Biomed. Opt. Exp.*, vol. 11, no. 6, pp. 3360–3373, 2020.

[27] J. Gröhl, M. Schellenberg, K. Dreher, and L. Maier-Hein, "Deep learning for biomedical photoacoustic imaging: A review," *Photoacoustics*, vol. 22, Jun. 2021, Art. no. 100241.

[28] S. Prahl. (1999). *Tabulated Molar Extinction Coefficient for Hemoglobin in Water*. [Online]. Available: http://omlc.ogi.edu/spectra/hemoglobin/summary.html

[29] D. J. Segelstein, *The Complex Refractive Index of Water*. Kansas City, MO, USA: Univ. of Missouri-Kansas City, 1981.

[30] R. L. P. van Veen, H. J. C. M. Sterenborg, A. Pifferi, A. Torricelli, E. Chikoidze, and R. Cubeddu, "Determination of visible near-IR absorption coefficients of mammalian fat using time- and spatially resolved diffuse reflectance and transmission spectroscopy," *J. Biomed. Opt.*, vol. 10, no. 5, 2005, Art. no. 054004.

[31] K. B. Chowdhury, J. Prakash, A. Karlas, D. Justel, and V. Ntziachristos, "A synthetic total impulse response characterization method for correction of hand-held optoacoustic images," *IEEE Trans. Med. Imag.*, vol. 39, no. 10, pp. 3218–3230, Oct. 2020.

[32] K. B. Chowdhury, M. Bader, C. Dehner, D. Jüstel, and V. Ntziachristos, "Individual transducer impulse response characterization method to improve image quality of array-based handheld optoacoustic tomography," *Opt. Lett.*, vol. 46, no. 1, pp. 1–4, 2021.

[33] M. Everingham *et al.*, "The PASCAL visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, pp. 303–338, 2010.

[34] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, in Lecture Notes in Computer Science, 2015, pp. 234–241.

[35] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2015, *arXiv:1412.6980*.

[36] L. A. Kunyansky, "Explicit inversion formulae for the spherical mean radon transform," *Inverse Problems*, vol. 23, no. 1, pp. 373–383, Feb. 2007.

[37] P. Kuchment and L. Kunyansky, "Mathematics of photoacoustic and thermoacoustic tomography," in *Handbook of Mathematical Methods in Imaging*, O. Scherzer, Ed. New York, NY, USA: Springer, 2011, pp. 817–865.

[38] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. 13th Int. Conf. Neural Inf. Process. Syst.*, Denver, CO, USA, 2000, pp. 535–541.

[39] H. Greenspan, B. van Ginneken, and R. M. Summers, "Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1153–1159, May 2016.

[40] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, Dec. 2017.

[41] G. Carleo *et al.*, "Machine learning and the physical sciences," *Rev. Mod. Phys.*, vol. 91, no. 4, 2019, Art. no. 045002.

[42] A. Hauptmann and B. Cox, "Deep learning in photoacoustic tomography: Current approaches and future directions," *J. Biomed. Opt.*, vol. 25, no. 11, Oct. 2020, Art. no. 112903.

[43] Q. Zhang, Y. N. Wu, and S.-C. Zhu, "Interpretable convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8827–8836.

[44] A. Karlas *et al.*, "Multispectral optoacoustic tomography of muscle perfusion and oxygenation under arterial and venous occlusion: A human pilot study," *J. Biophoton.*, vol. 13, no. 6, Jun. 2020, Art. no. e201960169.

[45] V. Ntziachristos and D. Razansky, "Molecular imaging by means of multispectral optoacoustic tomography (MSOT)," *Chem. Rev.*, vol. 110, no. 5, pp. 2783–2794, 2010.

[46] J. Aguirre *et al.*, "Precision assessment of label-free psoriasis biomarkers with ultra-broadband optoacoustic mesoscopy," *Nature Biomed. Eng.*, vol. 1, no. 5, p. 68, 2017.

[47] M. A. Pleitez *et al.*, "Label-free metabolic imaging by mid-infrared optoacoustic microscopy in living cells," *Nature Biotechnol.*, vol. 38, no. 3, pp. 293–296, Mar. 2020.

[48] A. Fenster, D. B. Downey, and H. N. Cardinal, "Three-dimensional ultrasound imaging," *Phys. Med. Biol.*, vol. 46, no. 5, pp. R67–R99, 2001.

[49] A. F. Fercher, W. Drexler, C. K. Hitzenberger, and T. Lasser, "Optical coherence tomography—Principles and applications," *Rep. Prog. Phys.*, vol. 66, no. 2, pp. 239–303, 2003.

[50] I. Robinson and R. Harder, "Coherent X-ray diffraction imaging of strain at the nanoscale," *Nature Mater.*, vol. 8, no. 4, pp. 291–298, 2009.

# Appendix B: A deep neural network for real-time optoacoustic image reconstruction with adjustable speed of sound

Research paper by Christoph Dehner, Guillaume Zahnd, Vasilis Ntziachristos, and Dominik Jüstel, as published in the journal Nature Machine Intelligence, 2023.

nature machine intelligence

Article

https://doi.org/10.1038/s42256-023-00724-3

# A deep neural network for real-time optoacoustic image reconstruction with adjustable speed of sound

Received: 22 March 2022

Accepted: 25 August 2023

Published online: 02 October 2023

Check for updates

Christoph Dehner[1,2,6], Guillaume Zahnd[1,3,6], Vasilis Ntziachristos[1,2,4,7] ✉ & Dominik Jüstel[1,2,5,7] ✉

Multispectral optoacoustic tomography is a high-resolution functional imaging modality that can non-invasively access a broad range of pathophysiological phenomena. Real-time imaging would enable translation of multispectral optoacoustic tomography into clinical imaging, visualize dynamic pathophysiological changes associated with disease progression and enable in situ diagnoses. Model-based reconstruction affords state-of-the-art optoacoustic images but cannot be used for real-time imaging. On the other hand, deep learning enables fast reconstruction of optoacoustic images, but the lack of experimental ground-truth training data leads to reduced image quality for in vivo scans. In this work we achieve accurate optoacoustic image reconstruction in 31 ms per image for arbitrary (experimental) input data by expressing model-based reconstruction with a deep neural network. The proposed deep learning framework, DeepMB, generalizes to experimental test data through training on optoacoustic signals synthesized from real-world images and ground truth optoacoustic images generated by model-based reconstruction. Based on qualitative and quantitative evaluation on a diverse dataset of in vivo images, we show that DeepMB reconstructs images approximately 1,000-times faster than the iterative model-based reference method while affording near-identical image qualities. Accurate and real-time image reconstructions with DeepMB can enable full access to the high-resolution and multispectral contrast of handheld optoacoustic tomography, thus adoption into clinical routines.

Multispectral optoacoustic tomography (MSOT) is an emerging functional imaging modality that uniquely enables non-invasive detection of optical contrast at high spatial resolution and centimetre-scale penetration depth in living tissue[1–7]. Accessing the multispectral contrast of endogenous chromophores, MSOT can quantify a broad range of pathophysiological surrogate biomarkers such as tissue fibrosis, inflammation, vascularization and oxygenation, and provide unmatched clinical information for multifarious diseases such as breast cancer[2,6], Duchenne muscular dystrophy[8] or inflammatory bowel disease[3].

[1]Institute of Biological and Medical Imaging, Helmholtz Zentrum München, Neuherberg, Germany. [2]Chair of Biological Imaging at the Central Institute for Translational Cancer Research (TranslaTUM), School of Medicine, Technical University of Munich, Munich, Germany. [3]iThera Medical GmbH, Munich, Germany. [4]Munich Institute of Robotics and Machine Intelligence (MIRMI), Technical University of Munich, Munich, Germany. [5]Institute of Computational Biology, Helmholtz Zentrum München, Neuherberg, Germany. [6]These authors contributed equally: Christoph Dehner, Guillaume Zahnd. [7]These authors jointly supervised this work: Vasilis Ntziachristos, Dominik Jüstel. ✉e-mail: bioimaging.translatum@tum.de; dominik.juestel@helmholtz-munich.de

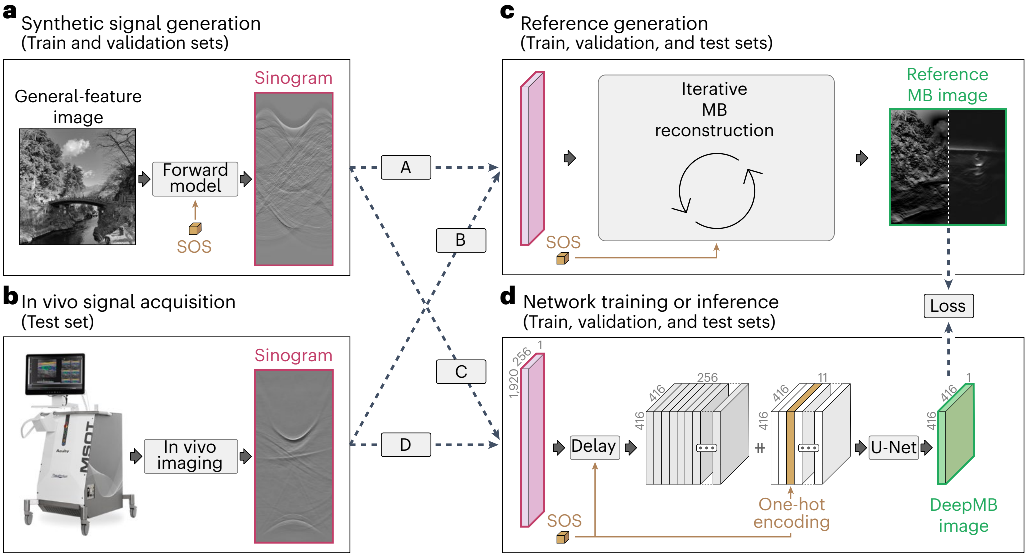


**Fig. 1 | DeepMB pipeline. a**, Real-world images—obtained from a publicly available dataset—are used to generate synthetic sinograms by applying an accurate physical forward model of the scanner. SOS, speed of sound. **b**, In vivo sinograms are acquired from six participants at diverse anatomical locations. **c**, Optoacoustic images are reconstructed via iterative model-based (MB) reconstruction to generate reference images for the synthetic (A) and in vivo (B) datasets. **d**, A deep neural network is trained using the synthetic data as training and validation sets (C), and the in vivo data as test set (D). In the network, input sinograms are first mapped into the image domain using a delay operation. Then, the SOS is one-hot encoded and concatenated as additional channels (represented by the + symbol). Finally, the output image is regressed from the channel stack using a U-Net convolutional neural network. The loss is calculated between the network output and the corresponding reference image (see 'Network training' section in the Methods for further details about the network training).

Real-time application is imperative to fully translate and integrate MSOT into clinical imaging[9–11]. Handheld MSOT imaging requires—similar to ultrasound imaging—live image feedback at sufficiently high frame rates (at least 24 fps for full-video rendering) to avoid hindering visio-tactile coordination, identify and localize relevant tissue structures using anatomical landmarks in their surroundings, and find the optimal transducer pose for the target region. Furthermore, real-time optoacoustic imaging is necessary to visualize dynamic pathophysiological changes associated with disease progression and enable in situ guidance and diagnosis during intra-operative and endoscopy imaging[12,13]. In practice, real-time reconstruction of optoacoustic images (that is, recovery of the initial pressure distribution in the imaged tissue) is generally conducted via the backprojection algorithm[14]; however, the backprojection formula is based on over-simplified modelling assumptions of the imaging process and cannot compensate for the ill-posedness of the underlying inverse problem arising from limited-angle acquisition, measurement noise and finite transducer bandwidth. Consequently, backprojection images systematically suffer from low spatial resolution and contrast, as well as negative pixel values that invalidate a physical interpretation of the image as an initial pressure distribution. By contrast, iterative model-based reconstruction[15,16] can provide accurate, state-of-the-art quality optoacoustic images by incorporating a physical model of the imaging device into the reconstruction process, constraining the reconstructed image to be non-negative, and introducing regularization to mitigate the ill-posedness of the inversion problem. Nevertheless, model-based reconstruction is computationally demanding due to the iterative and thus sequential nature of the algorithm, which is prohibitive for real-time imaging. Real-time model-based reconstruction has been demonstrated for a pre-clinical MSOT system by computing the reconstruction with a graphics processing unit (GPU)[17], but a similar acceleration is infeasible for state-of-the-art model-based reconstruction of data from modern clinical systems as these reconstructions are much more computationally demanding (larger images, more complex regularization functionals, inclusion of the total impulse response of the system in the model, necessity of a higher number of iterations until convergence[15,16]). The full imaging potential of MSOT is therefore only available offline after considerable computational time and currently remains inaccessible for clinical applications that require live image feedback.

Deep neural networks have recently been successfully applied to various inverse problems in imaging, utilizing their ability to capture suitable inverse transforms in a data-driven way and efficiently apply these transforms to new data[18–24]. Real-time image reconstruction with deep learning has been achieved using deep loop unfolding and direct inference. Deep loop unfolding involves interpreting the iterations of a variational reconstruction algorithm as the layers of a convolutional neural network, and training the resulting network end-to-end in a supervised fashion[25–29]. This methodology has been shown to facilitate accurate and efficient image reconstruction for various medical imaging modalities such as magnetic resonance imaging, computed tomography or intensity diffraction tomography. However, deep loop unfolding is unsuited for real-time optoacoustic image reconstruction as it requires repeated evaluations of the involved optoacoustic forward model (at least one forward and one adjoint model evaluation per data consistency block; see, for example, equation 11 in ref. 25), which is too computationally expensive to enable real-time processing (for example, with the imaging set-up from this paper, a single evaluation of the forward or adjoint model already takes more than 50 ms on a NVIDIA GeForce RTX3090 GPU). Conversely, deep-learning-based image reconstruction via direct inference can support real-time optoacoustic imaging because the approach does not require that the optoacoustic forward model is evaluated during image reconstruction. Over the past few years, several direct inference methods have been introduced to either directly infer high-quality images from recorded signals[30–36] or accelerate the minimization operation from iterative model-based reconstruction[37].

A key challenge in applying deep learning for optoacoustic image reconstruction is the generation of appropriate training data, that is, input sinograms and corresponding optoacoustic initial pressure reference images. In general, network training must rely on synthetized data because ground truth information on the initial pressure distribution in biological tissue is not available experimentally. Data synthesis involves hand-crafting reference distributions of the initial pressure and simulating the corresponding input sinograms using a physical forward model of the imaging process; however, such synthesized sinograms and reference images only partially represent the true properties of experimental data, and hence their use as input-target pairs for network training can lead to reductions in reconstruction accuracy for in vivo data.

In this work we show that learning a well-posed reconstruction operator facilitates accurate generalization from synthesized training data to experimental test data. We achieve real-time optoacoustic image reconstruction for arbitrary (experimental) input data by expressing model-based reconstruction using a deep neural network. The proposed deep learning framework, DeepMB, learns an accurate and universally applicable model-based optoacoustic reconstruction operator through training on optoacoustic signals synthesized from real-world images while using the optoacoustic images generated by model-based reconstruction of the corresponding signals as ground truth. DeepMB affords image quality nearly indistinguishable from state-of-the-art iterative model-based reconstructions at speeds enabling live imaging (32 fps, or 31 ms per image, versus 30–60 s per image for iterative model-based reconstruction). Furthermore, DeepMB is directly compatible with state-of-the-art clinical MSOT scanners because it supports high throughput data acquisition (sampling rate = 40 MHz; number of transducers = 256) and large image sizes (416 × 416 pixels). DeepMB also supports dynamic adjustments of the speed of sound (SOS) parameter during imaging, which enables the reconstruction of in-focus images for arbitrary tissue types. We demonstrate the performance of DeepMB both quantitatively and qualitatively on a diverse dataset of in vivo images (4,814 images, six participants, 25–29 scanned locations per participant).

## Results

To validate the capability of DeepMB to reconstruct images in real-time and with adjustable SOS, the framework was applied to a modern handheld optoacoustic scanner (MSOT Acuity Echo, iThera Medical GmbH) with SOS values ranging from 1,475 m s$^{-1}$ to 1,525 m s$^{-1}$ in 5 m s$^{-1}$ steps.

### DeepMB pipeline

Figure 1 illustrates the overall training and evaluation pipeline. DeepMB was trained similarly to the AUTOMAP framework[19], using input sinograms synthesized from general-feature images to facilitate the learning of an unbiased and universally applicable reconstruction operator. These sinograms were generated by employing a diverse collection of publicly available real-world images[38] as initial pressure distributions and simulating thereof the signals recorded by the scanner using an accurate physical forward model of the imaging process[15] (Fig. 1a and 'Synthesis of sinograms for training and validation' section in the Methods). The SOS values for the forward simulations were drawn uniformly at random from the considered range for each image. Ground-truth images for the synthesized sinograms were computed via model-based reconstruction (Fig. 1c). Figure 1d shows the deep neural network architecture of DeepMB, which inputs a sinogram (either synthetized or in vivo) and a SOS value, and outputs the final reconstructed image. The underlying design is based on the U-Net architecture[39] augmented with two extensions that promote the network to learn and express the effects of the different input SOS values onto the reconstructed images: (1) all signals were mapped from the input sinogram to the image domain with a linear delay operator based on the given input SOS value (no trainable weights), and (2) the input SOS value (one-hot encoded and concatenated as additional channels) was passed to the trainable convolutional layers of the network. A detailed description of the network training is given in the 'Network training' section in the Methods. After training, the applicability of DeepMB to clinical data was tested with a diverse dataset of in vivo sinograms acquired by scanning six participants at up to eight anatomical locations each (Fig. 1b). The corresponding ground-truth images of the acquired in vivo test sinograms were obtained analogously to the training data via model-based reconstruction. The inference time of DeepMB was 31 ms per sample on a modern GPU (NVIDIA GeForce RTX 3090).

### Qualitative evaluation

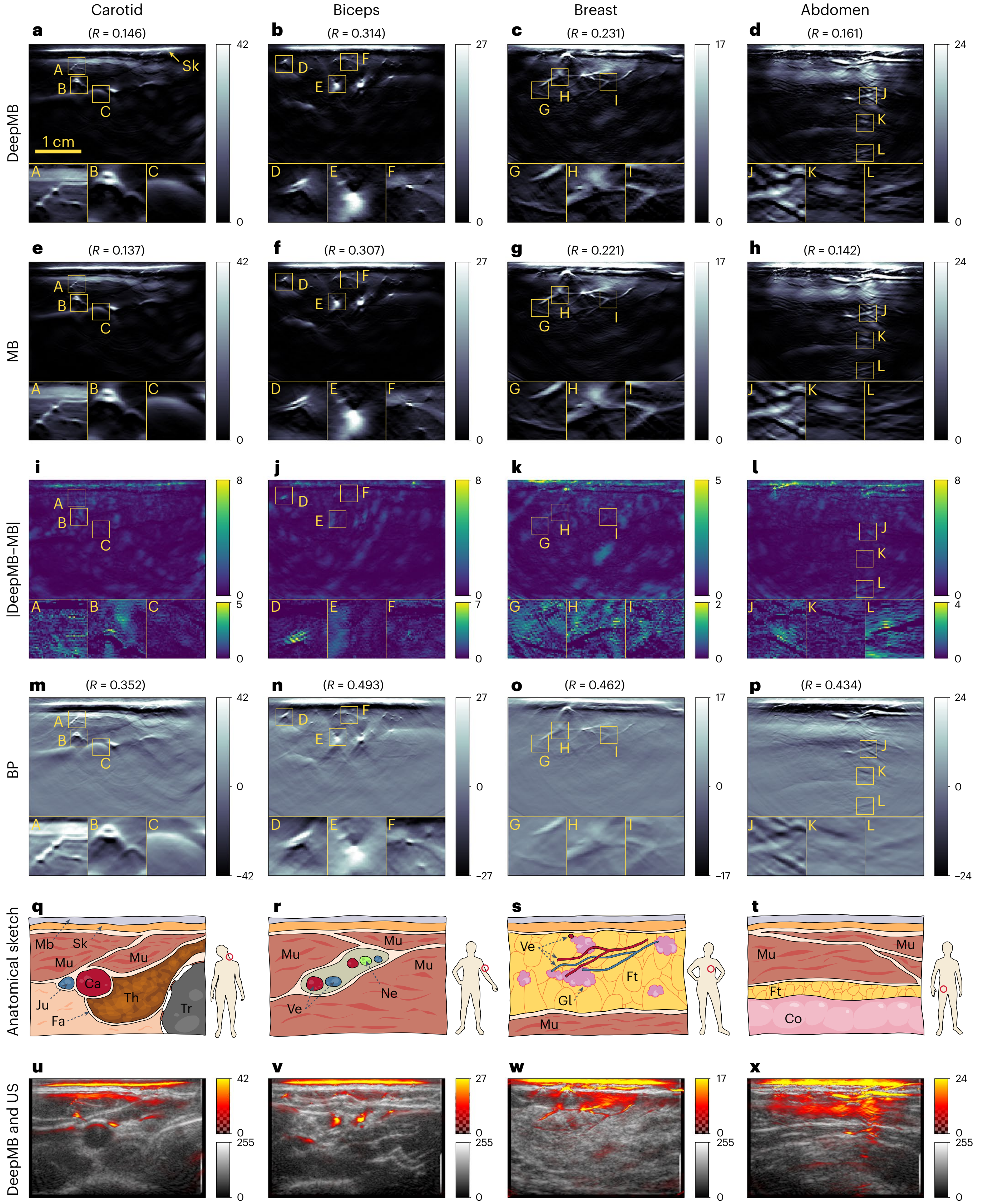
DeepMB successfully reconstructed high-quality optoacoustic images. To qualitatively evaluate DeepMB, all DeepMB images from the in vivo dataset (Fig. 1b) were thoroughly compared with their corresponding model-based reference images (Fig. 1c). Figure 2 shows four reconstructed images that correspond to scans of the carotid artery, biceps, breast and abdomen. DeepMB reconstructions (Fig. 2a–d) are systematically nearly indistinguishable from model-based references (Fig. 2e–h), with no noticeable failures, outliers or artefacts for any of the participants, anatomies, probe orientations, SOS values or laser wavelengths. The similarity between DeepMB and model-based images is also confirmed by their negligible pixel-wise absolute differences (Fig. 2i–l). The magnified region D in Fig. 2j depicts one of the largest observed discrepancies between DeepMB and model-based reconstructions, which manifests as minor blurring, showing that the DeepMB image is only marginally affected by these errors. In comparison, backprojection images (Fig. 2m–p) exhibit notable differences from reference model-based images and suffer from reduced spatial resolution and physically nonsensical negative initial pressure values. Finally, to facilitate relating the reconstructed optoacoustic images to the scanned anatomies, Fig. 2q–t depicts sketches of the rough anatomical context for all scans and Fig. 2u–x depicts the interleaved-acquired ultrasound images overlayed with the temporally corresponding DeepMB reconstructions. Extended Data Figs. 1 and 2 complement the qualitative comparison from Fig. 2: Extended Data Fig. 1 shows that the image quality of DeepMB is also superior to the backprojetion algorithm with negative values set to zero after the reconstruction, as well as to the delay-multiply-and-sum with coherence factor algorithm[40,41]. Extended Data Fig. 2 shows that DeepMB images are nearly indistinguishable from model-based references in the case of both very high and very low data residual norms.

Extended Data Videos 1 and 2 further illustrate the real-time optoacoustic imaging capabilities of DeepMB. Extended Data Video 1 shows a carotid artery continuously imaged in the transversal view at 800 nm, which demonstrates that DeepMB can be used to visualize motion at 25 Hz with state-of-the-art image quality. Extended Data Video 2 shows the optoacoustic image of a biceps in the transversal view at 800 nm while the SOS is gradually adjusted via a series of DeepMB reconstructions, which illustrates the importance of impromptu SOS tuning for optimal image quality.

**Fig. 2 | Examples from the in vivo test dataset for different anatomical locations. a**,**e**,**i**,**m**,**q**,**u**, Carotid artery. **b**,**f**,**j**,**n**,**r**,**v**, Biceps. **c**,**g**,**k**,**o**,**s**,**w**, Breast. **d**,**h**,**l**,**p**,**t**,**x**, Abdomen. The first four rows show DeepMB reconstructions, MB reconstructions, the pixel-wise absolute difference between DeepMB and MB reconstructions, and backprojection (BP) reconstructions. Data residual norm (*R*) values are shown above all reconstructed images. The last two rows display sketches of the rough anatomical context of the scans and the interleaved-acquired ultrasound (US) images overlayed with DeepMB reconstructions, respectively. All optoacoustic images and difference maps show the reconstructed initial pressure in arbitrary units and were slightly cropped to a field of view of 4.16 × 2.80 cm$^2$ to disregard the area occupied by the probe couplant above the skin line. Each enlarged region is 0.41 × 0.41 cm$^2$ and displays various anatomical details. All displayed scans were acquired at 800 nm. Mb, probe membrane; Sk, skin; Mu, muscle; Fa, fascia; Ca, common carotid artery; Ju, jugular vein; Th, thyroid; Tr, trachea; Ve, blood vessel; Ne, nerve; Ft, fat; Gl, glandular tissues; Co, colon.

### Quantitative evaluation

The ability of DeepMB to reconstruct images with equivalent fidelity to those afforded by model-based reconstruction was then confirmed by quantitative comparison. To quantify the image fidelity of DeepMB reconstructions, data residual norms were calculated for all in vivo test images (see the 'Data residual norm' section in the Methods for the precise definition). The data residual norm measures the fidelity of a reconstructed image by computing the mismatch between the image



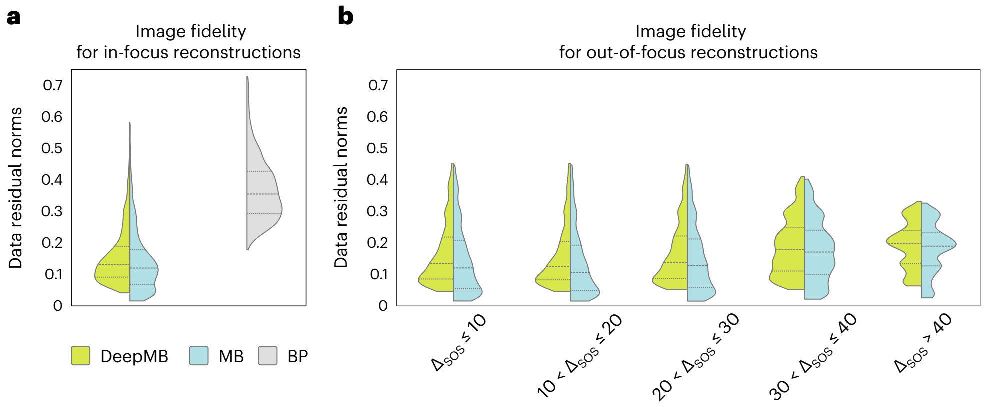


**Fig. 3 | Data residual norms of optoacoustic images from DeepMB, MB and BP reconstructions. a**, Data residual norms of in-focus images reconstructed with optimal SOS values, on all 4,814 samples from the in vivo test set. **b**, Data residual norms of out-of-focus images reconstructed with suboptimal SOS values, on a subset of 638 samples. The five plots depict the effect of SOS mismatch via a gradual increase of the offset ΔSOS in steps of 10 m s$^{-1}$. The inner bars indicate the 25th, 50th and 75th percentiles.

and the corresponding recorded acoustic signals with regard to the accurate physical forward model of the used scanner. The metric is demonstrably minimal for model-based reconstruction[42]. Data residual norms were also calculated for all other reconstruction methods for comparison purposes.

First, data residual norms were calculated with in-focus images (that is, reconstructed with optimal SOS values) to evaluate the fidelity of DeepMB images with the best possible quality (Fig. 3a). Data residual norms of DeepMB images (green; mean ± s.d. = 0.156 ± 0.088) are almost as low as the data residual norms of model-based images (blue; mean ± s.d. = 0.139 ± 0.095). The close agreement between data residual norms of DeepMB and model-based images confirms that both reconstruction approaches afford equivalent image qualities. By contrast, the data residual norms of backprojection images are markedly higher (grey, mean ± s.d. = 0.369 ± 0.098), which reaffirms the shortcomings of backprojection to accurately model the imaging process and explains the lower image quality observed in Fig. 2m-p. Table 1 summarizes the data residual norms of all reconstruction approaches evaluated in this paper. Extended Data Table 1 complements the quantitative comparison from Table 1 and confirms that the data residual norms of DeepMB images are almost as low as data residual norms of model-based images even when aggregated separately based on anatomical regions, participants, Fitzpatrick scale, body type, wavelength and SOS values.

Second, data residual norms were calculated for out-of-focus images (that is, reconstructed with sub-optimal SOS values) to evaluate the fidelity of DeepMB images during imaging applications with a priori-unknown SOS (Fig. 3b and Table 1). Data residual norms of DeepMB images remain close to those of model-based images for all considered levels of mismatch between the optimal and the employed SOS, thus confirming that DeepMB and model-based images are similarly trustworthy regardless of the selected SOS. Note that the two right-most distributions of data residual norms in Fig. 3b get narrower and include less extreme data residual norm values because they contain fewer data points.

In addition to the quantitative evaluation with data residual norms, the deviation of DeepMB and backprojection images from reference model-based reconstructions were also quantified by computing the mean absolute error, relative mean absolute error, mean squared error, relative mean squared error and structural similarity index. The obtained metrics for the in vivo test scans are reported in Table 1 and confirm that DeepMB images are very similar to model-based images, whereas backprojection images notably differ from the model-based references.

## Multispectral evaluation

The previously described experiments validate—using in vivo scans in the 700–980 nm range—that the single-wavelength image quality of DeepMB is nearly identical to model-based reconstruction and clearly superior to backprojection reconstruction. Further experiments were then conducted to show that the multispectral image contrast of DeepMB is comparable with model-based reconstruction, and superior to backprojection reconstruction.

To evaluate the multispectral image quality of DeepMB, model-based and backprojection reconstruction, all of the in-vivo scans from the test dataset were grouped into multispectral stacks of 29 images (one scan across the 700–980 nm range in steps of 10 nm, respectively) and linearly unmixed into oxyhaemoglobin, deoxyhaemoglobin, fat and water components[43]. Figure 4 visualizes the unmixed components from DeepMB, model-based and backproejction images for a representative breast scan, showing: the unmixed components for fat and water (Fig. 4a–c); the unmixed components for oxyhaemoglobin and deoxyhaemoglobin (Fig. 4d–f); the reference absorption spectra of the four chromophores used during unmixing (Fig. 4g); and a schematic sketch of the anatomical context for the depicted scan (Fig. 4h). The unmixed DeepMB images (Fig. 4a,d) are systematically nearly indistinguishable from the model-based references (Fig. 4b,e). Conversely, the unmixed backprojection images (Fig. 4c,f) exhibit considerably lower multispectral contrast (see, for example, magnifications A–C in Fig. 4c) and miss important image structures (see, for example, the fine vascularity in magnification B of Fig. 4f). Extended Data Figs. 3–5 visualize the unmixing results of three further in vivo scans and also display unmixed images from the delay-multiply-and-sum with coherence factor algorithm. Finally, the ability of DeepMB to obtain clearly superior multispectral images as backprojection and delay-multiply-and-sum with coherence factor was confirmed quantitatively by computing the structural similarity index, mean squared error, and mean absolute error for all unmixed images against the reference unmixed model-based images (see Table 2).

## Comparison with alternative training strategies for DeepMB

The evaluation experiments described so far thoroughly validate the ability of DeepMB to reconstruct high-quality images with adjustable SOS values from the range 1,475–1,525 m s$^{-1}$. Furthermore, alternative training strategies were assessed to better understand the effects of different specific aspects of the DeepMB methodology on the obtained image quality. Quantitative results from all conducted experiments are also reported in Tables 1 and 2.

**Table 1 | Quantitative evaluation of the image quality for all reconstruction methods assessed in this paper in comparison with the reference model-based reconstruction**

| | | Our method | Reference method | Traditional methods | | Alternative DeepMB training strategies | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | DeepMB | MB | BP | DMAS-CF | DeepMB$_{\text{no-sos}}$ | DeepMB$_{\text{scalar-sos}}$* | DeepMB$_{\text{initial-images}}$** | DeepMB$_{\text{in-vivo}}$*** |
| In-focus images | *R* (↓) | 0.156<br>(0.092, 0.189) | 0.139<br>(0.068, 0.180) | 0.369<br>(0.294, 0.428) | 0.982<br>(0.972, 0.996) | 0.164<br>(0.106, 0.193) | 0.169<br>(0.113, 0.197) | 0.267<br>(0.196, 0.324) | 0.155<br>(0.088, 0.193) |
| | MAE (↓) | 0.74<br>(0.43, 0.75) | n/a | 3.98<br>(2.60, 4.50) | 4.58<br>(2.80, 5.14) | 0.72<br>(0.41, 0.74) | 0.80<br>(0.47, 0.81) | 18.23<br>(13.49, 21.07) | 0.59<br>(0.30, 0.57) |
| | MAE$_{\text{rel}}$ (%) (↓) | 15.21<br>(12.90, 17.21) | n/a | 86.42<br>(82.43, 90.50) | 96.60<br>(95.58, 98.31) | 14.81<br>(12.35, 16.88) | 16.54<br>(14.11, 18.53) | 429.55<br>(357.69, 494.13) | 11.42<br>(9.57, 12.78) |
| | MSE (↓) | 9.45<br>(0.56, 2.41) | n/a | 84.98<br>(24.97, 85.20) | 254.85<br>(60.57, 236.06) | 8.51<br>(0.56, 3.15) | 10.16<br>(0.58, 3.41) | 703.59<br>(325.74, 837.51) | 5.35<br>(0.43, 1.70) |
| | MSE$_{\text{rel}}$ (%) (↓) | 1.34<br>(0.65, 1.49) | n/a | 37.01<br>(29.18, 43.85) | 94.85<br>(93.48, 97.46) | 1.47<br>(0.65, 1.84) | 1.71<br>(0.82, 2.00) | 455.05<br>(265.24, 574.32) | 0.93<br>(0.50, 1.03) |
| | SSIM (↑) | 0.98<br>(0.98, 0.99) | n/a | 0.73<br>(0.68, 0.79) | 0.65<br>(0.61, 0.69) | 0.98<br>(0.98, 0.99) | 0.98<br>(0.97, 0.99) | 0.37<br>(0.31, 0.42) | 0.99<br>(0.99, 0.99) |
| Out-of-focus images | *R* (↓) | 0.166<br>(0.087, 0.222) | 0.149<br>(0.059, 0.212) | 0.365<br>(0.281, 0.439) | 0.982<br>(0.971, 0.997) | 0.176<br>(0.105, 0.228) | 0.181<br>(0.111, 0.230) | 0.275<br>(0.196, 0.342) | 0.164<br>(0.081, 0.226) |
| | MAE (↓) | 0.78<br>(0.42, 0.76) | n/a | 4.10<br>(2.58, 4.49) | 4.72<br>(2.74, 5.14) | 0.77<br>(0.40, 0.76) | 0.83<br>(0.46, 0.82) | 18.43<br>(13.56, 20.87) | 0.61<br>(0.30, 0.59) |
| | MAE$_{\text{rel}}$ (%) (↓) | 15.12<br>(12.37, 17.21) | n/a | 86.34<br>(81.90, 90.79) | 96.53<br>(95.45, 98.01) | 14.85<br>(12.29, 17.29) | 16.49<br>(14.08, 18.55) | 425.41<br>(356.87, 486.35) | 11.42<br>(9.67, 12.81) |
| | MSE (↓) | 13.85<br>(0.51, 2.56) | n/a | 92.27<br>(24.21 85.79) | 295.52<br>(52.93 240.66) | 11.87<br>(0.52, 3.64) | 13.79<br>(0.63, 3.64) | 711.53<br>(320.39, 783.03) | 6.99<br>(0.39, 1.82) |
| | MSE$_{\text{rel}}$ (%) (↓) | 1.41<br>(0.65, 1.54) | n/a | 36.78<br>(28.22, 45.58) | 94.54<br>(93.25, 97.13) | 1.55<br>(0.64, 2.04) | 1.80<br>(0.84, 2.10) | 445.76<br>(259.31, 570.82) | 0.89<br>(0.52, 1.01) |
| | SSIM (↑) | 0.98<br>(0.97, 0.98) | n/a | 0.71<br>(0.65, 0.79) | 0.62<br>(0.58, 0.67) | 0.98<br>(0.98, 0.99) | 0.98<br>(0.97, 0.98) | 0.36<br>(0.31, 0.41) | 0.99<br>(0.99, 0.99) |

The table shows the mean values and in brackets the 25th and 75th percentiles for in focus (4,814 in vivo sinograms from the test dataset reconstructed with each one's optimal SOS values) and out-of-focus (638 in vivo sinograms from the test dataset reconstructed each with all 11 available SOS values) images. The arrow symbols (↑) and (↓) indicate for each metric whether a higher or lower value is better. R, data residual norm; MAE, mean absolute error; MAE$_{\text{rel}}$, relative man absolute error; MSE, mean squared error; MSE$_{\text{rel}}$, relative mean squared error; SSIM, structural similarity index; DMAS-CF, delay-multiply-and-sum with coherence factor; DeepMB$_{\text{no-sos}}$, training conducted without providing the SOS as additional input to the U-Net; DeepMB$_{\text{scalar-sos}}$, training conducted with encoding the SOS value into one additional input channel for the U-Net; DeepMB$_{\text{initial-images}}$, training conducted on the true synthetic initial pressure images instead of the corresponding MB reconstructions; DeepMB$_{\text{in-vivo}}$, training conducted on in vivo data instead of synthetic data. *All DeepMB$_{\text{scalar-sos}}$ images systematically have their overall brightness associated with the input SOS. **All DeepMB$_{\text{initial-images}}$ images suffer from strong reconstruction artefacts that manifest as intensity saturation (see Extended Data Fig. 6). ***Some DeepMB$_{\text{in-vivo images}}$ suffer from visible reconstruction artefacts that manifest as coffee-stain-like structures (see Extended Data Fig. 7).

**Advantages of one-hot-encoded SOS values.** Passing the one-hot-encoded input SOS value to the trainable layers of the network (as shown in Fig. 1d) slightly improves the image fidelity (that is, the data residual norms) of DeepMB reconstructions. To evaluate the benefits of this strategy, two other models with alternative SOS encoding schemes were trained and assessed: the first without providing the SOS to the U-Net (referred to as DeepMB$_{\text{no-sos}}$), and the second with the SOS encoded as a scalar value into one additional input channel for the U-Net (referred to as DeepMB$_{\text{scalar-sos}}$). In both models the SOS was used to apply the delay operator before the trainable U-Net layers, analogously to the standard DeepMB model (see Fig. 1d). Not providing the SOS as input to the U-Net was found to be a marginally inferior alternative to the standard one-hot-based SOS encoding with respect to image fidelity: DeepMB$_{\text{no-sos}}$ inferred high-quality and artefact-free images with visually the same quality as the standard DeepMB model, but with, on average, slightly higher data residual norms (0.164 versus 0.156). Further quantitative comparison of DeepMB and DeepMB$_{\text{no-sos}}$ reconstructions with image-based metrics did not identify a clearly superior approach, which corroborates that their overall visual appearance is very similar. Providing the SOS as scalar value to the U-Net was found to be a disadvantageous encoding scheme that impedes the ability of the neural network to learn an accurate reconstruction operator, as the overall brightness of images reconstructed with DeepMB$_{\text{scalar-sos}}$ was found to be associated with the input SOS values. More specifically, inferring DeepMB$_{\text{scalar-sos}}$ onto the same sinogram with different input SOS values obtained images of lower average intensities for higher input SOS values. These intensity differences were visually imperceptible with default colour maps but resulted in notably higher average data residual norms for the obtained images in comparison to DeepMB$_{\text{no-sos}}$ or the standard DeepMB model (0.169 versus 0.164 or 0.156).

**Advantages of model-based reference images.** Using model-based reference images as ground-truth references during training is essential to learn a generalizable model-based reconstruction operator. To compare the training strategy of DeepMB to the training methodology reported in previous deep-learning-based reconstruction methods for which the learning reference was true initial pressure images[30–35], another alternative model, referred to as DeepMB$_{\text{initial-images}}$, was trained using as ground-truth references the true synthetic initial pressure images (left side of Fig. 1a) instead of model-based reconstructions (right side of Fig. 1c). The reconstruction operator learnt by DeepMB$_{\text{initial-images}}$ was inferior in comparison to the standard DeepMB model: in vivo images reconstructed with DeepMB$_{\text{initial-images}}$ suffer from low resolution and contrast (see Extended Data Fig. 6) and have notably worse data residual norms (mean ± s.d. = 0.267 ± 0.094) than the standard DeepMB model.

**Advantages of synthesized training data.** Synthesized training data enables DeepMB to learn an accurate and general reconstruction operator. To contextualize the image quality of DeepMB with synthesized

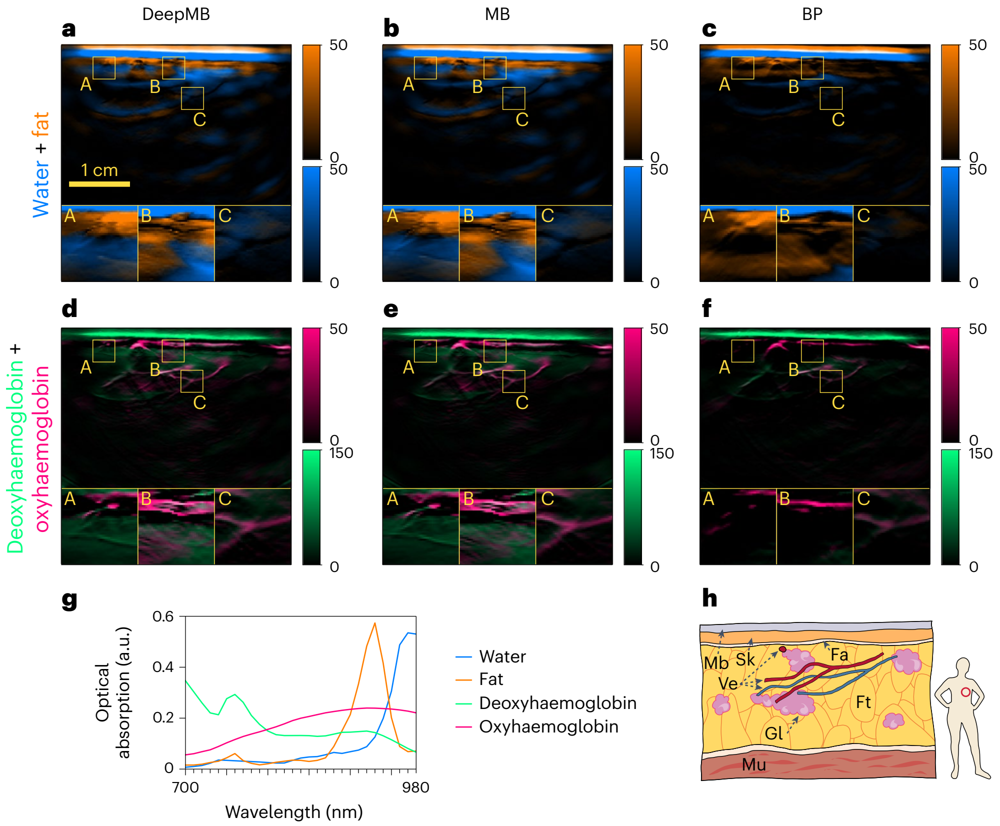


**Fig. 4 | Unmixing of a representative multispectral breast scan for DeepMB, MB and BP. a–f**, The unmixed components for fat and water (**a**–**c**), and for oxyhaemoglobin and deoxyhaemoglobin (**d**–**f**) for DeepMB (**a**,**d**), MB (**b**,**e**) and BP (**c**,**f**). **g**,**h**, The reference absorption spectra of the four chromophores used during unmixing (**g**) and a schematic sketch of the anatomical context for the depicted scan (**h**) are depicted underneath. All optoacoustic images show the unmixed components in arbitrary units and were slightly cropped to a field of view of 4.16 × 2.80 cm$^2$ to disregard the area occupied by the probe couplant above the skin line. Mb, probe membrane; Sk, skin; Fa, fascia; Mu, muscle; Ve, blood vessel; Ft, fat; Gl, glandular tissues.

training data, alternative DeepMB models were trained on in vivo data instead of real-world images. These models, referred to as DeepMB$_{\text{in-vivo}}$, inferred images with—on average—slightly better data residual norms than the standard DeepMB model (0.155 versus 0.156); however, approximately 20% of all DeepMB$_{\text{in-vivo}}$ images contained visible artefacts, either at the left or right image borders, or in regions showing strong absorption at the skin surface. Extended Data Fig. 7 shows representative examples of such artefacts. No artefacts were observed with the standard DeepMB model (trained using synthesized data), even when reducing the size of the synthetic training set from 8,000 to 3,500 to match the reduced amount of available in vivo training data.

## Discussion

We introduce a deep-learning-based reconstruction framework called DeepMB to learn the iterative model-based reconstruction operator and infer images with nearly identical quality as model-based reconstruction in 31 ms per image. We trained DeepMB on synthesized sinograms from real-world images instead of in vivo images, as these synthesized sinograms afford a large training dataset with a versatile set of image features, allowing DeepMB to accurately reconstruct images with diverse features. Such general-feature training datasets reduce the risk of encountering out-of-distribution samples (test data with features that are not contained in the training dataset) when applying the trained model to in vivo scans. By contrast, training a model on in vivo scans systematically introduces the risk of overfitting to specific characteristics of the training samples and could potentially lead to decreased image quality for never-seen-before scans that may involve different anatomical views or disease states. We indeed observed that the alternative models trained on in vivo data (DeepMB$_{\text{in-vivo}}$) failed to adequately generalize to some in vivo test scans and introduced artefacts within the reconstructed images (see Extended Data Fig. 7). Furthermore, using synthesized data instead of in vivo data alleviates the training of new DeepMB models as it obviates the need for recruiting and scanning a cohort of participants. Instead, training data can be automatically generated and used to straightforwardly obtain specifically trained DeepMB models for new scanners or different reconstruction parameters. On the other hand, our quantitative evaluation with data residual norms and image-based metrics showed that the use of more domain-specific training data (in our case in vivo scans) facilitated in aggregate slightly better images than the standard DeepMB model (for example, average data residual norms of 0.155 for DeepMB$_{\text{in-vivo}}$ versus 0.156 for DeepMB). Domain-specific training data can improve the reconstruction performance because it facilitates learning of a domain-specific data transform that exploits inherent characteristics and local spatial correlation of the considered data manifold[19]. Overall, the trade-off between domain-specific training data to improve accuracy and general training data to reduce the risk of out-of-distribution samples remains a fundamental challenge for real-world application of deep learning[44,45]. Subsequent research may therefore focus on strategies for balancing generality, accuracy and practicality during model training, for example, by employing hybrid training sets combining synthesized data from real-world images with in vivo optoacoustic images and synthesized data from other biomedical scenes or by applying domain-adaptation techniques[37,46,47].

Accurate generalization from synthesized training to in vivo test data is possible with DeepMB because the underlying inverse problem to solve (that is, regularized model-based reconstruction[42]) is well-posed; for each input sinogram there is a unique and stable solution (that is,

**Table 2 | Quantitative comparison of the unmixing components from DeepMB, BP and all alternative DeepMB models with the unmixing components from reference model-based reconstruction**

| | Our method | Traditional methods | | Alternative DeepMB training strategies | | | |
|---|---|---|---|---|---|---|---|
| | **DeepMB** | **BP** | **DMAS-CF** | **DeepMB$_{\text{no-sos}}$** | **DeepMB$_{\text{scalar-sos}}$** | **DeepMB$_{\text{initial-images}}$** | **DeepMB$_{\text{in-vivo}}$** |
| MAE (↓) | 1.26 (0.80, 1.50) | 5.34 (3.91, 6.04) | 7.78 (5.82, 8.78) | 1.26 (0.77, 1.54) | 1.39 (0.90, 1.61) | 29.22 (24.58, 32.07) | 1.06 (0.62, 1.20) |
| MAE$_{\text{rel}}$ (%) (↓) | 15.34 (13.20, 16.83) | 67.46 (65.47, 69.80) | 98.90 (98.39, 99.71) | 15.26 (13.10, 16.70) | 17.01 (15.07, 18.47) | 390.97 (342.72, 431.73) | 12.51 (10.50, 13.57) |
| MSE (↓) | 52.18 (3.52, 38.8) | 337.88 (96.38, 423.92) | 1,527.62 (476.96, 1,666.41) | 51.38 (4.05, 46.39) | 62.23 (4.67, 49.09) | 3,344.46 (1,704.83, 4146.30) | 31.07 (3.10, 19.14) |
| MSE$_{\text{rel}}$ (%) (↓) | 1.55 (0.72, 1.91) | 21.11 (18.21, 23.85) | 97.40 (95.93, 98.39) | 1.78 (0.86, 2.21) | 2.16 (1.02, 2.82) | 295.44 (199.81, 363.65) | 1.06 (0.60, 1.06) |
| SSIM (↑) | 0.99 (0.99, 1.00) | 0.90 (0.87, 0.93) | 0.83 (0.80, 0.87) | 0.99 (0.99, 1.00) | 0.99 (0.99, 1.00) | 0.59 (0.50, 0.70) | 1.00 (1.00, 1.00) |

The table shows the mean values and in brackets the 25th and 75th percentiles for the 166 multispectral stacks from the in vivo test dataset.

the reconstructed image). The network can thus learn a data transform that is agnostic to specific characteristics of the ground-truth images during training and generalizes to images with any content (be it synthesized or in vivo)[19]. By contrast, the alternative model DeepMB$_{\text{initial-images}}$ trained on true synthetic initial pressure images (left side in Fig. 1a) falls short to accurately generalize to experimental test data and ultimately results in decreased reconstruction image quality for in vivo data as the underlying inverse problem is ill-posed. More specifically, true synthetic initial pressure images contain information not available in the input sinograms due to limited angle acquisition, measurement noise, and finite transducer bandwidth. To restore the missing information, DeepMB$_{\text{initial-images}}$ must incorporate information from the training data manifold, which hinders the correct processing of test data not contained in the training data manifold.

DeepMB supports dynamic adjustments of the SOS parameter during imaging to reconstruct high-resolution and in-focus images for arbitrary tissue types. Information about the SOS in the imaged region is required during optoacoustic image reconstruction to compute the travel time of acoustic signals between the source chromophores and the transducers of the imaging system, and to account for the spatial impulse response of the imaging system[15,16]. In practice, the optimal SOS for a reconstruction is a priori unknown and needs to be manually tuned during imaging. Following previous efforts to automatically correct for SOS-related aberrations, especially in heterogeneous media[48], future research may also aim at automatically inferring the optimal SOS from the optoacoustic input sinogram—either in a distinct antecedent step or directly within the deep-learning-based reconstruction.

The presented methodology to accelerate iterative model-based reconstruction is also applicable to other optoacoustic reconstruction approaches. For instance, frequency-band model-based reconstruction[49] or Bayesian optoacoustic reconstruction[50,51] can disentangle structures of different physical scales and quantifying reconstruction uncertainty, respectively, but their long reconstruction times currently hinder their use in real-time applications. The underlying methodology of DeepMB could also be exploited to accelerate parametrized (iterative) inversion approaches for other imaging modalities, such as ultrasound[52], X-ray computed tomography[18,53], magnetic resonance imaging[27–29,54], computed tomography[26], or, more generally, for any parametric partial differential equation[25]. We are currently working on embedding DeepMB into the hardware of a next-generation MSOT scanner, to use DeepMB for real-time imaging in clinical applications.

# Methods

## Handheld MSOT imaging system

We evaluated DeepMB with a modern MSOT scanner (MSOT Acuity Echo, iThera Medical GmbH). The system was equipped with a multiwavelength laser that illuminates tissues with short laser pulses (<10 ns) at a repetition rate of 25 Hz. The scanner featured a custom-made ultrasound detector (IMASONIC SAS) with the following characteristics: number of piezoelectric elements = 256; concavity radius = 4 cm; angular coverage = 125°; central frequency = 4 MHz. The parasitic noise generated by light-transducer interference was reduced via optical shielding of the matching layer, yielding an extended 153% frequency bandwidth. The raw channel data for each optoacoustic scan were recorded with a sampling frequency of 40 MHz in 50.75 µs, yielding a sinogram of 2,030 × 256 samples. Co-registered B-mode ultrasound images were acquired interleaved at approximately 6 Hz for live guidance and navigation. During imaging, optoacoustic backprojection images as well as B-mode ultrasound images were displayed in real-time on the scanner monitor for guidance.

## Acquisition of in vivo test sinograms

We scanned six healthy volunteers to collect in vivo data for DeepMB evaluation. Three females and three males particpated, aged from 20 to 36 years (mean age = 28.3 ± 5.7). Self-assessed skin colour, according to the Fitzpatrick scale, was type II (2 participants), type III (3 participants) and type IV (1 participant). Self-assessed body type was ectomorph (2 participants), mesomorph (3 participants) and endomorph (1 participant). We have complied with all relevant ethical regulations following the guidelines provided by Helmholtz Center Munich. All participants gave written informed consent upon recruitment.

For each participant, we scanned between 25 and 29 different combinations of anatomical locations and probe orientations: biceps, thyroid, carotid, calf (each left/right and transversal/longitudinal), elbow, neck, colon (each left/right) and breast (each left/right and top/bottom, female participants only). We conducted between one and four acquisitions for each combination of anatomical location and probe orientation. During each acquisition, we recorded sinograms for approximately 10 s at wavelengths cyclically iterating from 700 to 980 nm in steps of 10 nm. We then selected, per acquisition, the 29 consecutively acquired sinograms for which we observed minimal motion in the interleaved ultrasound images, amounting to a total of 4,814 in vivo test sinograms.

Finally, we band-pass filtered all selected in vivo sinograms between 100 kHz and 12 MHz to remove frequency components beyond the transducer bandwidth and cropped the first 110 time samples to remove device-specific noise present at the beginning of the sinograms.

## Determination of the SOS values

We manually tuned the SOS values of all in vivo test scans to evaluate DeepMB reconstructions under both in-focus and out-of-focus conditions. We used a SOS step size of 5 m s$^{-1}$ to enable SOS adjustments slightly below the system spatial resolution (approximatively 200 µm). We found that the optimal range of SOS values was 1,475–1,525 m s$^{-1}$ for the in vivo dataset, and we therefore used the same range to define the supported input SOS values of the DeepMB network.

For each scan, we manually selected the SOS value that resulted in the most well-focused reconstructed image. To speed up tuning, we selected the optimal SOS values on the basis of approximate and high-frequency-dominated reconstructions that we computed by applying the transpose model of the system to the recorded sinograms. Furthermore, we tuned the SOS for scans at 800 nm only, and adopted the values for all scans at other wavelengths acquired at the time, exploiting their spatial co-registration due to the absence of motion (see the above sections for details).

### Synthesis of sinograms for training and validation

For network training and validation, optoacoustic sinograms were synthesized with an accurate physical forward model of imaging process that incorporates the total impulse response of the system[15], parametrized by a SOS value drawn uniformly at random from the range 1,475–1,525 m s$^{-1}$ in 5 m s$^{-1}$ steps. Real-world images serving as initial pressure distributions for the forward simulations were randomly selected from the publicly available PASCAL Visual Object Classes Challenge 2012 (VOC2012) dataset[38], converted to monochannel grayscale and resized to 416 × 416 pixels. After the application of the forward model, each synthesized sinogram was scaled by a factor drawn uniformly at random from the 0–450 range to better match the variance observed in in vivo sinograms.

### Image reconstruction

We reconstructed all sinograms (synthetic as well as in vivo) via iterative-model-based reconstruction to generate the ground-truth optoacoustic images. We used Shearlet L$^1$ regularization to tackle the ill-posedness of the inverse problem. Shearlet L$^1$ regularization is a convex relaxation of Shearlet sparsity, which can reduce limited-view artefacts in reconstructed images, as Shearlets provide a maximally sparse approximation of a larger class of images (known as cartoon-like functions) with a mathematically proven optimal encoding rate[55]. The optimal pressure field is characterized as

$$p_0 \ := \ \underset{p\geq 0}{\operatorname{argmin}} ||M_{\mathrm{SOS}}p - s||_2^2 + \lambda ||\mathrm{SH}\,(p)\,||_1,$$

where $p_0$ is the reconstructed image, $M_{\mathrm{SOS}}$ is the forward model of the imaging process for the selected reconstruction SOS, $s$ is the input sinogram, $\lambda$ is the regularization parameter tuned via an L-curve, SH is the Shearlet transform and $\|\cdot\|_n$ is the $n$-norm. The minimization problem was solved via bound-constrained sparse reconstruction by separable approximation[56–58]. All images were reconstructed with a size of 416 × 416 pixels and a field of view of 4.16 × 4.16 cm$^2$. For comparison purposes, we also reconstructed all images using the backprojection formula[59,60] and the delay-multiply-and-sum with coherence factor algorithm[40,41].

### Network training

The DeepMB network was implemented in Python and PyTorch. It was trained—either on synthetic or in vivo data—for 300 epochs using stochastic gradient descent with batch size of 4, learning rate of 0.01, momentum of 0.99 and a per-epoch learning rate decay factor of 0.99. The network loss was calculated as the mean square error between the output image and the reference image. The final model was selected based on the minimal loss on the validation dataset, and compiled into an ONNX model for speed-up.

To facilitate training, all input sinograms were scaled by $K = 450^{-1}$ to ensure that their values never exceed the range [−1, 1]. The same scaling factor was also applied to all target images. Furthermore, the square root was applied to all target reference images used during training and validation to reduce the network output values and limit the influence of high intensity pixels during loss calculation. When applying the trained network on in vivo test data, inferred images were first squared then scaled by K$^{-1}$ to revert the preprocessing operation.

When training on synthetic data to build the standard DeepMB model, we used 8,000 sinograms as train split and 2,000 sinograms as validation split. The alternative scenario involving training on in vivo data to build the DeepMB$_{\text{in-vivo}}$ models was performed as described hereafter: six different permutations were conducted, with a 4/1/1 participants division between the train, validation and test splits, respectively, each participant being part of the validation and test splits once.

The DeepMB network is based upon the U-Net architecture[39] with a depth of five layers and a width of 64 features. To gradually reduce the total number of data channels from 267 (that is, 256 transducer elements, and one-hot encoding of 11 possible SOS values) down to 64, three 2D convolutional layers with 208, 160 and 112 features, respectively, were added prior to the U-Net. All kernel and padding size were (3, 3) and (1, 1), respectively. Biases were accounted for, and the final activation was the absolute value function.

### Data residual norm

To quantify the image fidelity of reconstructions from DeepMB, model-based, or backprojection, we evaluated the data residual norm R, defined as

$$R := \frac{||M_{\mathrm{SOS}}\,p_0 - s||_2^2}{||s||_2^2},$$

where $p_0$ is the reconstructed image, $M_{\mathrm{SOS}}$ is the forward model from model-based reconstruction, s is the input sinogram and $\|\cdot\|_2$ is the two-norm. Time sample values from the input sinogram that are outside the reach of the applied forward model are set to zero before computing the data residual norm to avoid distortions by signals originating from outside the field of view. We employed data residual norms as the primary evaluation metric for our experiments because it respects the underlying physics of the imaging process and is demonstrably minimal for model-based reconstruction. To constrain the solutions space for all reconstruction methods in a similar way and enable a meaningful comparison between backprojection on one hand, versus non-negative model-based and DeepMB on the other hand, negative pixel values were set to zero prior to residual calculation for backprojection images. All images were individually scaled using the linear degree of freedom of optoacoustic image reconstruction so that their data residual norms are minimal.

For the evaluation of in-focus images, data residual norms were calculated for the reconstructions with the optimal SOS values of all 4,814 samples from the in vivo test set. For the evaluation of out-of-focus images, data residuals were calculated for the reconstructions with all 11 SOS values of a subset of 638 randomly selected in vivo samples.

### Unmixing

To evaluate the multispectral image quality of DeepMB, model-based and backprojection, all reconstructed in-vivo scans from the test dataset were grouped into multispectral stacks of 29 images (respectively one scan from the range 700–980 nm in steps of 10 nm) and unmixed into oxyhaemoglobin, deoxyhaemoglobin, fat and water components:

$$\widehat{W} := \arg\min_{W\geq 0} ||S - WH||_{\mathrm{F}}^2,$$

where $S$ (size 173,056 × 29) denotes all pixels of a multispectral stack, $H$ (size 4 × 29) denotes the reference absorption spectra of water, fat, oxyhaemoglobin and deoxyhaemoglobin in the wavelength range 700–980 nm, and $\widehat{W}$ (size 173,056 × 4) denotes the unmixed components for the four considered chromophores; $||\mathrm{M}||_F := \left(\sum_{i,j} m_{i,j}^2\right)^{0.5}$ denotes the Frobenius norm and $M \geq 0$ refers to entry wise inequality.

All negative pixel values in the backprojection images were set to zero before unmixing.

### Image-based evaluation metrics

We also quantified the deviation of standard DeepMB, all alternative DeepMB, backprojection and delay-multiply-and-sum with coherence factor images from reference model-based reconstructions by computing the MAE, $\mathrm{MAE_{rel}}$, MSE, $\mathrm{MSE_{rel}}$ and SSIM, defined as

$$\mathrm{MAE}\ := \ ||i_{\mathrm{rec}} - i_{\mathrm{mb}}||_1,$$

$$\mathrm{MAE_{rel}}\ := \ \frac{||i_{\mathrm{rec}} - i_{\mathrm{mb}}||_1}{||i_{\mathrm{mb}}||_1},$$

$$\mathrm{MSE}\ := \ ||i_{\mathrm{rec}} - i_{\mathrm{mb}}||_2^2,$$

$$\mathrm{MSE_{rel}}\ := \ \frac{||i_{\mathrm{rec}} - i_{\mathrm{mb}}||_2^2}{||i_{\mathrm{mb}}||_2^2},$$

$$\mathrm{SSIM}\ := \ \frac{(2\mu_{\mathrm{rec}}\mu_{\mathrm{mb}} + c_1)(2\sigma_{\mathrm{rec,mb}} + c_2)}{(\mu_{\mathrm{rec}}^2 + \mu_{\mathrm{mb}}^2 + c_1)(\sigma_{\mathrm{rec}}^2 + \sigma_{\mathrm{mb}}^2 + c_2)},$$

where $i_{\mathrm{rec}}$ (size 173,056 × 1) is the vectorization of a reconstructed image from either standard DeepMB, any alternative DeepMB, backprojection or delay-multiply-and-sum with coherence factor and $i_{\mathrm{mb}}$ (size 173,056 × 1) is the vectorization of the corresponding reference image from model-based reconstruction. The SSIM is calculated as the average over sliding windows of size 21 × 21 pixels, where $\mu_{\mathrm{rec}}$ and $\mu_{\mathrm{mb}}$ are the averages of $i_{\mathrm{rec}}$ and $i_{\mathrm{mb}}$; $\sigma_{\mathrm{rec}}^2$ and $\sigma_{\mathrm{mb}}^2$ are the variances of $i_{\mathrm{rec}}$ and $i_{\mathrm{mb}}$; $\sigma_{\mathrm{rec,mb}}$ is the covariance of $i_{\mathrm{rec}}$ and $i_{\mathrm{mb}}$; and $c_1 = (0.01\max(i_{\mathrm{mb}}))^2$ and $c_2 = (0.03\max(i_{\mathrm{mb}}))^2$are two empirical variables to stabilize the division with weak denominators. All backprojection images were also preprocessed to enable a meaningful comparison with the model-based reference images: negative pixels were set to zero and all images were individually scaled using the linear degree of freedom in reconstructed optoacoustic images so that the respectively calculated metric is minimal.

Image-based metrics were computed analogously to the data residual norms using all 4,814 in vivo test samples (each reconstructed with the optimal SOS value) for the in-focus case and a subset of 638 in vivo test samples (each reconstructed with all 11 available SOS values) for the out-of-focus case.

### Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

## Data availability

In vivo data from two of the six scanned volunteers, the trained DeepMB model used in this work, and a download link for Pascal VOC 2012 dataset[38] used to synthesize training data for DeepMB are provided along with the source code on Github (https://github.com/juestellab/deepmb)[61]. In vivo data from the other four scanned volunteers cannot be shared due to privacy and consent restrictions.

## Code availability

The source code for DeepMB is publicly available on GitHub (https://github.com/juestellab/deepmb)[61].

## References

1. Ntziachristos, V. & Razansky, D. Molecular imaging by means of multispectral optoacoustic tomography (MSOT). *Chem. Rev.* **110**, 2783–2794 (2010).
2. Diot, G. et al. Multispectral optoacoustic tomography (MSOT) of human breast cancer. *Clin. Cancer Res.* **23**, 6912–6922 (2017).
3. Knieling, F. et al. Multispectral optoacoustic tomography for assessment of Crohn's disease activity. *N. Engl. J. Med.* **376**, 1292–1294 (2017).
4. Karlas, A. et al. Multispectral optoacoustic tomography of muscle perfusion and oxygenation under arterial and venous occlusion: a human pilot study. *J. Biophoton.* **13**, e201960169 (2020).
5. Dehner, C., Olefir, I., Chowdhury, K. B., Justel, D. & Ntziachristos, V. Deep-learning-based electrical noise removal enables high spectral optoacoustic contrast in deep tissue. *IEEE Trans. Med. Imaging* **41**, 3182–3193 (2022).
6. Kukacka, J. et al. Image processing improvements afford second-generation handheld optoacoustic imaging of breast cancer patients. *Photoacoustics* **26**, 100343 (2022).
7. Jüstel, D. et al. Spotlight on nerves: portable multispectral optoacoustic imaging of peripheral nerve vascularization and morphology. *Adv. Sci.* **10**, 2301322 (2023).
8. Regensburger, A. P. et al. Detection of collagens by multispectral optoacoustic tomography as an imaging biomarker for Duchenne muscular dystrophy. *Nat. Med.* **25**, 1905–1915 (2019).
9. Dima, A. & Ntziachristos, V. Non-invasive carotid imaging using optoacoustic tomography. *Opt. Express* **20**, 25044–25057 (2012).
10. Taruttis, A. & Ntziachristos, V. Advances in real-time multispectral optoacoustic imaging and its applications. *Nat. Photon.* **9**, 219–227 (2015).
11. Ivankovic, I., Mercep, E., Schmedt, C. G., Dean-Ben, X. L. & Razansky, D. Real-time volumetric assessment of the human carotid artery: handheld multispectral optoacoustic tomography. *Radiology* **291**, 45–50 (2019).
12. Sethuraman, S., Aglyamov, S. R., Amirian, J. H., Smalling, R. W. & Emelianov, S. Y. Intravascular photoacoustic imaging using an IVUS imaging catheter. *IEEE Trans. Ultrason. Ferroelectr. Freq. Control* **54**, 978–986 (2007).
13. Yang, J. M. et al. Photoacoustic endoscopy. *Opt. Lett.* **34**, 1591–1593 (2009).
14. Xu, M. & Wang, L. V. Universal back-projection algorithm for photoacoustic computed tomography. *Phys. Rev. E* **71**, 016706 (2005).
15. Chowdhury, K. B., Prakash, J., Karlas, A., Jüstel, D. & Ntziachristos, V. A synthetic total impulse response characterization method for correction of hand-held optoacoustic images. *IEEE Trans. Med. Imaging* **39**, 3218–3230 (2020).
16. Chowdhury, K. B., Bader, M., Dehner, C., Justel, D. & Ntziachristos, V. Individual transducer impulse response characterization method to improve image quality of array-based handheld optoacoustic tomography. *Opt. Lett.* **46**, 1–4 (2021).
17. Ding, L., Dean-Ben, X. L. & Razansky, D. Real-time model-based inversion in cross-sectional optoacoustic tomography. *IEEE Trans. Med. Imaging* **35**, 1883–1891 (2016).
18. Jin, K. H., McCann, M. T., Froustey, E. & Unser, M. Deep convolutional neural network for Inverse problems in imaging. *IEEE Trans. Image Process.* **26**, 4509–4522 (2017).
19. Zhu, B., Liu, J. Z., Cauley, S. F., Rosen, B. R. & Rosen, M. S. Image reconstruction by domain-transform manifold learning. *Nature* **555**, 487–492 (2018).
20. Ongie, G. et al. Deep learning techniques for inverse problems in imaging. *IEEE J. Sel. Areas Information Theory* **1**, 39–56 (2020).
21. Lucas, A., Iliadis, M., Molina, R. & Katsaggelos, A. K. Using deep neural networks for inverse problems in imaging: beyond analytical methods. *IEEE Signal Process Mag.* **35**, 20–36 (2018).
22. Gröhl, J., Schellenberg, M., Dreher, K. & Maier-Hein, L. Deep learning for biomedical photoacoustic imaging: a review. *Photoacoustics* **22**, 100241 (2021).

23. Hauptmann, A. & Cox, B. Deep learning in photoacoustic tomography: current approaches and future directions. *J. Biomed. Opt.* **25**, 112903 (2020).
24. Reiter, A. & Bell, M. A. L. A machine learning approach to identifying point source locations in photoacoustic data. In *Photons Plus Ultrasound: Imaging and Sensing* 100643J (SPIE, 2017).
25. Aggarwal, H. K., Mani, M. P. & Jacob, M. MoDL: model-based deep learning architecture for inverse problems. *IEEE Trans. Med. Imaging* **38**, 394–405 (2019).
26. Liu, J. et al. SGD-Net: efficient model-based deep learning with theoretical guarantees. *IEEE Trans. Comput. Imaging* **7**, 598–610 (2021).
27. Genzel, M., Macdonald, J. & Marz, M. Solving inverse problems with deep neural networks—robustness Included. *IEEE Trans. Pattern Anal. Mach. Intell.* **45**, 1119–1134 (2022).
28. Schlemper, J., Caballero, J., Hajnal, J. V., Price, A. N. & Rueckert, D. A deep cascade of convolutional neural networks for dynamic MR image reconstruction. *IEEE Trans. Med. Imaging* **37**, 491–503 (2018).
29. Hammernik, K. et al. Learning a variational network for reconstruction of accelerated MRI data. *Magn. Reson. Med.* **79**, 3055–3071 (2018).
30. Kim, M., Jeng, G. S., Pelivanov, I. & O'Donnell, M. Deep-learning image reconstruction for real-time photoacoustic system. *IEEE Trans. Med. Imaging* **39**, 3379–3390 (2020).
31. Lan, H., Jiang, D., Yang, C., Gao, F. & Gao, F. Y-Net: hybrid deep learning image reconstruction for photoacoustic tomography in vivo. *Photoacoustics* **20**, 100197 (2020).
32. Waibel, D. et al. Reconstruction of initial pressure from limited view photoacoustic images using deep learning. In *Photons Plus Ultrasound: Imaging and Sensing* 104942S (SPIE, 2018).
33. Feng, J. et al. End-to-end Res-Unet based reconstruction algorithm for photoacoustic imaging. *Biomed. Opt. Express* **11**, 5321–5340 (2020).
34. Tong, T. et al. Domain transform network for photoacoustic tomography from limited-view and sparsely sampled data. *Photoacoustics* **19**, 100190 (2020).
35. Guan, S., Khan, A. A., Sikdar, S. & Chitnis, P. V. Limited-view and sparse photoacoustic tomography for neuroimaging with deep learning. *Sci. Rep.* **10**, 8510 (2020).
36. Guo, M., Lan, H., Yang, C., Liu, J. & Gao, F. AS-Net: fast photoacoustic reconstruction with multi-feature fusion from sparse data. *IEEE Trans. Comput. Imaging* **8**, 215–223 (2022).
37. Hauptmann, A. et al. Model-based learning for accelerated, limited-view 3-D photoacoustic tomography. *IEEE Trans. Med. Imaging* **37**, 1382–1393 (2018).
38. Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J. & Zisserman, A. The Pascal visual object classes (VOC) challenge. *Int. J. Comput. Vision* **88**, 303–338 (2010).
39. Ronneberger, O., Fischer, P. & Brox, T. U-Net: convolutional networks for biomedical image segmentation. In *International Conference on Medical image Computing and Computer-Assisted Intervention* 234–241 (Springer, 2015).
40. Jeon, S. et al. Real-time delay-multiply-and-sum beamforming with coherence factor for in vivo clinical photoacoustic imaging of humans. *Photoacoustics* **15**, 100136 (2019).
41. Matrone, G., Savoia, A. S., Caliano, G. & Magenes, G. The delay multiply and sum beamforming algorithm in ultrasound B-mode medical imaging. *IEEE Trans. Med. Imaging* **34**, 940–949 (2015).
42. Rosenthal, A., Ntziachristos, V. & Razansky, D. Acoustic inversion in optoacoustic tomography: a review. *Curr. Med. Imaging Rev.* **9**, 318–336 (2013).
43. Prahl, S. *Assorted Spectra* (accessed 19 January 2023); https://omlc.org/spectra/
44. Tobin, J. et al. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems* 23–30 (IEEE, 2017).
45. Mårtensson, G. et al. The reliability of a deep learning model in clinical out-of-distribution MRI data: a multicohort study. *Med. Image Anal.* **66**, 101714 (2020).
46. Susmelj, A. K. et al. Signal domain learning approach for optoacoustic image reconstruction from limited view data. In *Proc. 5th International Conference on Medical Imaging with Deep Learning* 1173–1191 (PMLR, 2022).
47. Schellenberg, M. et al. Photoacoustic image synthesis with generative adversarial networks. *Photoacoustics* **28**, 100402 (2022).
48. Jeon, S., Choi, W., Park, B. & Kim, C. A deep learning-based model that reduces speed of sound aberrations for improved in vivo photoacoustic imaging. *IEEE Trans. Image Process.* **30**, 8773–8784 (2021).
49. Longo, A., Justel, D. & Ntziachristos, V. Disentangling the frequency content in optoacoustics. *IEEE Trans. Med. Imaging* **41**, 3373–3384 (2022).
50. Tick, J., Pulkkinen, A. & Tarvainen, T. Image reconstruction with uncertainty quantification in photoacoustic tomography. *J. Acoust. Soc. Am.* **139**, 1951 (2016).
51. Tick, J. et al. Three dimensional photoacoustic tomography in Bayesian framework. *J. Acoust. Soc. Am.* **144**, 2061 (2018).
52. Hyun, D., Brickson, L. L., Looby, K. T. & Dahl, J. J. Beamforming and Speckle Reduction Using Neural Networks. *IEEE Trans. Ultrason. Ferroelectr. Freq. Control* **66**, 898–910 (2019).
53. Kang, E., Min, J. & Ye, J. C. A deep convolutional neural network using directional wavelets for low-dose X-ray CT reconstruction. *Med. Phys.* **44**, e360–e375 (2017).
54. Moya-Sáez, E., Peña-Nogales, Ó., Luis-García, R. D. & Alberola-López, C. A deep learning approach for synthetic MRI based on two routine sequences and training with synthetic data. *Comput. Methods Programs Biomed.* **210**, 106371 (2021).
55. Kutyniok, G. & Lim, W.-Q. Compactly supported shearlets are optimally sparse. *J. Approx. Theory* **163**, 1564–1589 (2011).
56. Wright, S. J., Nowak, R. D. & Figueiredo, M. A. T. Sparse reconstruction by separable approximation. *IEEE Trans. Signal Process.* **57**, 2479–2493 (2009).
57. Chartrand, R. & Wohlberg, B. Total-variation regularization with bound constraints. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing* 766–769 (IEEE, 2010).
58. Kutyniok, G., Lim, W.-Q. & Reisenhofer, R. ShearLab 3D: faithful digital shearlet transforms based on compactly supported shearlets. In *ACM Transactions on Mathematical Software* 1–42 (ACM, 2016).
59. Kunyansky, L. A. Explicit inversion formulae for the spherical mean Radon transform. *Inverse Prob.* **23**, 373–383 (2007).
60. Kuchment, P. & Kunyansky, L. in *Handbook of Mathematical Methods in Imaging* (ed. Scherzer, O.) 817–865 (Springer, 2011).
61. Dehner, C. & Zahnd, G. *DeepMB v1.0.0* (Zenodo, 2023); https://doi.org/10.5281/zenodo.8169175

## Acknowledgements

We would like to thank A. Longo for her precious contribution during in vivo image acquisition and the conception of Fig. 2, and R. Wilson for his attentive reading and improvements of the manuscript. This project has received funding from the Bavarian Ministry of Economic Affairs, Energy and Technology (StMWi) (DIE-2106-0005// DIE0161/02, DeepOPUS, granted to D.J.) and from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme under grant agreement no. 694968 (PREMSOT, granted to V.N.).

## Author contributions

C.D. and G.Z. contributed equally to this work. C.D., G.Z., and D.J. conceptualized the initial idea. C.D. and G.Z. implemented the algorithm, conducted the experiments, analysed the results, and wrote the manuscript. D.J. and V.N. supervised the work. All authors provided feedback and approved the manuscript.

## Competing interests

V.N. is an equity owner in and consultant for iThera Medical GmbH. G.Z. and C.D. are employees of iThera Medical GmbH. D.J., C.D. and G.Z. are inventors in patent applications related to DeepMB (patent nos. EP22177153.8 and PCT/EP2023/064714).

## Additional information

**Extended data** is available for this paper at https://doi.org/10.1038/s42256-023-00724-3.

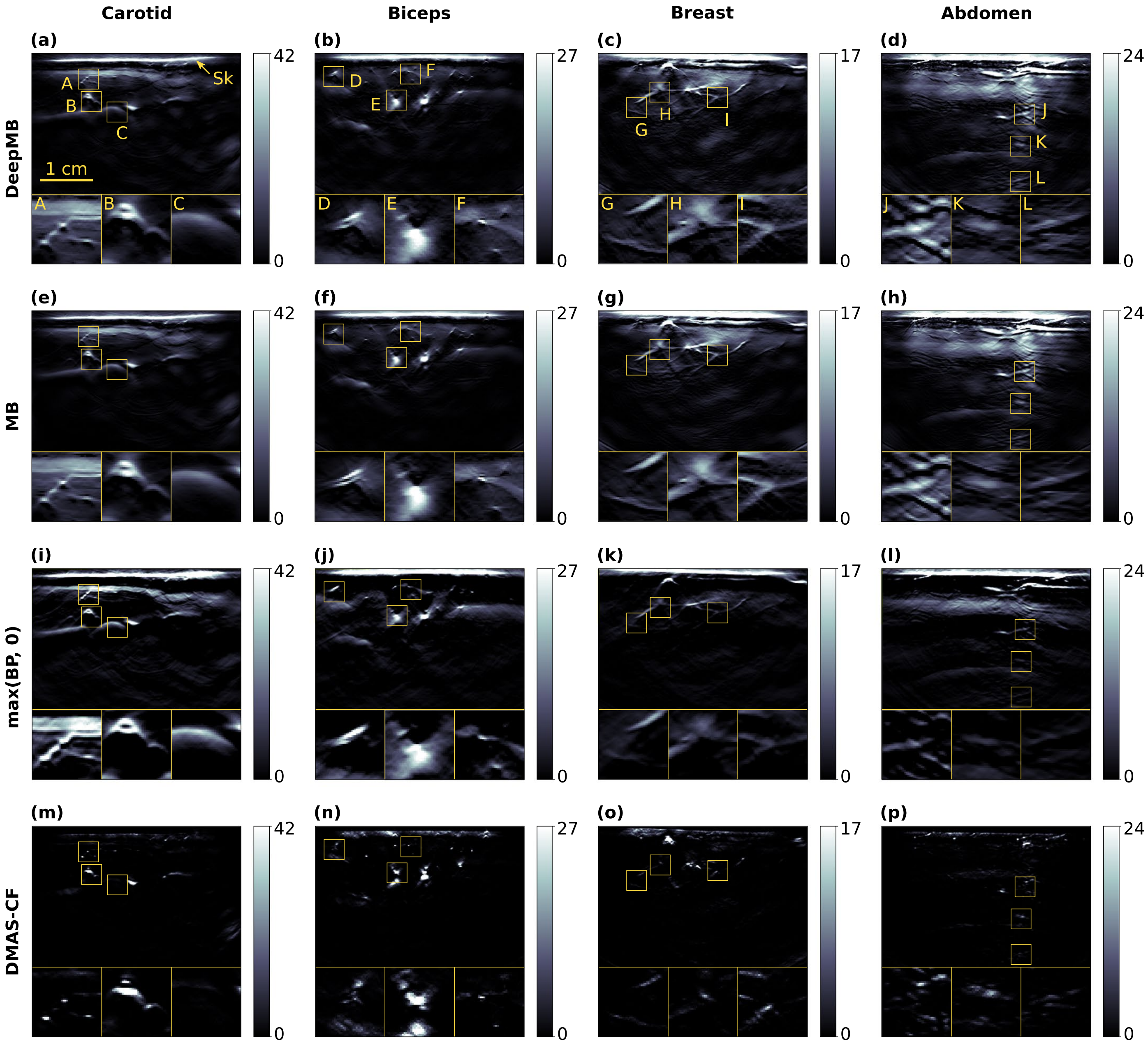
**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s42256-023-00724-3.

**Correspondence and requests for materials** should be addressed to Vasilis Ntziachristos or Dominik Jüstel.

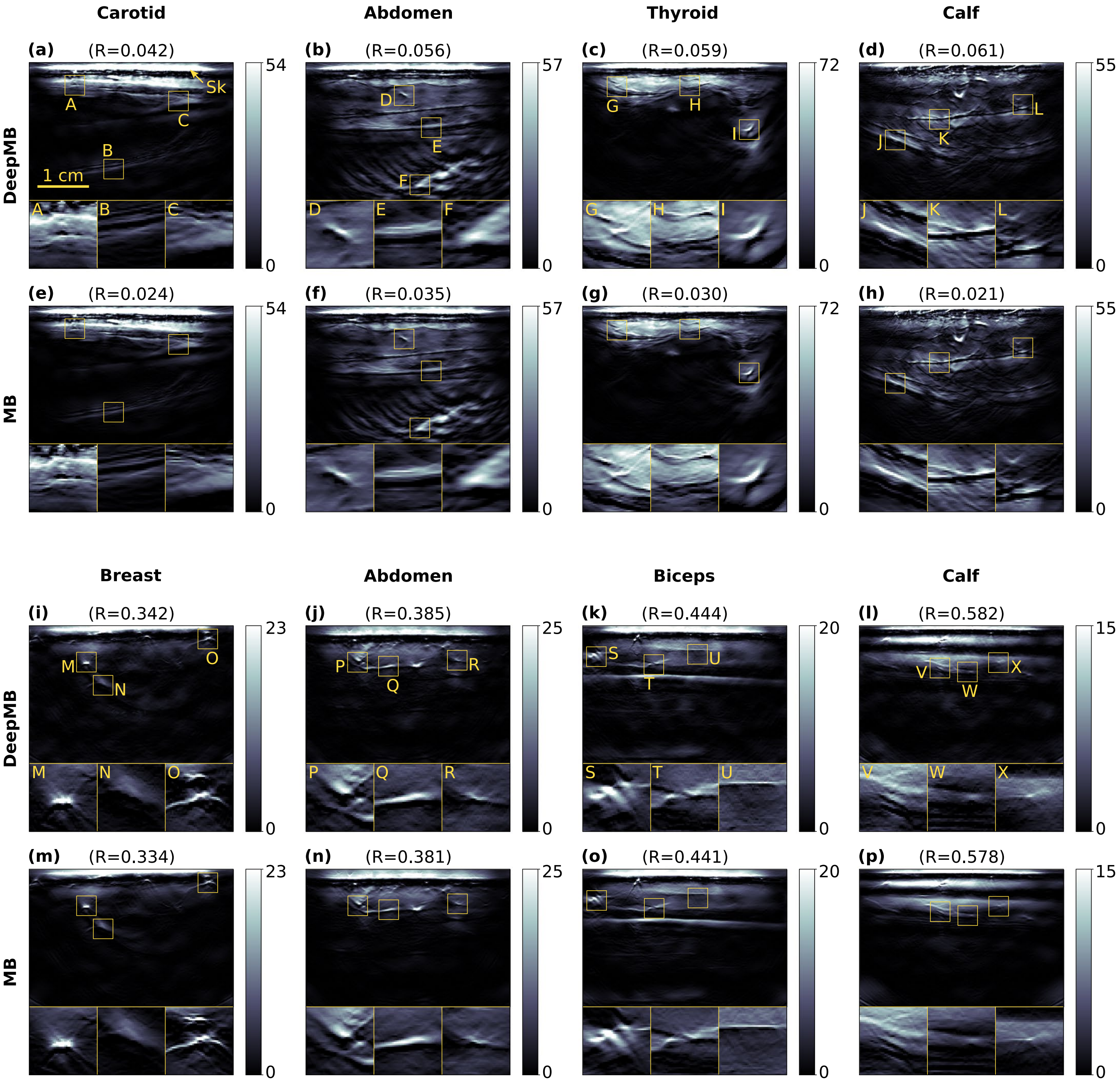
**Peer review information** *Nature Machine Intelligence* thanks Andreas Hauptmann, and the other, anonymous, reviewers for their contribution to the peer review of this work. Primary Handling Editor: Jacob Huth, in collaboration with the *Nature Machine Intelligence* editorial team.

**Reprints and permissions information** is available at www.nature.com/reprints.

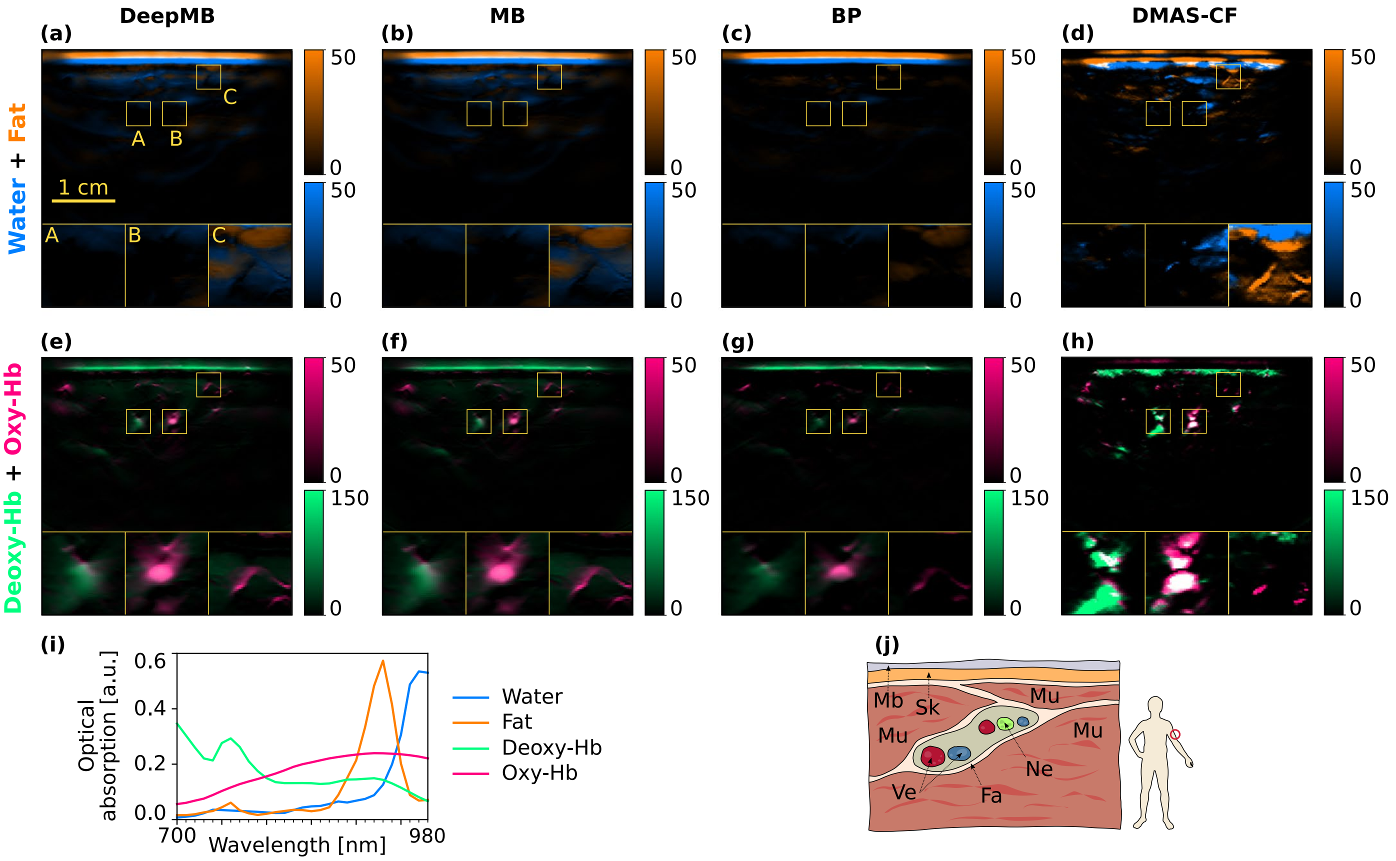
**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



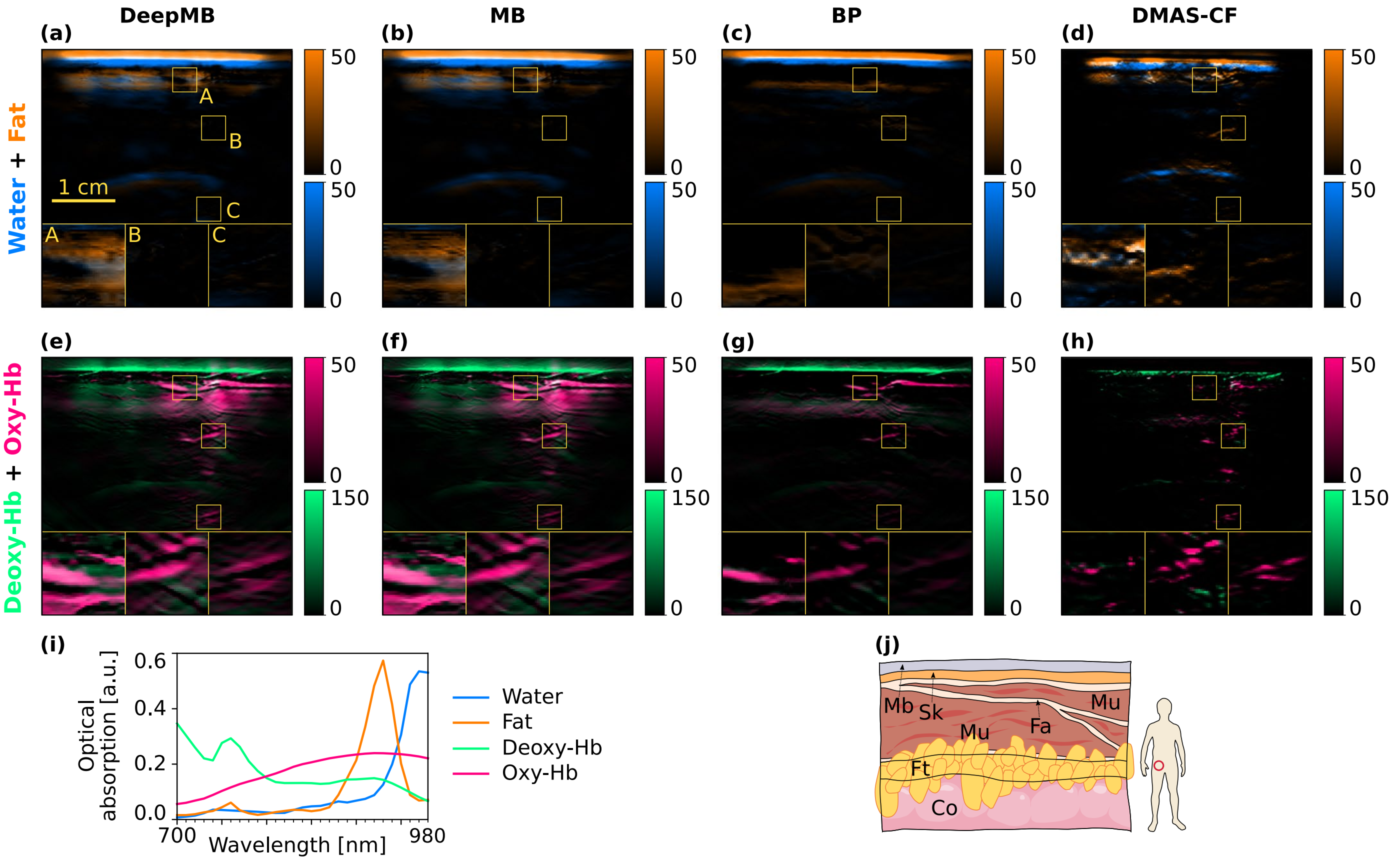
**Extended Data Fig. 1 | Visual comparison of backprojection images with negative pixel values set to zero after the reconstruction (BP, i-l) and delay-multiply-and-sum with coherence factor (DMAS-CF, m-p) images, against the corresponding deep model-based (DeepMB, a-d) and model-based (MB, e-h) images.** Visual comparison of backprojection (BP) images with negative pixel values set to zero after the reconstruction (third row) and delay-multiply-and-sum with coherence factor (DMAS-CF, fourth row) images, against the corresponding deep model-based (DeepMB) and model-based (MB) images (first two rows). The presented samples are the same as those depicted in Fig. 2. DeepMB and MB images are nearly identical; BP images notably differ from reference model-based reconstructions suffering from lower resolution (see for example structures shown in zoom A of tile i and zoom D of tile j), missing structures in image regions that contained negative pixel values (see for example zoom F of tile j, or the entire region below the skin line (Sk) in tile k and l), and reduced contrast (see for example structures shown in zoom I of tile k and zoom J of tile l). All images show the reconstructed initial pressure in arbitrary units and were slightly cropped to a field of view of 4.16 × 2.80 cm$^2$ to disregard the area occupied by the probe couplant above the skin line.



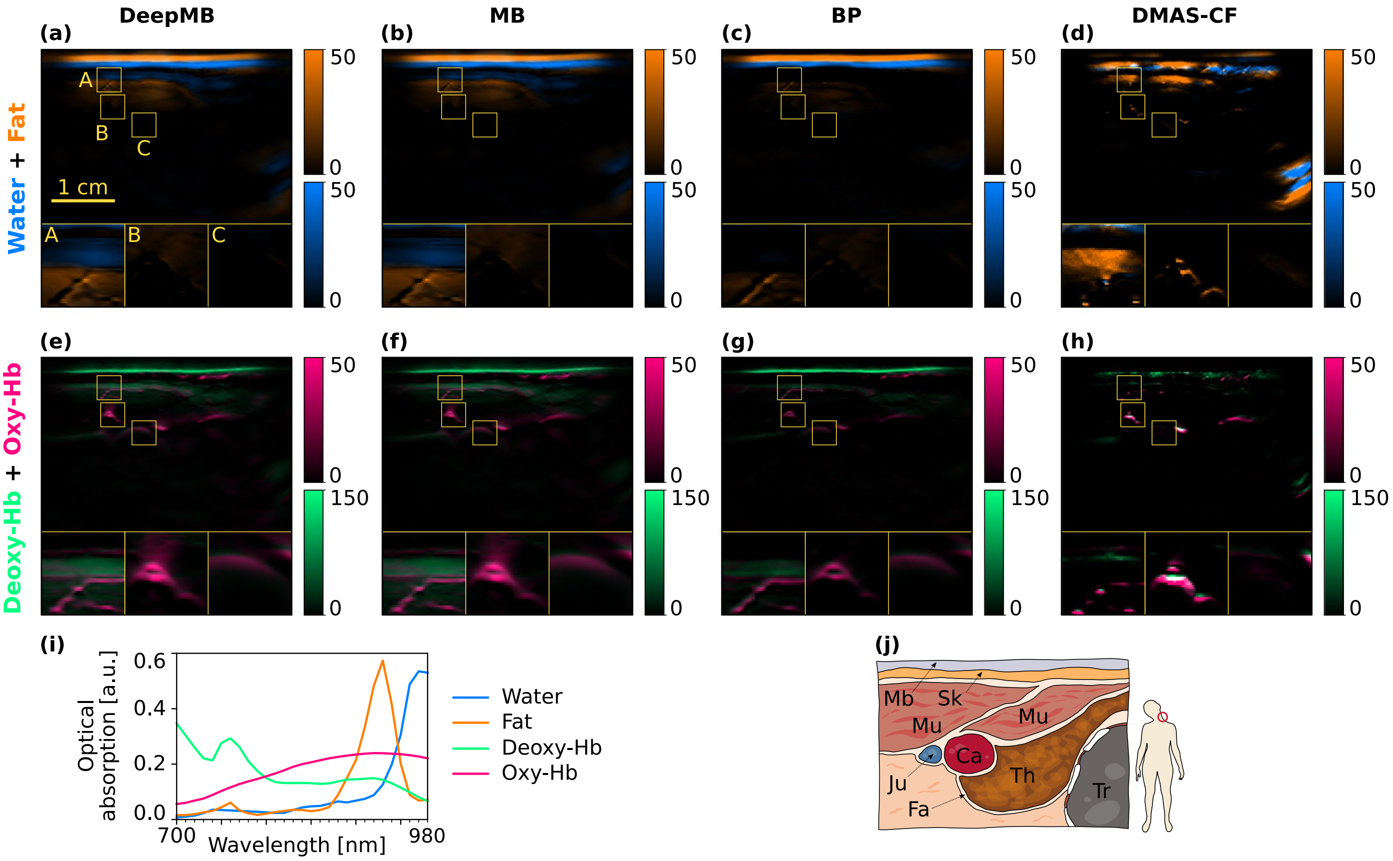
**Extended Data Fig. 2 | Examples of deep model-based and model-based images with low and high data residual norms.** Examples from the in vivo test dataset with low and high data residual norms (namely, below the 5th percentile (**a-h**) and above the 95th percentile (**i-p**) of all 4814 test samples, respectively), for deep model-based (DeepMB) and model-based (MB). The data residual norm (R) is indicated between round brackets above each image. Panels (**a, e**) and (**l, p**) correspond to the samples for which DeepMB afforded the overall lowest and highest data residual norms, respectively. All images show the reconstructed initial pressure in arbitrary units and were slightly cropped to a field of view of 4.16 × 2.80 cm$^2$ to disregard the area occupied by the probe couplant above the skin line (Sk).



**Extended Data Fig. 3 | Unmixing of a multispectral biceps scan for deep model-based, model-based, backprojection, and delay-multiply-and-sum with coherence factor reconstructions.** Unmixing of a representative multispectral biceps scan for deep model-based (DeepMB; **a, e**), model-based (MB; **b, f**), backprojection (BP; **c, g**), and delay-multiply-and-sum with coherence factor (DMAS-CF; **d, h**). The unmixed components for fat and water and for oxyheamoglobin and deoxyhaemoglobin are shown in the first two rows, respectively. The third row depicts the reference absorption spectra of the four chromophores used during unmixing **(i)** and a schematic sketch of the anatomical context for the depicted scan **(j)**. All optoacoustic images show the unmixed components in arbitrary units and were slightly cropped to a field of view of 4.16 × 2.80 cm$^2$ to disregard the area occupied by the probe couplant above the skin line. Mb: probe membrane, Sk: skin, Fa: fascia, Mu: muscle, Ve: blood vessel, Ne: nerve.

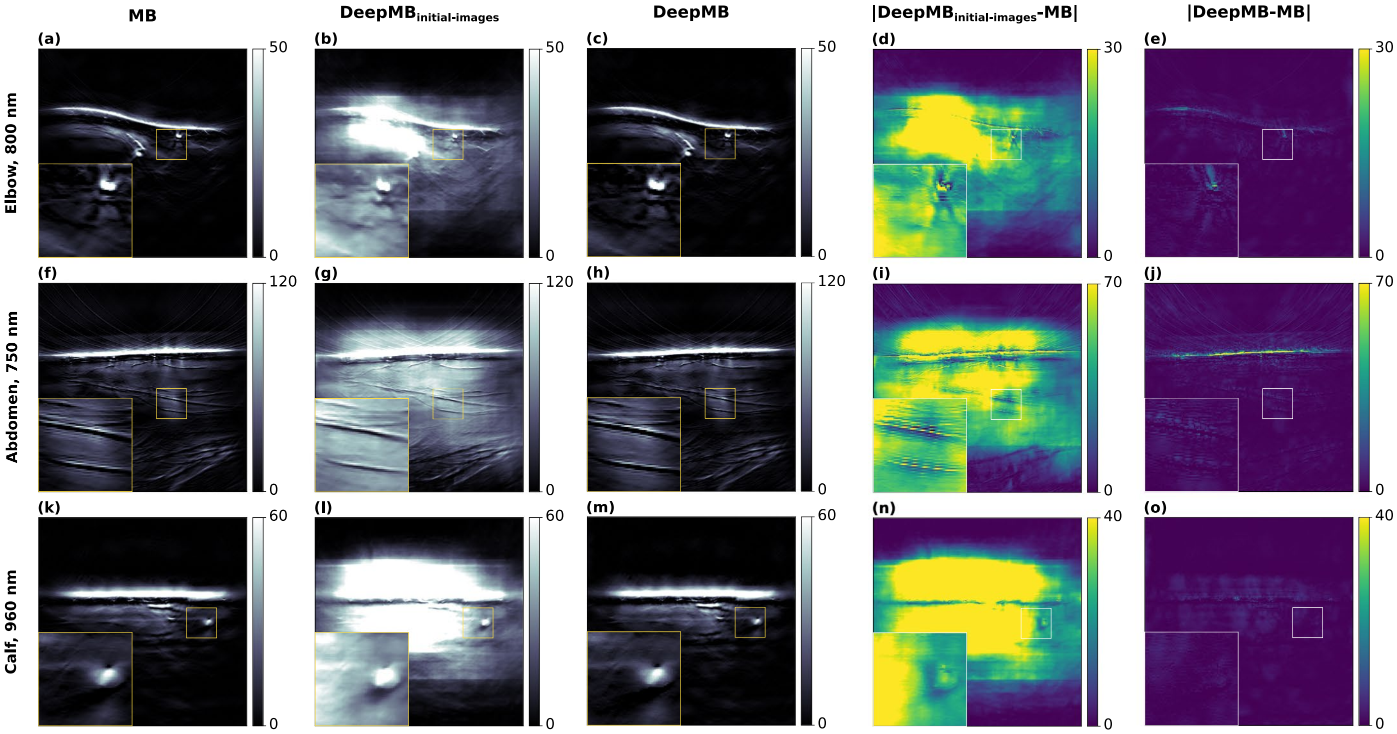


**Extended Data Fig. 4 | Unmixing of a multispectral abdomen scan for deep model-based, model-based, backprojection, and delay-multiply-and-sum with coherence factor reconstructions.** Unmixing of a representative multispectral abdomen scan for deep model-based (DeepMB; **a, e**), model-based (MB; **b, f**), backprojection (BP; **c, g**) and delay-multiply-and-sum with coherence factor (DMAS-CF; **d, h**). The unmixed components for fat and water and for oxyhaemoglobin and deoxyhaemoglobin are shown in the first two rows, respectively. The third row depicts the reference absorption spectra of the four chromophores used during unmixing **(i)** and a schematic sketch of the anatomical context for the depicted scan **(j)**. All optoacoustic images show the unmixed components in arbitrary units and were slightly cropped to a field of view of 4.16 × 2.80 cm$^2$ to disregard the area occupied by the probe couplant above the skin line. Mb: probe membrane, Sk: skin, Fa: fascia, Mu: muscle, Ft: fat, Co: colon.



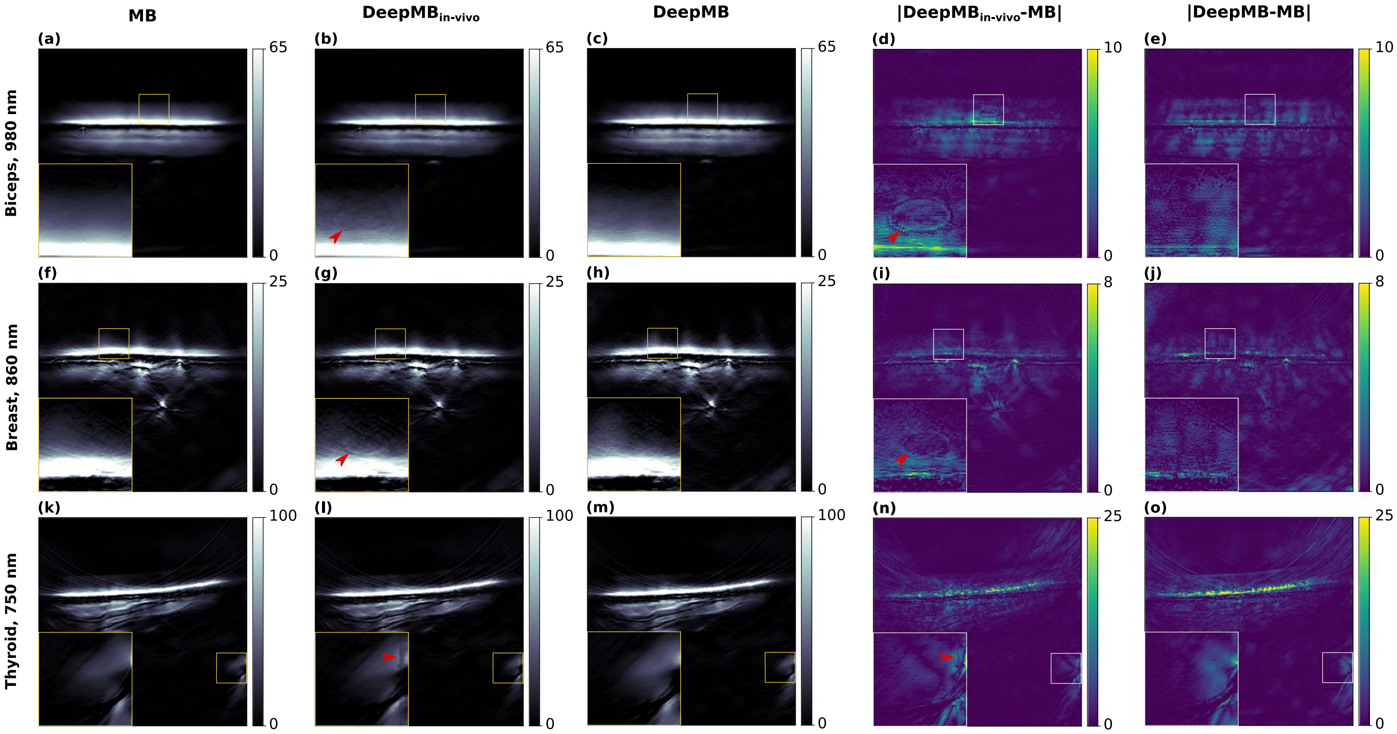
**Extended Data Fig. 5 | Unmixing of a multispectral carotid scan for deep model-based, model-based, backprojection, and delay-multiply-and-sum with coherence factor reconstructions.** Unmixing of a representative multispectral carotid scan for deep model-based (DeepMB; **a, e**), model-based (MB; **b, f**), backprojection (BP; **c, g**) and delay-multiply-and-sum with coherence factor (DMAS-CF; **d, h**). The unmixed components for fat and water and for oxyhaemoglobin and deoxyhaemoglobin are shown in the first two rows, respectively. The third row depicts the reference absorption spectra of the four chromophores used during unmixing **(i)** and a schematic sketch of the anatomical context for the depicted scan **( j)**. All optoacoustic images show the unmixed components in arbitrary units and were slightly cropped to a field of view of 4.16 × 2.80 cm$^2$ to disregard the area occupied by the probe couplant above the skin line. Mb: probe membrane, Sk: skin, Fa: fascia, Mu: muscle, Ca: common carotid artery, Ju: jugular vein, Th: thyroid, Tr: trachea.

https://doi.org/10.1038/s42256-023-00724-3



**Extended Data Fig. 6 | Example images from the alternative model DeepMB$_{\text{initial-images}}$ trained using true initial pressure reference images.** Representative examples showing the inaptitude of the alternative model DeepMB$_{\text{initial-images}}$ (that is, trained on true initial pressure images) to reconstruct in vivo images. The three rows depict different anatomies (elbow: **a**–**e**, abdomen: **f**–**j**, calf: **k**–**o**). The three leftmost columns correspond to images reconstructed via model-based (MB), alternative DeepMB$_{\text{initial-images}}$, and standard DeepMB. The two rightmost columns show the absolute differences between the reference model-based image and the image inferred from DeepMB$_{\text{initial-images}}$ and DeepMB, respectively. The field of view is 4.16 × 4.16 cm$^2$, the enlarged region is 0.61 × 0.61 cm$^2$.

https://doi.org/10.1038/s42256-023-00724-3



**Extended Data Fig. 7 | Example images from the alternative model DeepMB$_{\text{in-vivo}}$ trained using in vivo data.** Representative examples of reconstruction artefacts (red arrows) from alternative models DeepMB$_{\text{in-vivo}}$ (that is, trained on in vivo data instead of synthesized data). The three rows depict different anatomies (biceps: **a**–**e**, breast: **f**–**j**, thyroid: **k**–**o**). The three leftmost columns correspond to images reconstructed via model-based (MB), alternative DeepMB trained on in vivo data (DeepMB$_{\text{in-vivo}}$), and standard DeepMB (DeepMB). The two rightmost columns show the absolute differences between the reference model-based image and the image inferred from DeepMB$_{\text{in-vivo}}$ and DeepMB, respectively. The field of view is 4.16 × 4.16 cm$^2$, the enlarged region is 0.61 × 0.61 cm$^2$.

**Extended Data Table 1 | Quantitative evaluation of deep model-based and model-based reconstructions for different aggregations of the in vivo test dataset**

| Aggregation | Categories | DeepMB | MB |
|---|---|---|---|
| Entire in vivo test set | All images (n=4814) | 0.156 [0.092, 0.189] | 0.139 [0.068, 0.180] |
| Anatomical regions | Biceps (n=725) | 0.214 [0.118, 0.296] | 0.202 [0.108, 0.287] |
| | Breast (n=435) | 0.179 [0.118, 0.224] | 0.166 [0.106, 0.213] |
| | Calf (n=696) | 0.152 [0.087, 0.175] | 0.133 [0.058, 0.168] |
| | Carotid (n=754) | 0.128 [0.089, 0.159] | 0.115 [0.075, 0.148] |
| | Colon (n=870) | 0.152 [0.094, 0.193] | 0.136 [0.068, 0.183] |
| | Elbow (n=377) | 0.129 [0.087, 0.157] | 0.094 [0.037, 0.133] |
| | Neck (n=203) | 0.128 [0.083, 0.156] | 0.109 [0.053, 0.145] |
| | Thyroid (n=754) | 0.143 [0.079, 0.165] | 0.128 [0.059, 0.155] |
| Participants | 01 (n=667) | 0.148 [0.094, 0.174] | 0.132 [0.068, 0.163] |
| | 02 (n=638) | 0.084 [0.065, 0.097] | 0.050 [0.028, 0.069] |
| | 03 (n=1015) | 0.101 [0.074, 0.123] | 0.083 [0.045, 0.113] |
| | 04 (n=899) | 0.254 [0.174, 0.318] | 0.245 [0.166, 0.310] |
| | 05 (n=986) | 0.183 [0.130, 0.218] | 0.172 [0.120, 0.208] |
| | 06 (n=609) | 0.141 [0.102, 0.164] | 0.126 [0.088, 0.153] |
| Fitzpatrick scale | 2 (n=1566) | 0.209 [0.131, 0.280] | 0.197 [0.120, 0.270] |
| | 3 (n=2610) | 0.141 [0.094, 0.168] | 0.127 [0.076, 0.159] |
| | 4 (n=986) | 0.084 [0.065, 0.097] | 0.050 [0.028, 0.069] |
| Body type | Endomorph (n=1914) | 0.173 [0.094, 0.229] | 0.159 [0.075, 0.217] |
| | Mesomorph (n=1914) | 0.125 [0.080, 0.149] | 0.103 [0.045, 0.137] |
| | Ectomorph (n=986) | 0.183 [0.130, 0.218] | 0.172 [0.120, 0.208] |
| Wavelengths (nm) | 700 (n=166) | 0.142 [0.082, 0.181] | 0.108 [0.034, 0.158] |
| | 710 (n=166) | 0.140 [0.080, 0.178] | 0.111 [0.036, 0.163] |
| | 720 (n=166) | 0.142 [0.079, 0.179] | 0.114 [0.037, 0.164] |
| | 730 (n=166) | 0.142 [0.076, 0.178] | 0.116 [0.035, 0.165] |
| | 740 (n=166) | 0.142 [0.076, 0.183] | 0.118 [0.036, 0.172] |
| | 750 (n=166) | 0.142 [0.077, 0.189] | 0.120 [0.038, 0.180] |
| | 760 (n=166) | 0.144 [0.076, 0.194] | 0.122 [0.041, 0.181] |
| | 770 (n=166) | 0.150 [0.077, 0.200] | 0.130 [0.042, 0.188] |
| | 780 (n=166) | 0.159 [0.080, 0.214] | 0.139 [0.044, 0.203] |
| | 790 (n=166) | 0.167 [0.082, 0.225] | 0.147 [0.047, 0.215] |
| | 800 (n=166) | 0.172 [0.085, 0.238] | 0.153 [0.051, 0.229] |
| | 810 (n=166) | 0.175 [0.088, 0.238] | 0.157 [0.055, 0.227] |
| | 820 (n=166) | 0.178 [0.090, 0.247] | 0.161 [0.058, 0.233] |
| | 830 (n=166) | 0.181 [0.089, 0.245] | 0.165 [0.065, 0.236] |
| | 840 (n=166) | 0.180 [0.092, 0.248] | 0.164 [0.068, 0.241] |
| | 850 (n=166) | 0.185 [0.096, 0.255] | 0.170 [0.076, 0.247] |
| | 860 (n=166) | 0.188 [0.104, 0.260] | 0.173 [0.082, 0.252] |
| | 870 (n=166) | 0.186 [0.109, 0.258] | 0.172 [0.092, 0.250] |
| | 880 (n=166) | 0.181 [0.112, 0.251] | 0.168 [0.097, 0.243] |
| | 890 (n=166) | 0.191 [0.120, 0.263] | 0.178 [0.102, 0.254] |
| | 900 (n=166) | 0.181 [0.122, 0.241] | 0.169 [0.109, 0.234] |
| | 910 (n=166) | 0.148 [0.121, 0.181] | 0.138 [0.111, 0.174] |
| | 920 (n=166) | 0.126 [0.111, 0.142] | 0.117 [0.103, 0.136] |
| | 930 (n=166) | 0.115 [0.103, 0.129] | 0.107 [0.095, 0.121] |
| | 940 (n=166) | 0.130 [0.117, 0.148] | 0.122 [0.107, 0.142] |
| | 950 (n=166) | 0.152 [0.129, 0.178] | 0.146 [0.120, 0.173] |
| | 960 (n=166) | 0.133 [0.105, 0.156] | 0.127 [0.099, 0.153] |
| | 970 (n=166) | 0.123 [0.095, 0.144] | 0.118 [0.089, 0.140] |
| | 980 (n=166) | 0.119 [0.094, 0.138] | 0.114 [0.088, 0.134] |
| Speed of sound (m/s) | 1475 (n=58) | 0.241 [0.188, 0.300] | 0.233 [0.183, 0.291] |
| | 1480 (n=203) | 0.199 [0.127, 0.262] | 0.187 [0.119, 0.252] |
| | 1485 (n=261) | 0.241 [0.179, 0.287] | 0.233 [0.174, 0.276] |
| | 1490 (n=406) | 0.190 [0.121, 0.248] | 0.176 [0.106, 0.235] |
| | 1495 (n=464) | 0.146 [0.077, 0.178] | 0.125 [0.043, 0.166] |
| | 1500 (n=1131) | 0.156 [0.086, 0.193] | 0.137 [0.056, 0.182] |
| | 1505 (n=754) | 0.131 [0.089, 0.160] | 0.112 [0.059, 0.148] |
| | 1510 (n=725) | 0.123 [0.083, 0.148] | 0.105 [0.056, 0.138] |
| | 1515 (n=493) | 0.162 [0.094, 0.192] | 0.149 [0.078, 0.191] |
| | 1520 (n=174) | 0.135 [0.099, 0.161] | 0.124 [0.087, 0.155] |
| | 1525 (n=145) | 0.139 [0.088, 0.192] | 0.129 [0.080, 0.185] |

Quantitative evaluation of the image fidelity of deep model-based (DeepMB) and model-based (MB) reconstructions for different aggregations of the in-focus in vivo test dataset. For each considered category, the table provides the mean data residual norms, the 25th and 75th percentiles of the data residual norms (in square brackets), and the number of included images (in parentheses).