Deep Orthogonal Multi-Frequency Fusion for Tomogram-Free Diagnosis in Diffuse Optical Imaging

Hanene Ben Yedder, Ben Cardoen, Majid Shokoufi, Farid Golnaraghi, and Ghassan Hamarneh

***Abstract*— Identifying breast cancer lesions with a portable diffuse optical tomography (DOT) device can improve early detection while avoiding otherwise unnecessarily invasive, ionizing, and more expensive modalities such as CT, as well as enabling pre-screening efficiency. Critical to this capability is not just the identification of lesions but rather the complex problem of discriminating between malignant and benign lesions. To accurately capture the highly heterogeneous tissue of a cancer lesion embedded in healthy breast tissue with non-invasive DOT, multiple frequencies can be combined to optimize signal penetration and reduce sensitivity to noise. However, these frequency responses can overlap, capture common information, and correlate, potentially confounding reconstruction and downstream end tasks. We show that an orthogonal fusion loss of multi-frequency DOT can improve reconstruction. More importantly, the orthogonal fusion leads to more accurate end-to-end identification of malignant versus benign lesions, illustrating its regularization properties in the multi-frequency input space. While the deployment of portable DOT probes requires a severely constrained computational budget, we show that our raw-to-task model, for direct prediction of the end task from signal, significantly reduces computational complexity without sacrificing accuracy, enabling a high real-time throughput, desired in medical settings. Furthermore, our results indicate that image reconstruction is not necessary for unbiased classification of lesions with a balanced accuracy of** 77% **and** 66% **on the synthetic dataset and clinical dataset, respectively, using the raw-to-task model. Code is available at** `https://github.com/sfu-mial/FuseNet`.

***Index Terms*— Diffuse optical tomography, image reconstruction, deep learning, multi-frequency, tissue estimation, lesion classification, diagnosis, multitask learning, transfer learning, handheld probe.**

I. INTRODUCTION

Breast cancer is the most frequently diagnosed cancer among women [1]. Pre-screening is usually carried out using self-breast examinations, which can suffer from high false-positive rates, or clinical breast examinations [2]. Although breast lumps are often benign, such as lipoma, cyst, or hamartoma, lesion malignancies may appear with a non-palpable sign; hence, regular screenings are critical [3]. While mammography is the most commonly used screening tool today, it has potential cumulative health risks due to its reliance on ionizing radiation and low sensitivity in patients with thick breast tissue [4]. Furthermore, the acquisition device's complexity and size limit patient screening throughput [5].

Imaging modalities based on near-infrared light are emerging as tools for biomedical diagnosis, given the non-ionizing nature of infrared light as well as their ability to penetrate a few centimeters into human structures, such as the skull, brain, and breast [6]. The recent progress of optical sensors makes optical-based modalities increasingly attractive. Diffusion optical tomography (DOT) uses near-infrared light to image soft tissues, offering several advantages in terms of safety, costs, portability, and sensitivity to functional changes [7]. This technique has shown great potential in investigating functional brain imaging [8], [9] and breast cancer screening [10], [11]. Fig. 1-A shows a typical breast screening workflow in the medical setting.

DOT measures the distribution of tissue optical properties as a function of absorption and scattering coefficients. These properties are closely correlated to physiological markers and allow indirect quantitative assessment of tissue malignancy [7], [12]. Indeed, marked variations between healthy and tumor tissue are observed in terms of optical properties and chromophore components (e.g., oxy/deoxy hemoglobin and collagen) [13]. In particular, normal breast tissue and lesions can be separated in terms of optical coefficients at several wavelengths [12], [14].

These properties make DOT a potential promising tool in pre-screening of patients in a clinical setting, saving them from unnecessary exposure to more precise but potentially harmful ionizing modalities such as CT. In such a setting, there is a clear need for both low latency, i.e., method inference

---

Correspondence to be addressed to {hbenyedd, hamarneh}@sfu.ca

This study was conducted through the Fraser Health-ethical approval reference number FHREB 2014-065 – PI Dr. Farid Golnaraghi and supported in part by grants from Natural Sciences and Engineering Research Council of Canada (NSERC) and Simon Fraser University graduate fellowships and graduate dean's entrance scholarship. The authors are grateful for the reviewers' feedback, which greatly improved the manuscript. This research was enabled in part by support provided by Westgrid (www.westgrid.ca). The authors thank Weina Jin, from the Medical Image Analysis Lab, for her suggestions during paper preparation.

Hanene Ben Yedder, Ben Cardoen, and Ghassan Hamarneh are with the Medical Image Analysis Lab, School of Computing Science, Simon Fraser University, BC Canada V5A 1S6. e-mail: {hbenyedd, bcardoen, hamarneh}@sfu.ca

Farid Golnaraghi is with School of Mechatronic Systems Engineering, Simon Fraser University, BC Canada V5A 1S6. Majid Shokoufi is with School of Mechatronic Systems Engineering, Simon Fraser University, BC Canada V5A 1S6. e-mail: {mfgolnar, mshokouf}@sfu.ca

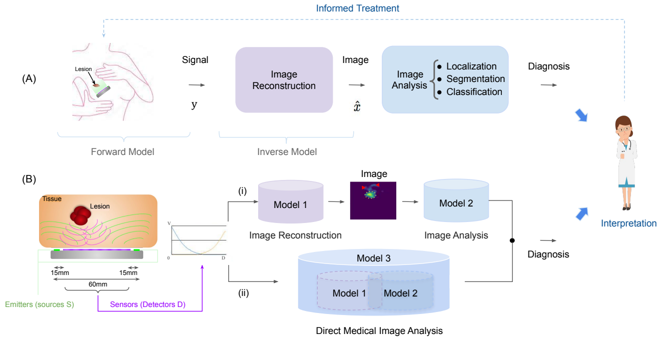

Fig. 1. Typical breast cancer screening workflow. (A) Images reconstructed by an inverse model, from signals collected by the acquisition hardware, are analyzed for assessment, diagnosis, and treatment prognosis. (B) Screening pipelines can be divided into two paradigms: (i) Accurate reconstruction followed by image based classification. (ii) A direct prediction model omits image reconstruction to focus solely on the ultimate task and can help overcome sub-task errors, e.g., reconstruction induced false positives, marked by red triangles in this scenario, in paradigm (i).

TABLE I
REGULARIZATION APPROACHES IN STATE-OF-THE-ART DL-RECONSTRUCTION METHODS. D: DESIGN APPROACH; FF: FEED-FORWARD, I: ITERATIVE UNROLLED BASED MODEL; M: MULTI-MODAL/FREQUENCY; S/P: IN SILICO/PHANTOM DATA; AND C: CLINICAL PATIENT DATA.

	D	M	S/P	C	Approach to mitigate ill-posedness
[29]–[31]	FF	×	✓	×	CNN learns the nonlinear end-to-end mapping
[25]	FF	×	✓	×	Promote appearance similarity
[27], [32]	I	×	✓	×	Augment Gauss-Newton algorithm with deep learning
[33]	FF	✓	✓	×	Micro-CT structural prior
[26]	FF	×	✓	×	Model based on Lippmann-Schwinger equation
[34]	FF	×	✓	×	Reflection model as sum of features from different depths
[35]	I	×	✓	×	Data-driven unrolled network promoting appearance similarity
[36]–[38]	FF	✓	✓	✓	Multi-modal representation learning (US+DOT)
[16]	FF	×	✓	✓	Deep spatial-wise attention network
Ours	FF	✓	✓	✓	Orthogonal multi-frequency representation learning

speed, preferably real-time, and accurate reconstruction and classification.

A DOT scanner is comprised of an array of emitters and receivers, using low-powered LEDs or lasers, to measure the optical transmission [15] or reflection [10] of light beamed into the tissue at various locations on the tissue surface. While an optimized probe design enables a reduced hardware complexity and better portability, it increases the complexity of the reconstruction task, especially when the number of sources is limited [16]. DOT can be classified into three modes: continuous wave, frequency domain, and time domain. Given the tradeoff between imaging performance and cost, frequency domain methods tend to be the most cost-effective mode where tissue optical properties can be directly inferred from the back-scattered signal's amplitude and phase [17]. Furthermore, sampling at different frequencies in a sufficiently broad bandwidth enables converting frequency-domain signals to the time-domain using the inverse Fourier transformation [18]. In this work, we focus on the frequency-domain DOT.

A. DOT Reconstruction Algorithms

Given that a photon can experience many alterations of its path in random directions until it is absorbed, DOT image reconstruction is an ill-posed inverse problem, subject to artifacts [19]. Reconstruction quality and depth sensitivity are inversely proportional to the distance between source and detector and noise level, and strongly depend on the reconstruction method [20]. In addition, the highly heterogeneous nature of malignant cancerous tissue further complicates the reconstruction task [21]. A portable design with limited power budget, significantly reducing the number of available sensors-detectors pairs and the available computational envelope for reconstruction, complicate matters further, by increasing the ill-posedness of the reconstruction problem.

While difference imaging approaches rely on a reference measurement to recover the change in the tissue's optical properties, e.g., a reference tissue, a phantom, or the previous condition like a rest stage in brain DOT, absolute imaging approaches use a single set of measurements to reconstruct optical coefficients. In this manuscript, we focus on absolute imaging.

Traditional image reconstruction techniques commonly rely on non-linear methods minimizing an objective function, iteratively until convergence, e.g., gradient and Newton-type methods [22]. Based on an initial homogeneous tissue optical properties estimate, the difference between the measured signal and the modelled data is used to iteratively update the estimate until achieving convergence within acceptable limits with the measured data. Regularization terms are leveraged to ensure convergence by restricting the space of all possible solutions into only a subset of physically plausible ones. A comprehensive review is presented in [20].

Even though non-linear methods follow directly from the underlying mathematical problem formulation, in practice they have a high computational cost as each iteration needs to be optimized independently at reconstruction time, prohibiting real-time reconstruction. Furthermore, the reconstruction accuracy is easily compromised as the number of sources and detectors is reduced, and reconstruction of complex shapes can become challenging [19]. To address these shortcomings, researchers have explored deep learning (DL) as an alternative approach [23], [24]. A deep learning model for DOT reconstruction is typically trained in a supervised setting on in silico or phantom training data pairs. By incorporating complex and diversified data samples, the model can selectively enrich its feature space to improve performance on real-world data. Recent studies, e.g., [16], [25]–[27], have shown image reconstruction and classification are faster and more accurate when deep learning algorithms are used instead of conventional reconstruction methods. One advantage deep learning based algorithms have over classical reconstruction methods is that they can exploit implicitly learned feature encodings from the DOT sensor data, whereas classical reconstruction algorithms can exploit only priors encoded by human designers [23]. Recent advances tackle the problem of ill-posedness in a variety of ways. We summarize the closely related approaches in a tabulated overview (Table I), and refer the interested reader to [24], [28] for a more in-depth review.

B. Multi-frequency DOT

Frequency-domain systems use intensity modulated sources, ranging from a few MHz to 1 GHz, to illuminate the tissue and collect the amplitude and phase of diffusing waves. A multi-spectral image can be obtained using several LEDs or lasers of multiple wavelengths as illumination. The different LEDs are used consecutively to capture an image per wavelength or combined as one multi-spectral image [39].

The primary motivation for multi-frequency DOT is to exploit the different but complementary responses of chromophore, tissue components, to multi-frequency excitation, given that chromophores absorb photons at different rates at different modulation frequencies [40]. This wavelength sensitivity is leveraged to analyze optical spectra and reconstruct images of the exposed tissue for diagnostic purposes, given that recovered chromophore concentration changes can convey information about functional brain vascular events and the characterization and monitoring of breast lesions [41]. The captured multi-frequency data can provide more spatial and contextual information, enabling more robust and accurate identification and discrimination of disease-correlated biological anomalies [12].

While higher frequencies allow for a better separation of optical properties, such as absorption and scattering coefficients, as well as a better detection of small and shallow objects, the limit of the signal-to-noise ratio (SNR) decreases with increased modulation frequency [42], [43], and penetration decreases as frequency increases. Utilizing multi-frequency data for improving DOT image reconstruction and diagnosis has been an active field of research, illustrating that the accuracy of the optical coefficient can be improved using measurements with multiple modulation frequencies [39], [44]–[46]. Recent studies, summarized in a tabulated overview (Table II) have shown that using measurements with multiple modulation frequencies can improve the recovery of optical coefficients and provide higher SNR and lower error [39], [47]. Improvement, however, depends on frequencies selection scheme and utilized instrument given the specific noise impact [46]. This finding is supported by Zimek et al. [48], who reported that adding dimensions can harm discriminative potential if those dimensions do not improve the signal-to-noise ratio.

Augmenting DOT with ultrasound is finding recent adoption as well, an example of multi-modal fusion [49], [50], [51]. The aforementioned art is based on conventional reconstruction algorithms. To the best of our knowledge, no deep learning-based method has explored the merit of exploiting multiple frequencies in DOT-reconstruction and diagnosis.

TABLE II

Multi-Frequency DOT for Image Reconstruction and Diagnosis. Di: dimension; Freq: frequency range; D: dataset, S: simulation, P: phantom, C: clinical; and MF:Multi-Frequency.

	Leveraging different modulation frequency schemes	DI	Freq (MHz)	D
[47]	Improve joint optical coefficients recons.	2D	100-250	S
[39]	Enhance fused MF image quality	2D	100-1000	S
[52]	Compensate physiological and noise interference in recons.	3D	361-382	P
[53]	Frequency shifting for reduced recons. ill-posedness	2D	$100+5*i, \forall i \in \{0, .., 100\}$	S
[54]	Minimize the effect of phase data and improve contrast	2D	$\{78, 141, 203\}$	P
[46]	Evalutate the impact of modulation frequency selection	2D	50-500	S
[14]	Discrimination between malignant and benign tissue	2D	283-472	C

C. Multi-frequency as Data Fusion

Data fusion models mimic higher cognitive abstraction in the human brain by synthesizing information from multiple sources for improved decision-making. While data fusion is non-trivial, the resulting contribution of multiple data sources or multimodal data can significantly improve the performance of deep learning models [55], [56]. The underlying motivation for collecting multi-modal data is to learn the optimal joint representation from rich and complementary features of the same object or scene. In the context of combining multiple information sources to learn more powerful representations, the terms 'early' and 'late' fusion are commonly used [57]. Early fusion refers to concatenating input data from multiple sources in separate channels before presenting it as input to the network, while late fusion involves processing each input data individually and aggregating their output. Mid-fusion restricts cross-data flow to later layers of the network, allowing early layers to specialize in learning and extracting data-specific patterns [58].

Attention mechanisms have been shown to be suitable for the fusion of features that usually suffer from confounding issues such as conflicting or cancelling information, correlation, and noise. Attention provides an approach to learn to select informative subsets of the data, as well as the relationship between data streams, before fusing them into a single comprehensive representation [56], [59]. Transformer based models, based on a multi-head attention architecture, have recently gained increased adoption [58], [60]. However, the high computational cost and complexity, scaling adversely with input sequence length, remain a significant challenge, especially given the real-time requirement.

Self-supervised learning (SSL) based on a joint embedding architecture, driven by the maximization of the information content of the network branches' embedding, opened the door to the application of joint-embedding SSL to multi-modal signals [61]. The idea is to produce independent embedding variables, removing confounding effects such as partial correlation and avoiding modal collapse between data streams by encouraging architecture diversity between branches, using loss based normalized cross-correlation matrix [62] or explicit variance-preservation term for each embedding [61].

Imposing orthogonal constraints in linear and convolutional neural network layers can act as a form of regularization that can help improve task performance and be beneficial for the network's generalization [63], [64]. Orthogonality in feature space was proposed to encourage intra-class compactness and inter-class separation of the deep features, and has shown improvement in classification tasks [65]. Multi-modal orthogonalization has been used to force uni-modal embeddings to provide independent and complementary information to the fused prediction [66]. Another advantage is that an orthogonal encoding can enforce the learning of a more sparse correlation-free representation. The resulting smaller encoding can reduce architecture dimensions, and serve as an implicit regularization.

D. Towards Direct Medical Image Analysis in DOT

Traditional computational pipelines in biomedical imaging involve solving tasks sequentially (Fig. 1-B.i, e.g., segmentation followed by classification or detection). Although each of these two tasks is usually solved separately, the useful clinical information extracted by the second task is highly dependent on the first task's results. While a 'joint' or multi-stage model where different tasks are lumped together, for example, image reconstruction then classifying diagnosis, can benefit from feature sharing and joint parameters tuning for both tasks, significant computational resources are required to optimize sub-tasks that may not necessarily lead to end-task improvements. In contrast, in the direct medical image analysis [67] (DMIA) paradigm, end task results are directly inferred from raw/original data (e.g., raw sensors or whole image/volume) as illustrates Fig. 1-B.ii. Therefore, the model can focus solely on the end task, reclaiming some of the computational resources for improved results while requiring fewer resources. For instance, Wu et al. [68] trained a neural network for joint reconstruction and lung nodule detection from raw acquisitions and showed performance improvement compared to a two-stage approach. Hussain et al. [69] had shown that a segmentation-free kidney volume estimation can help overcome segmentation errors and limitations and reduce the false-positive area estimates. In a similar perspective, Taghanaki et al. [70] investigated a segmentation-free tumor's volume and activity estimation in PET images. Recently, Abhishek et al. [71] illustrated that, in the context of cancerous skin lesions, predicting the management decisions directly can be a simpler problem to address than predicting the diagnosis followed by management decisions, as one action can be prescribed to multiple subsets of disease classes.

E. Contributions

We make the following contributions in this paper:

(i) We investigate the benefit of multi-frequency data on the quality of DOT reconstruction and breast lesion diagnosis. Previously, many works have addressed the multi-frequency reconstruction problem or diagnosis, albeit using conventional methods. Despite the importance of multi-frequency acquisition for chromophore reconstruction, no deep learning framework has investigated multi-frequency fusion nor joint reconstruction and diagnosis to date. Here, we present a novel approach designed to recover the optical properties of breast tissue from multi-frequency data with a deep orthogonal fusion model followed by a diagnosis.

(ii) To the best of our knowledge, this is the first deep learning-based method that investigates the merits of tackling the diagnosis prediction task from raw sensor data directly without image reconstruction in DOT (direct prediction [1]). Results with and without reconstruction are contrasted using a modular pipeline, highlighting the potential of the raw-to-task model for improved accuracy, while reducing computational complexity.

(iii) We extend a fusion network [59] by training models using an orthogonalization loss function [65] to maximize the independent contribution of each modulation frequency data and emphasize their collective strength, with improved predictive performance compared to a single frequency model.

Section II, introduces our proposed model for multi-frequency DOT fusion and defines the two prediction pipelines (raw-to-task and joint reconstruction and diagnosis). Physics-based computational simulation and real patient datasets are detailed in Section III-A.1. In silico performance results are presented in Section III-B and results on real-world data in Section III-C. We conclude the paper by discussing insights and limitations on interpretability, speed, and adaptive dynamic treatment in Section IV.

II. METHODOLOGY

Solving the inverse problem in DOT recovers the spatial distribution of a tissue's optical properties $x \in \mathbb{R}^{W \times H}$ based on the measured boundary data $y^i \in \mathbb{R}^{S \times D \times N}$, from S sources (emitters) with D sensors (detectors) at different modulation frequencies $i \in \{1, N\}$. The learned inverse function $\mathcal{F}^{-1}(\cdot)$ maps the raw measurements y to an image estimate $\hat{x}$ while remaining faithful to the underlying physics constraints. Learning the inverse function $\mathcal{F}^{-1}(\cdot)$ is carried out by solving:

$$\theta^* = \underset{\theta}{\operatorname{argmin}}\, \mathcal{L}\left(\mathcal{F}^{-1}(y^i;\theta);x\right) + \lambda\mathcal{R}(\mathcal{F}^{-1}(y^i;\theta))\,, \qquad (1)$$

where $\mathcal{L}$ and $\mathcal{R}$ are the network loss function and the regularization, θ are the optimized network weights that parameterize $\mathcal{F}^{-1}$. The reconstruction of an image based on the fusion of all raw signals from diverse modulation frequencies is considered as well by using the fusion network described in Section II-A. While reconstructing an accurate 2D/3D image/volume from collected measurements has been the mainstream task in DOT, in a clinical setting, the ultimate purpose is not necessarily obtaining the image itself but rather making an informed clinical diagnosis or management decision, such as lesion detection and classification into predefined classes. To compare the impact of omitting the reconstruction and directly predicting the end task, we implemented two architectures: The first, FuseNet, reflects classical approaches, i.e., a classification module is appended to the output of the reconstruction layer to make a prediction, where the result of the multi-spectral reconstruction is used to supervise the classification task (Section II-B). Whereas the second, Raw-to-Task, uses the same classification module to make a prediction based on the fused raw data directly, i.e., no reconstruction is considered in between. The ultimate goal is to study the ability of deep learning to provide superior prediction based on the raw signal only while reducing model complexity and computational cost (Section II-C).

A. Fusion Network

Given multi-frequency raw data paired with known diagnosis outcomes, the objective is to learn a robust multi-frequency

---

[1] While our direct prediction contribution omits the 'tomogram' part of DOT, and thus works directly on near-infrared (NIR) sensor data, our fusion contribution applies both to tomogram reconstruction as well as tomogram-free reconstruction. Thus, we continue to use DOT throught the paper instead of NIR.

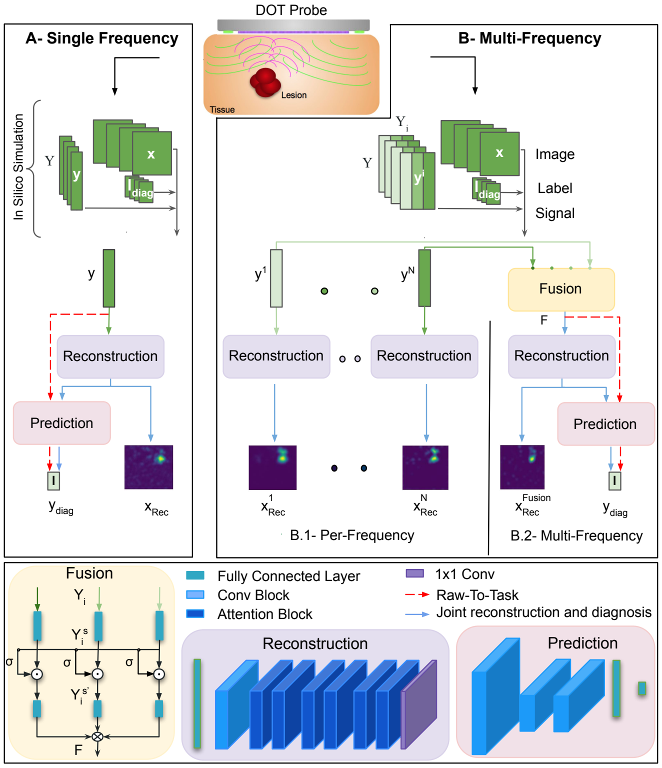

Fig. 2. Architecture overview of the proposed DOT image reconstruction and diagnosis method. (A) single-frequency and (B) multi-frequency signals ($\boldsymbol{y}$) along with corresponding ground truth diagnosis labels ($\boldsymbol{l_{diag}}$) and images ($\boldsymbol{x}$) are used to train the model. In the single-frequency variant of our method (A), $\boldsymbol{y}$, is used as input to the image reconstruction, then the resulting image is used for diagnosis prediction. In multi-frequency, note the two variants: (B.1) per-frequency reconstruction and (B.2) multi-spectral reconstruction and diagnosis. For both single and multi-frequency, the red dashed lines depict the raw-to-task flow, where the image reconstruction is skipped and the diagnosis is predicted directly from $\boldsymbol{y}$. The bottom panel shows the details of the multi-frequency fusion, reconstruction, and prediction modules.

representation in a supervised learning setting. While many fusion strategies have been proposed in computer vision, natural language processing, and multimodal biomedical data, strategies for fusing data in multi-frequency DOT data remain unexplored in deep learning-based approaches. Inspired by recent methods for multimodal data fusion [59], [66], we adopt a similar attention-based mechanism to control the expressiveness of features from each input frequency before constructing the multi-frequency embedding, while uniquely feeding the raw data directly with no further pre-processing. Let $Y \in \mathbb{R}^{S\times D\times N\times M}$ be a training mini-batch including M tissue samples, each collected using N frequencies such that $Y = [Y_1, Y_2, ..., Y_N]$ where for each frequency i, $Y_i = [y^i_1, ...y^i_M]$ includes data for M samples. When $N > 1$, input measurements from each frequency are combined using the fusion branch (Fusion, Fig. 2). To reduce the impact of noisy input features and compress the size of the feature space, each Y_i is first passed through a fully connected layer of length l, with ReLU activation, outputting $Y^s_i \in \mathbb{R}^{l\times 1\times M}$, followed by an attention mechanism that scores the relevance of each feature in Y_i. We define frequencies $\not i$ as the set $\{j\}$ such that $j \in \{1, .., N\} \setminus \{i\}$, i.e., for frequencies other than i. A linear transformation W_A of frequencies $Y_{\not i}$ that would score the relative importance of each feature in i, is learned. W_A is a learned weight matrix parameters for feature gating. The attention weights vector a_i is then applied to Y^s_i, an element-wise product of scores and features, to form the attention-weighted embedding $Y^{s'}_i \in \mathbb{R}^{l\times 1\times M}$:

$$Y^{s'}_i = a_i * Y^s_i = \sigma\left(W_A * [Y_{\not i}]\right) * Y^s_i. \qquad (2)$$

Finally, attention-weighted embeddings are passed through a fully connected layer of length $l2$, with ReLU activation, then combined through a Kronecker product between all frequency embeddings to capture possible interactions. Each vector is appended by 1 to capture partial interactions between frequencies [59]. The final fused embedding is then defined as:

$$F = \begin{bmatrix} 1 \\ Y^{s'}_1 \end{bmatrix} \otimes \begin{bmatrix} 1 \\ Y^{s'}_2 \end{bmatrix} \otimes \ldots \otimes \begin{bmatrix} 1 \\ Y^{s'}_N \end{bmatrix}. \qquad (3)$$

$F \in \mathbb{R}^{l'\times l'\times l'\times M}$, for $N = 3$ and $l' = l2 + 1$, is a N-dimensional hypercube of all frequency interactions.

B. Joint Multi-frequency Reconstruction and Diagnosis

The task is to recover tissue optical properties and diagnosis outcome given raw signal data. While a single frequency model SF-JRD (Fig. 2-A), used as a baseline, relies on a single frequency measurement to reconstruct spatially distributed optical coefficients and predict diagnosis, a multi-frequency model (FuseNet) relies on a joint representation from multiple frequency measurements (Fig. 2-B). The multi-frequency model, including a network with multiple branches as shown in (Fig. 2-B), inputs N measurements of the same scanned tissue at N modulation frequencies. A multi-spectral image that combines all frequency measurements, using the fusion branch encoding (Fig. 2-B.2), is reconstructed and passed to a classification module for diagnosis prediction. Furthermore, a per-frequency image is reconstructed using each modulated frequency signal. As depicted in (Fig. 2-B), the FuseNet model outputs are x^i_{Rec} $\forall i \in \{1, .., N\}$, x^{Fusion}_{Rec}, and y_{diag} which denote the per-frequency reconstructed image (Fig. 2-B.1), the multi-spectral reconstructed image, and the predicted diagnosis label (Fig. 2-B.2), respectively.

Using multiple inputs, per frequency network reconstruction branches (Fig. 2-B) learn independent representations, where features derived from each input measurement (Y_i) are only useful for the corresponding output x^i_{Rec}. Furthermore, given the differences in initialization, the branches can converge to disconnected modes in weight space, thereby behaving as independently trained neural networks. Empirically, we observe that they converge to distinct optima. For this multi-task reconstruction and prediction model, we extend the multi-task framework [16] and train a model to simultaneously reconstruct a per-frequency image, localize the lesion, and predict the diagnosis.

The reconstruction branch (Fig. 2) implements the design detailed in the multi-task framework [16] with a fully connected layer, 128×128, followed by a convolutional layer and 4 residual attention blocks with 32 channels, filters of size of 3 and ReLU activation, to produce the final reconstruction image. While the first and last layers are shallow feature extractors, the attention blocks extract hierarchical attention-aware features with modules of the form: two convolutions followed by squeeze and excite modules. This deep spatial-wise attention network attends to the most important features by reweighting features according to their interdependencies in feature space and filtering noisy ones. In contrast to the difference approach, which uses a reference measurement of healthy tissue to compute the contrasted inverse image, our reconstruction module is designed to learn the mapping directly, from the measured data to the desired output, without the need for any prior knowledge or references that can bias the search space. Furthermore, obtaining such a reference measurement from a homogeneous background in a clinical setting, such as a breast cancer screening, is not trivial; hence, we consider absolute imaging, where the network learns the inverse mapping between sensor measurements and the image domain directly.

The prediction branch (Fig. 2) includes 2 convolutional layers with max pooling and two final classification layers. Raw data from different frequencies are passed to the reconstruction branch except for the multi-spectral subnetwork, where raw data from different frequencies are first fused via the fusion branch. The fused features are passed to the reconstruction branch, which outputs a multi-spectral image followed by a classification layer to output the final classification prediction. The multi-task loss (L_{MULTI}) encompasses all three tasks: reconstruction, lesion localization, and diagnosis as a sum of losses for each task is defined as follows:

$$L_{MULTI} = L_{REC} + L_{DIAG} \tag{4}$$

where L_{REC} and L_{DIAG} denote the reconstruction loss and the diagnosis losses, respectively.

1) Reconstruction loss: We adopt the reconstruction loss defined by Ben Yedder et al. [16]. The mean square error loss L_{MSE} combined with the location loss L_{LOC} guide the image reconstruction and lesion localization of the network as per (5). L_{MSE} recovers the pixel-wise representation of the image.

$$\begin{aligned} L_{\text{REC}} &= L_{\text{MSE}} + \beta\, L_{\text{LOC}}, \\ L_{\text{LOC}} &= ||DT(\mathcal{F}^{-1}(y_i, \theta), x) - DT(x)||, \end{aligned} \tag{5}$$

where DT denotes the distance transform and computes the Euclidean distance between the image pixel location and the lesion boundaries, θ denotes the parameters of the multi-task model, and $\beta \in [0, 1]$ is a hyper-parameter controlling the contribution of L_{LOC}.

2) Diagnosis loss: The diagnosis loss, L_{DIAG}, is a weighted sum of the categorical cross entropy loss L_{CE}, and the orthogonal projection loss L_{OPL}:

$$L_{\text{DIAG}} = L_{CE} + \gamma L_{OPL}, \tag{6}$$

where:
$$\begin{aligned} L_{CE} &= L_{CE}\left(x, l_{\text{diag}} \mid \Theta\right) \\ &= -\sum_{j=1}^{n_{\text{diag}}} l_{\text{diag},j} \cdot \log\left(\phi\left(x \mid \Theta\right)_j\right), \\ L_{OPL} &= (1 - s) + |d| \\ s &= \sum_{\substack{i,j \in B \\ y_i = y_j}} \left\langle \mathbf{f}_i, \mathbf{f}_j \right\rangle, d = \sum_{\substack{i,k \in B \\ y_i \neq y_k}} \left\langle \mathbf{f}_i, \mathbf{f}_k \right\rangle, \end{aligned} \tag{7}$$

n_{diag}, l_{diag} denote the number of classes in the diagnosis prediction tasks and ground truth label, respectively. $\phi(x|\Theta)_j$ denotes the predicted probability for the j^{th} class by the model parameterized by Θ. $\gamma \in [0, 1]$ is a hyper-parameter balancing the contribution of the L_{OPL}. $|x|$ is the absolute value operator, $< x, y >$ the cosine similarity operator applied on two vectors, and B denotes the mini-batch size.

The orthogonal projection loss L_{OPL}, as defined in [65], is used to maximize separability between classes by enforcing class-wise orthogonality in the intermediate feature space and simultaneously ensuring inter-class orthogonality (d term) and intra-class clustering ((1-s) term) within a mini-batch.

TABLE III

SUMMARY OF VARIANTS OF OUR METHOD ARCHITECTURES INPUT AND OUTPUT DETAILS'. N: NUMBER OF MODULATION FREQUENCIES; S: SOURCES; D: DETECTORS; H: HEIGHT; W: WIDTH.

	Input	Output	
		Direct prediction	Joint reconstruction and diagnosis
Single-Freq	$Y_1 \in \mathbb{R}^{S \times D}$	$y_{diag} \in \mathbb{R}$	$x_{Rec} \in \mathbb{R}^{W \times H}$
			$y_{diag} \in \mathbb{R}$
		SF-DP	***SF-JRD***
Multi-Freq	$Y \in \mathbb{R}^{S \times D \times N}$	$y_{diag} \in \mathbb{R}$	$x^i_{Rec} \in \mathbb{R}^{W \times H} \quad \forall i \in \{1, .., N\}$
			$x^{Fusion}_{Rec} \in \mathbb{R}^{W \times H}$
			$y_{diag} \in \mathbb{R}$
		Raw-to-Task	***FuseNet***

C. Direct Prediction: Raw to Task Model

The ultimate aim of DOT-based screening is the early identification and classification of breast cancer lesions. Therefore, we investigate if focusing exclusively on the end task, at the cost of omitting the reconstruction of a 2D image, can perform better or worse compared to classification with the intermediate reconstruction. Without the need to reconstruct a 2D image, the architecture and computational complexity reduce significantly, leading to a reduction in power consumption and data computation latency. The classification module is used to make predictions based on the fused raw data, where combined features, extracted from different frequencies using the fusion branch (Section II-A), are passed to a convolutional layer for the prediction task and a final classification layer with the associated loss (Fig. 2-dashed lines). The diagnosis loss function, L_{DIAG}, is used to train the model given the raw input measurement where:

$$\begin{aligned} L_{\text{CE}} &= L_{\text{CE}}\left(\left(y_i, .., y_N\right), l_{\text{diag}} \mid \Theta\right) \\ &= -\sum_{j=1}^{n_{\text{diag}}} l_{\text{diag},j} \cdot \log\left(\phi\left(y \mid \Theta\right)_j\right), \end{aligned} \tag{8}$$

TABLE IV

OPTICAL COEFFICIENTS DISTRIBUTIONS ON THE IN SILICO DATASET FOR WAVELENGTHS IN 690-850 NM SPECTRUM [72]

		Healthy tissue	Benign	Malignant
Absorption $\nu_a(cm^{-1})$	690	0.042 ± 0.013	0.08 ± 0.04	0.110 ± 0.066
	750	0.046 ± 0.024		0.100 ± 0.060
	800	0.052 ± 0.015		0.118 ± 0.096
	850	0.032 ± 0.005		0.124 ± 0.089
Scattering $\nu_s(cm^{-1})$	690	12.9 ± 2.3	19.4 ± 8.4	13.5 ± 4.7
	750	8.70 ± 2.2		11.6 ± 3.9
	800	10.5 ± 1.2		12.2 ± 1.7
	850	8.40 ± 0.4		9.10 ± 1.9

y_i denotes the i^{th} measurement of the raw data and $\phi(y^{(i)}|\Theta)_j$ denotes the predicted probability for the j^{th} class given an input $y^{(i)}$ by the model parameterized by Θ. The orthogonal projection loss L_{OPL} (7) is used to maximize separability between classes in the feature space.

FuseNet, Raw-to-Task, SF-JRD and SF-DP models are trained separately while using the same modules: fusion, reconstruction, and prediction modules. Table III summarises different models input and output details.

D. Transfer Learning Network

In medical imaging settings, transfer learning [73] can be used to bridge the gap between simulated and clinical data by transferring knowledge learned from simulated data to improve the performance of models on clinical data [74]. This is particularly important in medical imaging and relatively new imaging devices such as DOB probes, where obtaining large quantities of annotated clinical data can be challenging and expensive. Similar to Ben Yedder et al. [16], we use transfer learning to render an in silico trained network applicable to real world data and reduce the disparities between real-world acquisition y^p and in silico simulated data y^s. A multi-layer perceptron (MLP) network is used to tackle the domain shift by minimizes the transfer learning loss $\mathcal{L}_{TL}$ over N_p sets of real data measurements obtained using a phantom solution and their corresponding tissue-equivalent simulated data:

$$\theta^* = \underset{\theta}{\operatorname{argmin}}\, \mathcal{L}_{\mathrm{TL}}(\theta)$$

where

$$\begin{aligned}\mathcal{L}_{\mathrm{TL}}(\theta) &= \sum_{i=1}^{N_p} ||\phi(y_i^p;\theta) - y_i^s|| \\ &\quad + \alpha \sum_{i=1}^{N_p} \sum_{j=1}^{D-w+1} ||\phi(y^p_{[j-w\ \ j+w]};\theta) - y^s_{[j-w\ \ j+w]}||\end{aligned} \tag{9}$$

w is the size of the sliding window, D is the number of detectors, α is a hyper-parameter that is used to control the contribution of the windowed mean absolute error loss. At inference time, the final reconstructed and diagnosis results are computed as:

$$\hat{x}^* = \mathcal{F}^{-1}(\phi(y^p)). \tag{10}$$

III. RESULTS

We present results on both in-silico and clinical data. Results were obtained by training the model on the in-silico data. A transfer learning network, adapted from [16] and trained on a phantom dataset, bridges the distributions shift that is unavoidable when switching between in silico and real world data. A Gaussian noise was added to the signal, mimicking real world signal fluctuation, to improve model robustness to sensor noise and mimic the real-world drift of device characteristics on different clinics in between calibrations. This noise model depicts the highly variable noise to each individual detector as caused by sensor noise and interference of refracting light. Consistent with previous work [16], [26], [75], we set σ = 10% of the maximum sensor value. Besides the simulated noise, the probe accounts for ambient light, the predominant source of noise, as well by capturing a frame without any active emitters and then subtracting it from the actual data measurement, taken during clinical tests, prior feeding it into our model. Performance evaluation captures image reconstruction quality, diagnosis accuracy, and speed. The next section provides details.

A. Experimental Design

1) Dataset: We simulate light propagation into tissue at different light wavelengths, 690, 750, 800, and 850 nm, illuminating the tissue sequentially, using the physics-based Toast++ software [76]. Probe geometry, with two LED sources and a row of 128 detectors placed in the same straight line, as illustrated in Fig. 1-B, was configured to reflect real physical DOT probe geometry [10], used clinically, in terms of the number and geometry of sources and detectors and the used frequencies. We collect training samples from synthesized tissues with known optical properties and labels. Lesions are modeled as tissue with perturbed optical coefficients embedded in an otherwise homogeneous diffusive medium. A set of 2D images with various lesion sizes, shapes, and positions discretized into finite element nodes (triangular meshes) is synthesized. In order to mimic real breast tissue optical parameters', we base the optical properties on realistic optical coefficient values [14], [72] as summarized in Table IV. A total of 4000 sample data pairs (256-D × 4 vectors, 2D image, label) are used to train and test our method. Each sample includes the collected measurement vectors, one for each frequency, the ground truth image, and the diagnosis label. Training dataset size is chosen as a compromise between training time and in-silico performance. We focused on the diversity of simulated scenarios while also being mindful of computational resources. While we note that in silico and phantom data can result in very large datasets [77], we focused on the diversity of simulated scenarios and adopted a dataset with various lesions number, size, and depth to emulate realistic conditions where the optical properties of the anomaly and surrounding tissues were taken from available in vivo breast tissue experimental data.

Our recently developed hand-held breast scanner (DOB-probe) [10], [78] was used to collect real patient data to test our method. The probe includes two source LEDs, with

TABLE V
SUMMARY OF CLINICAL DATA

	Tumor Position	Tumor size (cm)	Tumor Type
Patient 1	Left Breast	$1 \times 0.8 \times 0.7$ $2.5 \times 0.8 \times 0.8$	BI-RADS 7
	Right Breast	$1.1 \times 0.8 \times 0.7$	Benign
Patient 2	Left Breast	$2.2 \times 1.7 \times 1.7$	BI-RADS 4
Patient 3	Left Breast	$1 \times 1 \times 1$	Non-invasive ductal
Patient 4	Left Breast	$2.5 \times 1.7 \times 3.5$	BI-RADS 5
Patient 5	Right Breast	2.4	BI-RADS 4
Patient 6	Right Breast	$2.3 \times 2.2 \times 1.5$	BI-RADS 4
Patient 7	Left Breast	$1.7 \times 1.4 \times 1.2$	BI-RADS 5
Patient 8	Left Breast	$1.6 \times 0.8 \times 0.8$	BI-RADS 5
Patient 9	Right Breast	$2.2 \times 2.1 \times 2.3$	Invasive ductal

wavelengths of 690, 750, 800, and 850 nm illuminating the tissue consecutively and a row of 128 co-linear detectors. Note that the frequencies share variable overlap in the spectrum [14], motivating further the need for orthogonal encoding. To train the transfer learning module breast-mimetic phantoms, with known inhomogeneity locations, and DOB-probe were used to collect measurements [16].

Following the ethics and institutional review board approval protocol, clinical data were collected from 9 participants diagnosed with breast tumors [79]. In a normal clinical pre-screening exam, a breast is usually divided into four quadrants, and different measurements are collected on each quadrant. Given that the used probe is in clinical trials [10], [16], [39], patients with known cancer localization are considered, and sweeps over the lesion location and the opposite healthy breast are collected. This step was essential to proving that the technology we introduced works well with human tissue. For each patient, height, weight, age, and gender, as well as details of the subjects' breast cancer, briefly summarized in Table V, were recorded. Patients were placed in a supine position, and scans at multiple points over the lesion location and healthy breast were collected. On average, four different measurements (scans) were taken on each breast. Even though no reconstruction ground truth is available for real-world data, it is invaluable to detect robustness and real-world performance, with partial ground truth known from other modalities on the same patients. The precise location, size, and type of the tumor lesion were determined via mammography, ultrasound, or biopsy. Note that these details were only used for model performance evaluation and metrics calculation, while raw sensor signal only was inputted to the model. Another advantage of our direct prediction approach is that the absence of pixel-wise ground truth is less problematic compared to reconstruction based classification, as only the diagnosis label is required.

2) Implementation: Models were implemented in the Keras TensorFlow framework and trained for 100 epochs on an NVIDIA Titan X GPU. By optimizing the model's performance on the validation set, we set all hyper-parameters as follows: batch size to 16, learning rate to 10^{-4}, optimizer set to Adam, and initialization to Xavier. Early stopping was used if the validation loss had not improved within 10 epochs. The in silico data was divided in a 80/10/10% training/validation/test split, and hyper-parameters β (5), α and D (9), and γ (6) were set to 0.2, 0.5, 4 and 0.5, respectively. The fully connected units for the fusion branch were set to 32 and 16 for l and l2, respectively.

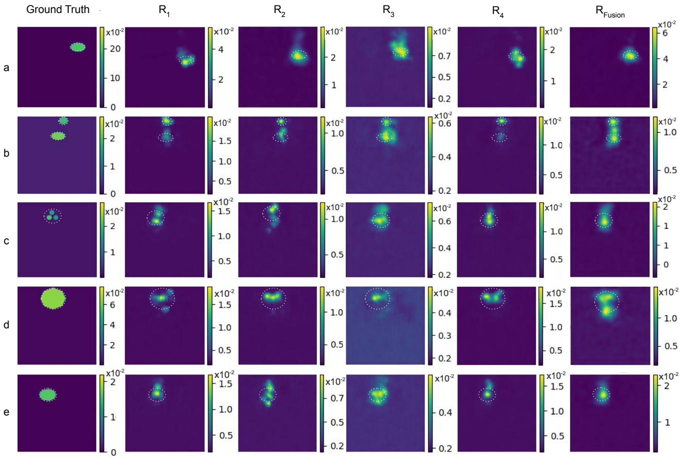

Fig. 3. Qualitative reconstruction performance of absorption coefficients using the FuseNet++ on in silico samples with varying ground truth lesion sizes, locations, and numbers. Our multi-spectral results ($\boldsymbol{R}_{\text{Fusion}}$) show an overall superiority in terms of generally improved background/foreground contrast and a better differentiation between lesion sizes and lesion localization compared to per-frequency reconstruction results ($\boldsymbol{R_1}$ to $\boldsymbol{R_4}$) at wavelengths 690, 750, 800, and 850 nm, respectively.

3) Evaluation metrics: To quantify the models' robustness, we look at (i) lesion localization error (Loc. Error); (ii) peak signal-to-noise ratio (PSNR); (iii) structural similarity index (SSIM); and (iv) Fuzzy Jaccard for reconstruction quantification, while the balanced accuracy (BA), F1 score (F1), precision P, recall R, Matthews correlation coefficient (MCC), and confusion matrix are reported for the classification task quantification.

$$\begin{aligned} BA &= \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right) \\ P &= \frac{TP}{(TP+FP)}, \\ R &= \frac{TP}{(TP+FN)}, \\ F1 &= 2\frac{P*R}{P+R} \\ MCC &= \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \end{aligned} \tag{11}$$

True positive (TP) is the number of correctly predicted samples as positive, while false positive (FP) is the number of wrongly predicted samples as positive. False negative (FN) is the number of wrongly predicted samples as negative, while true negative (TN) is the number of correctly predicted negative class samples over the number of classes in the prediction tasks. Recall quantifies the number of positive class samples properly identified by the model, while precision measures the number of correct positive predictions made by the model. BA, used when quantifying performance on imbalanced data, measures the average accuracy obtained from all classes. MCC measures the quality of multi-class classifications and is informative in cases of skewed class distributions.

For the computational cost at inference, we quantify the forward pass of the model, measured in ms per example.

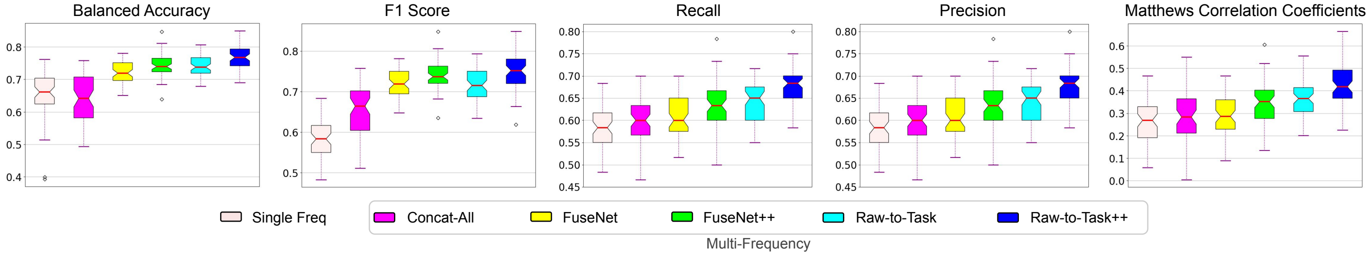

Fig. 4. Quantitative diagnosis performance of different models when one vs multi-frequency are used. Overall results show improved prediction performances in multi-frequency models. Note the significant improvement when FuseNet is used compared to a simple concatenation (Concat-All). Results using the FuseNet++ enforce the benefit of feature space orthogonality. Raw-to-task++, in which all network capacity is dedicated to the end task, shows an overall performance gain.

TABLE VI

Quantitative results on in silico test dataset.Losses are defined in Section 2-B; Loc.error: lesion localization error; PSNR: peak signal-to-noise ratio; SSIM:structural similarity index; BA: balanced accuracy; F1: F1 score. †: value not supported by method, ‡: image reconstruction skipped.

	Loss			Loc. Error (pixel, ↑)	PSNR (dB, ↑)	SSIM (↑)	Fuzzy Jaccard (↑)	Runtime (ms, ↓)	BA ↑	F1 ↑
	L_{REC}	L_{CE}	L_{OPL}							
Single-Freq	✓	✓	†	17.7 ± 21.9	19.1 ± 4.8	0.80 ± 0.05	0.60 ± 0.17	23	0.65	0.65
Concat-All	✓	✓	†	20.4 ± 18.4	19.6 ± 6.2	0.73 ± 0.17	0.61 ± 0.18	28	0.63	0.65
FuseNet	✓	✓	-	17.6 ± 23.3	20.2 ± 4.1	0.88 ± 0.05	0.62 ± 0.19	31	0.72	0.72
FuseNet++	✓	✓	✓	**15.7 ± 12.7**	**21.2 ± 4.4**	**0.89 ± 0.03**	**0.64 ± 0.18**	32	0.74	0.74
Raw-to-Task	†	✓	-	‡				**15**	0.74	0.72
Raw-to-Task++	†	✓	✓	‡				**15**	**0.77**	**0.75**

To evaluate the performance of our models, we contrast the results when using one frequency with many frequencies in the FuseNet and the Raw-to-task model. We present results on in-silico data and clinical data.

B. Results on Synthetic Data

Trained on the in silico data and tested on a separate test set of 240 images, we compare the reconstruction and prediction performance of our FuseNet and the prediction performance with the Raw-to-Task counterpart.

1) Joint reconstruction and diagnosis: Figure 3, illustrates reconstruction results on selected in silico samples with different lesion sizes, numbers, locations, and depths. In order to offer clinicians more details, results based on each frequency separately (R_i) as well as results that use all frequencies are shown, with the latter showing more consistent performance. The joint model successfully exploits the presence of the different frequencies and generally shows an improved background/foreground contrast. For example, the difference in signature for 3 small but proximate lesions is marked in different frequency results (R_1 to R_4) (row c), while a more accurately reconstructed sphere size is provided by the fusion result R_{Fusion} in row (d). Detecting heterogeneity in lesions is critical for correct treatment estimation given that it is a proxy indicator of evolutionary pressure in the lesion, selecting for more resistant cancer sub-populations. Table VI presents the quantitative results of the ablation study, where the contribution of different losses and modular choices of the architecture to model performance are quantified. Rows 1 to 4 highlight the benefit of using multi-frequency fusion on the reconstruction task. A naive multiple frequencies concatenation will not necessarily improve results, which agrees with the findings reported by Applegate et al. [46], illustrating the impact of adding noisy dimensions on performances. Nonetheless, we see improved results for FuseNet. When fusion branch and L_{OPL} are used jointly (FuseNet++), the features contribution from each frequency is maximized in contrast to simple features concatenation (Concat-All) at the price of a minimal computational increase (only 9%).

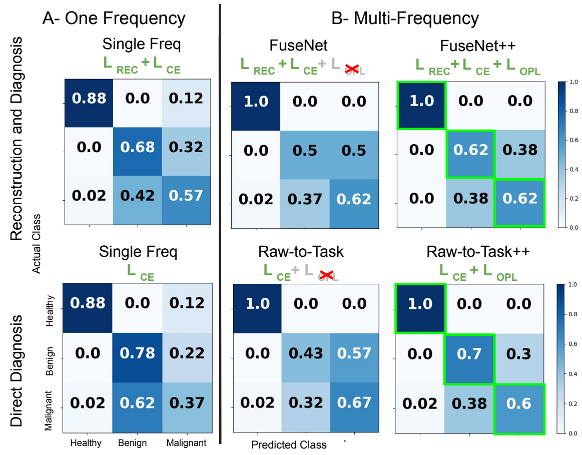

Fig. 5. Diagnosis prediction confusion matrices when (A) one vs (B) multi-frequency inputs are used. Note the improvement in accuracy of unbiased lesion classification (benign, malign) vs healthy when multiple frequencies are used, as illustrated by the higher values along the diagonal. Results of FuseNet++ highlight the benefit of encouraging orthogonality in enhancing benign vs malignant separability while reducing healthy false negative. Raw-to-task++ further improves separability at the expense of minimal false negative (2%).

Prediction performance highlighted in Table VI and Fig. 4 show an overall improvement when more input frequencies are available, with a boost in performance when FuseNet and FuseNet++ are used. Confusion matrices (Fig. 5-A,B) show a clear discrimination between healthy and lesion features when more data, in the form of more frequencies, is available. Further, improved benign and malignant discrimination is

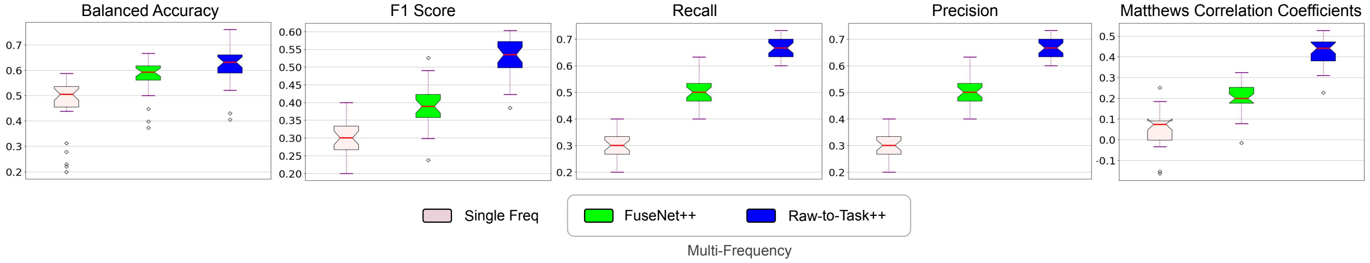

Fig. 6. Quantitative diagnosis performance when one vs multi-frequency are used on clinical dataset. Overall results show improved prediction performances in multi-frequency models compared to single frequency and indicate that image reconstruction is not necessary for unbiased classification and can even lead to biased results. Note the marked improvement when Raw-to-task++ is used compared to FuseNet++.

observed when feature orthogonality is leveraged (Fig. 5-B) as well as a reduction in healthy false negative.

2) Direct prediction: In Figure 5, similarly to the joint model, the direct prediction model results using a single frequency input (SF-DP) (Fig. 5-A) are contrasted with raw-to-task prediction results using multiple frequencies as input (Fig. 5-B). A clear discrimination between features is apparent when more data, in the form of multiple frequencies, is available, especially when discriminating between healthy and lesion; the primary application in DOT-based screening deployments. Raw-to-task model significantly reduces computational complexity (Table VI-Runtime), enabling lower latency and higher throughput in real medical settings. Next, we tested the contribution of individual loss function terms and architecture component on overall diagnosis performance. Figure 4 shows the diagnosis performance on the test set for the best value of γ and highlights the benefits of the feature orthogonality constraint in breast cancer diagnosis, where tumoral and non-tumoral breast lesion differentiation is challenging. Contrasting FuseNet++ and Raw-to-task++ (Fig. 4-5) illustrates performance gain when all network capacity is dedicated to the end task rather than intermediate ones.

C. Results on Clinical Data

Figure 7 presents the reconstruction performance on breast scans from patients diagnosed with breast tumors. The probe is placed close to the likely location of each identified lesion, and a set of scans are made. The opposite healthy breast, for each patient, is scanned as a contrastive reference. Weak labels were attributed to each set of scans regardless of the probe's closeness to the tumor localization. As a partial ground truth, patients underwent mammography and/or Ultrasound scans to obtain estimated lesion dimensions and biopsies to confirm tumor type. While lesions are accurately reconstructed in most cases, as shown in Fig. 7, with clear foreground and background discrimination in R_{Fusion} as well as R_1 to R_4, healthy cases, capturing only background readings, highlight a better robustness of orthogonal fusion, R_{Fusion}, to noise.

Figure 6 reports quantitative prediction performance on single vs. multi-frequency data and highlights the overall improved performance when more frequencies are used. Note the biased classification results when image reconstruction supervises the prediction task, FuseNet++, compared to direct prediction from raw data, Raw-to-task++.

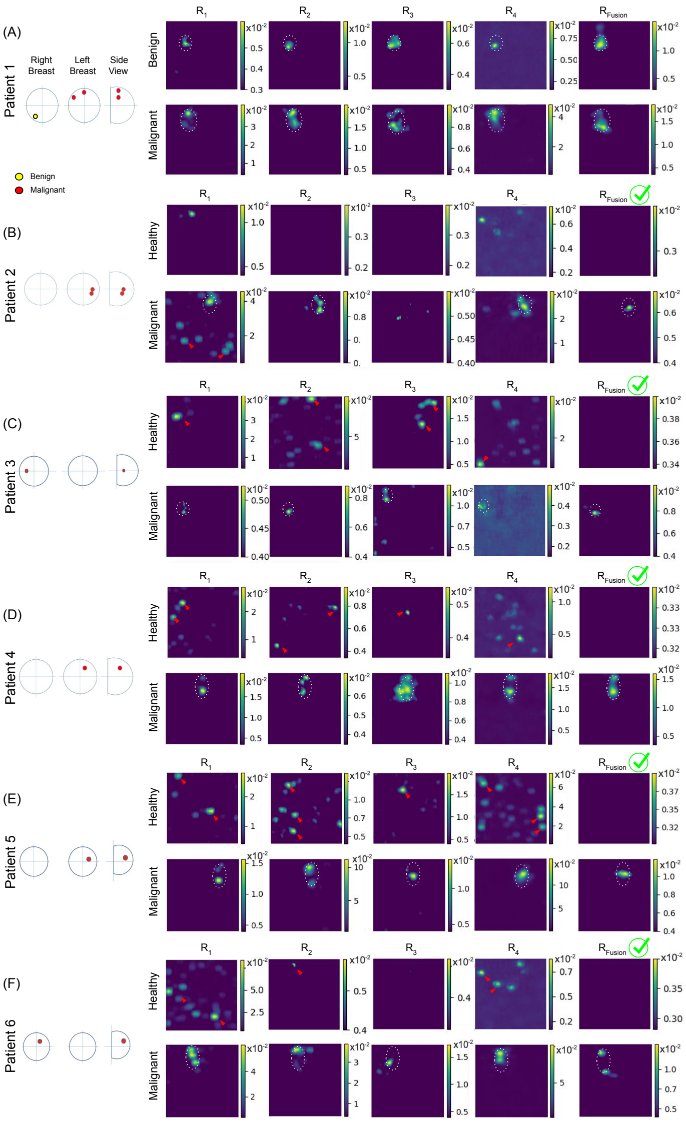

Fig. 7. Qualitative reconstruction results in clinical patients with benign and malignant tumors. Approximate lesion sizes and locations were obtained with joint modalities (details in Table V). Note, in (A), the ability of FuseNet++ to reconstruct lesions, while, in (B-F), the robustness of orthogonal fusion to noise $(\boldsymbol{R}_{\text{Fusion}})$ compared to $(\boldsymbol{R_1}$ to $\boldsymbol{R_4})$ (healthy row) is highlighted.

TABLE VII
QUANTITATIVE RESULTS ON CLINICAL DATASET USING RAW-TO-TASK++

	Precision	Recall	F1-score	Number of scans
Healthy	0.71	0.65	0.68	32
Benign	0.11	0.5	0.18	2
Malignant	0.78	0.68	0.73	44
Weighted-Avg	0.73	0.66	0.69	78

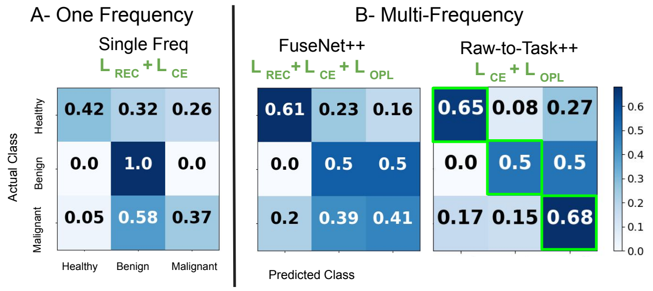

Fig. 8. Clinical data diagnosis prediction confusion matrices when (A) one vs (B) multi-frequency inputs are used. Note the improvement in accuracy of lesion classification in the Raw-to-Task model, despite the imbalance in the data.

The confusion matrix, Fig. 8, shows improved discrimination between healthy and lesion features with the raw-to-task model. If we consider that a key feature of the reconstruction based classification is the interpretable angle of such results, we note that the raw-to-task model has the added advantage, in addition to improved performances, that it omits potentially confounding explanations, where reconstruction artifacts can mislead experts. Indeed, in recent work on explainable artificial intelligence, such confounding explainers were identified as a roadblock [80]. Table VII reports raw-to-task model diagnosis performances on each data class to highlight the clinical data imbalance compared to a balanced training data scheme.

In Figure 9, we illustrate some failure cases at the limit of detection capability, with false positives (Fig. 9-A,B, marked with red triangles) and false negatives (Fig. 9-C). For screening purposes, false negatives are more critical; false positives would eventually be resolved by follow-up diagnosis. Note that the discriminatory power of the detection is limited by tumor depth, shape, and noise level. It may require several scans over breast tissue in order to be captured. The failure cases here are from a single scan measurement only, not aggregated

Although the transfer learning network, trained using phantom data, bridges to some extent the disparity between in silico (training) data and real-world data, its performance on clinical data reveals that it can still be misled by significant real-world variations, such as differences in illumination and noise levels. Additionally, since each tumor is unique, tumor heterogeneity can result in distinct acquisition signatures that may not be present in the training data. These failure cases highlight the need for more clinical data (patient data) to better train the transfer learning module. Current results present a proof of concept, where validation on larger and more diverse datasets is still required.

D. Effect of Lesion Localization on Accuracy

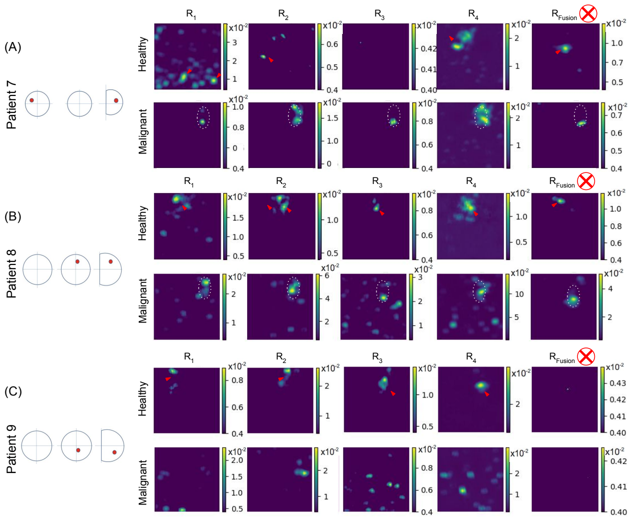

Fig. 9. Examples of reconstruction failure cases. (A,B) highlight false positive reconstruction cases, marked with red triangles, that remain less critical than false negative cases where a tumor is missed (C). Note the noisy reconstruction in $\boldsymbol{R_1}$ to $\boldsymbol{R_4}$, suggesting a quite noisy input signal.

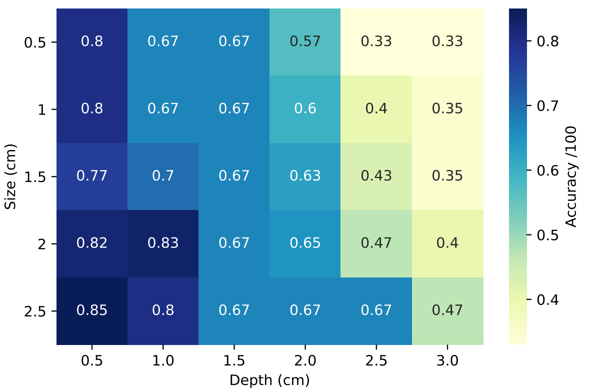

Fig. 10. Effect of lesion depth and radius on model prediction accuracy. Note how the more superficial (closer to the skin surface) and larger lesions are more accurately detected.

We quantify the effect of lesion location on lesion detection accuracy in Figure 10, where we classify whether a lesion is present or not. The penetration depth into breast tissue is approximately half the distance between the source and detectors [18], ∼2.5 cm for our DOT probe. Our results confirm the expected reduction in lesion detection accuracy as the lesions decrease in size or increase in depth.

IV. DISCUSSION

In order to be an effective tool in clinical settings, a clinician's trust is essential. A combination of good performance, as quantified by accuracy and other metrics, and an interpretable model increases trust. Neither deep learning based reconstruction nor classical iterative algorithms provide a path from pixel to sensor value in a way that a clinician can easily understand. While a reconstructed image may seem to increase interpretability, it is typically not created in an interpretable way and is not necessarily causally related to the classification decision. Omitting the reconstructed image, while increasing performance, would not therefore reduce the trust a clinician has in our direct to task contribution.

Cancer treatment regimens, especially for treatment-resistant lesions, are shifting towards adaptive or dynamic treatment models, such as the recent game theory-driven treatment of resistant prostate tumor patients [81]. However, these require accurate, unbiased, and specialized task-specific models. Our raw-to-task approach can be extended to develop models specializing in multiple tasks, not just diagnostics. Examples are prediction of lesion type, progression, localization, and tumor heterogeneity, all the way to successful treatment regimens ahead of time, paving the way for adaptive personalized medicine and disease management [71].

A key focus of this work was to leverage orthogonality in mitigating confounding factors induced by multi-frequency fusion. However, as noted as early as 1936 by Fisher et al. [82], orthogonal representations need not be informative, and thus, in a deep learning setting can also lead to orthogonal or independent encodings that are less or uninformative, as we encountered in our own experiments. The heterogeneity of lesions, especially malignant ones, ensures that no two malignant lesions will likely be the same, thus driving the need for diagnostic capability that focuses on identifying the diverse lesion types, not necessarily the reconstructed image.

V. CONCLUSION

We introduce deep learning based multi-frequency orthogonal fusion for diffuse optical tomography with end-to-end classification of malignancy of breast lesions. Orthogonal fusion of multi-frequency improved both image reconstruction quality and accuracy of tumoral and non-tumoral breast lesions' discrimination. In addition, we show that raw-to-task learning can improve classification without requiring reconstruction in a real time setting.

REFERENCES

[1] M. Akram, M. Iqbal, M. Daniyal, and A. U. Khan, "Awareness and current knowledge of breast cancer," *Biological research*, vol. 50, pp. 1–23, 2017.

[2] C. Coleman, "Early detection and screening for breast cancer," in *Seminars in oncology nursing*, vol. 33, no. 2. Elsevier, 2017, pp. 141–155.

[3] A. G. Waks and E. P. Winer, "Breast cancer treatment: a review," *Jama*, vol. 321, no. 3, pp. 288–300, 2019.

[4] M. K. Shetty, "Screening for breast cancer with mammography: current status and an overview," *Indian journal of surgical oncology*, vol. 1, pp. 218–223, 2010.

[5] S. Iranmakani, T. Mortezazadeh, F. Sajadian, M. F. Ghaziani, A. Ghafari, D. Khezerloo *et al.*, "A review of various modalities in breast imaging: technical aspects and clinical outcomes," *Egyptian Journal of Radiology and Nuclear Medicine*, vol. 51, no. 1, pp. 1–22, 2020.

[6] A. Dalla Mora, D. Contini, S. Arridge, F. Martelli, A. Tosi, G. Boso *et al.*, "Towards next-generation time-domain diffuse optics for extreme depth penetration and sensitivity," *Biomedical optics express*, vol. 6, no. 5, pp. 1749–1760, 2015.

[7] M. Applegate, R. Istfan, S. Spink, A. Tank, and D. Roblyer, "Recent advances in high speed diffuse optical imaging in biomedicine," *APL Photonics*, vol. 5, no. 4, p. 040802, 2020.

[8] C. Chen, F. Tian, H. Liu, and J. Huang, "Diffuse optical tomography enhanced by clustered sparsity for functional brain imaging," *IEEE transactions on medical imaging*, vol. 33, no. 12, pp. 2323–2331, 2014.

[9] H. Zhao, E. M. othersand Frijia, E. V. Rosas, L. Collins-Jones, G. Smith, R. Nixon-Hill *et al.*, "Design and validation of a mechanically flexible and ultra-lightweight high-density diffuse optical tomography system for functional neuroimaging of newborns," *Neurophotonics*, vol. 8, no. 1, p. 015011, 2021.

[10] M. Shokoufi and F. Golnaraghi, "Handheld diffuse optical breast scanner probe for cross-sectional imaging of breast tissue," *Journal of Innovative Optical Health Sciences*, vol. 12, no. 02, p. 1950008, 2019.

[11] M. L. Altoe, A. Marone, H. K. Kim, K. Kalinsky, D. L. Hershman, A. H. Hielscher *et al.*, "Diffuse optical tomography of the breast: a potential modifiable biomarker of breast cancer risk with neoadjuvant chemotherapy," *Biomedical Optics Express*, vol. 10, no. 8, pp. 4305–4315, 2019.

[12] S. Mahdy, O. Hamdy, M. A. Hassan, and M. A. Eldosoky, "A modified source-detector configuration for the discrimination between normal and diseased human breast based on the continuous-wave diffuse optical imaging approach: a simulation study," *Lasers in Medical Science*, pp. 1–10, 2021.

[13] T. Durduran, R. Choe, W. B. Baker, and A. G. Yodh, "Diffuse optics for tissue monitoring and tomography," *Reports on progress in physics*, vol. 73, no. 7, p. 076701, 2010.

[14] P. Taroni, A. M. Paganoni, F. Ieva, A. Pifferi, G. Quarto, F. Abbate *et al.*, "Non-invasive optical estimate of tissue composition to differentiate malignant from benign breast lesions: A pilot study," *Scientific reports*, vol. 7, no. 1, pp. 1–11, 2017.

[15] F. S. Azar, K. Lee, A. Khamene, R. Choe, A. Corlu, S. D. Konecky *et al.*, "Standardized platform for coregistration of nonconcurrent diffuse optical and magnetic resonance breast images obtained in different geometries," *Journal of biomedical optics*, vol. 12, no. 5, pp. 051 902–051 902, 2007.

[16] H. Ben Yedder, B. Cardoen, M. Shokoufi, F. Golnaraghi, and G. Hamarneh, "Multitask deep learning reconstruction and localization of lesions in limited angle diffuse optical tomography," *IEEE Transactions on Medical Imaging*, vol. 41, no. 3, pp. 515–530, 2021.

[17] V. Mudeng, G. Ayana, S.-U. Zhang, and S.-w. Choe, "Progress of near-infrared-based medical imaging and cancer cell suppressors," *Chemosensors*, vol. 10, no. 11, p. 471, 2022.

[18] C.-W. Sun, "Biophotonics for tissue oxygenation analysis," in *Biophotonics for Medical Applications*. Elsevier, 2015, pp. 301–320.

[19] Y. Hoshi and Y. Yamada, "Overview of diffuse optical tomography and its clinical applications," *Journal of Biomedical Optics*, vol. 21, no. 9, p. 091312, 2016.

[20] S. R. Arridge and J. C. Schotland, "Optical tomography: forward and inverse problems," *Inverse problems*, vol. 25, no. 12, p. 123010, 2009.

[21] F. Beca and K. Polyak, "Intratumor heterogeneity in breast cancer," in *Novel Biomarkers in the Continuum of Breast Cancer*. Springer, 2016, pp. 169–189.

[22] L. Zhang and G. Zhang, "Brief review on learning-based methods for optical tomography," *Journal of Innovative Optical Health Sciences*, vol. 12, no. 06, p. 1930011, 2019.

[23] H. Ben Yedder, B. Cardoen, and G. Hamarneh, "Deep learning for biomedical image reconstruction: A survey," *Artificial Intelligence Review*, pp. 1–33, 2020.

[24] G. M. Balasubramaniam, B. Wiesel, N. Biton, R. Kumar, J. Kupferman, and S. Arnon, "Tutorial on the use of deep learning in diffuse optical tomography," *Electronics*, vol. 11, no. 3, p. 305, 2022.

[25] H. Ben Yedder, M. Shokoufi, B. Cardoen, F. Golnaraghi, and G. Hamarneh, "Limited-angle diffuse optical tomography image reconstruction using deep learning," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 66–74.

[26] J. Yoo, S. Sabir, D. Heo, K. H. Kim, A. Wahab, Y. Choi *et al.*, "Deep learning diffuse optical tomography," *IEEE Transactions on medical imaging*, vol. 39, no. 4, pp. 877–887, 2019.

[27] M. Mozumder, A. Hauptmann, I. Nissilä, S. R. Arridge, and T. Tarvainen, "A model-based iterative learning approach for diffuse optical tomography," *IEEE Transactions on Medical Imaging*, vol. 41, no. 5, pp. 1289–1299, 2021.

[28] J. T. Smith, M. Ochoa, D. Faulkner, G. Haskins, and X. Intes, "Deep learning in macroscopic diffuse optical imaging," *Journal of Biomedical Optics*, vol. 27, no. 2, pp. 020 901–020 901, 2022.

[29] S. Sabir, S. Cho, Y. Kim, R. Pua, D. Heo, K. H. Kim *et al.*, "Convolutional neural network-based approach to estimate bulk optical properties in diffuse optical tomography," *Applied Optics*, vol. 59, no. 5, pp. 1461–1470, 2020.

[30] H. Ben Yedder, A. BenTaieb, M. Shokoufi, A. Zahiremami, F. Golnaraghi, and G. Hamarneh, "Deep learning based image reconstruction

for diffuse optical tomography," in *International Workshop on Machine Learning for Medical Image Reconstruction*. Springer, 2018, pp. 112–119.

[31] N. I. Nizam, M. Ochoa, J. T. Smith, and X. Intes, "Wide-field diffuse optical tomography using deep learning," in *Optical Tomography and Spectroscopy*. Optica Publishing Group, 2022, pp. OW4D–7.

[32] M. Mozumder, A. Hauptmann, S. R. Arridge, and T. Tarvainen, "Diffuse optical tomography utilizing model-based learning," in *Optics and the Brain*. Optica Publishing Group, 2022, pp. JTu3A–10.

[33] N. I. Nizam, M. Ochoa, J. T. Smith, and X. Intes, "Deep learning-based fusion of widefield diffuse optical tomography and micro-ct structural priors for accurate 3d reconstructions," *Biomedical Optics Express*, vol. 14, no. 3, pp. 1041–1053, 2023.

[34] X. Fang, C. Gao, Y. Li, and T. Li, "Solving heterogenous region for diffuse optical tomography with a convolutional forward calculation model and the inverse neural network," in *Advanced Optical Imaging Technologies III*, vol. 11549. SPIE, 2020, pp. 50–60.

[35] Y. Zhao, A. Raghuram, F. Wang, S. H. Kim, A. H. Hielscher, J. T. Robinson *et al.*, "Unrolled-dot: an interpretable deep network for diffuse optical tomography," *Journal of Biomedical Optics*, vol. 28, no. 3, p. 036002, 2023.

[36] Y. Zou, Y. Zeng, S. Li, and Q. Zhu, "Machine learning model with physical constraints for diffuse optical tomography," *Biomedical Optics Express*, vol. 12, no. 9, pp. 5720–5735, 2021.

[37] S. Li, M. Zhang, M. Xue, and Q. Zhu, "Difference imaging from single measurements in diffuse optical tomography: a deep learning approach," *Journal of Biomedical Optics*, vol. 27, no. 8, p. 086003, 2022.

[38] Y. Zou, Y. Zeng, S. Li, and Q. Zhu, "Unsupervised machine learning model for dot reconstruction," in *Optical Tomography and Spectroscopy of Tissue XIV*, vol. 11639. SPIE, 2021, pp. 23–36.

[39] N. Z. Shifa, M. M. R. Sayem, and M. A. Islam, "Improved image reconstruction using multi frequency data for diffuse optical tomography," in *2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD)*. IEEE, 2021, pp. 264–268.

[40] A. Godavarty, S. Rodriguez, Y.-J. Jung, and S. Gonzalez, "Optical imaging for breast cancer prescreening," *Breast Cancer: Targets and Therapy*, vol. 7, p. 193, 2015.

[41] M. Doulgerakis, A. T. Eggebrecht, and H. Dehghani, "High-density functional diffuse optical tomography based on frequency-domain measurements improves image quality and spatial resolution," *Neurophotonics*, vol. 6, no. 3, p. 035007, 2019.

[42] M. B. Unlu, O. Birgul, R. Shafiiha, G. Gulsen, and O. Nalcioglu, "Diffuse optical tomographic reconstruction using multifrequency data," *Journal of Biomedical Optics*, vol. 11, no. 5, p. 054008, 2006.

[43] H. K. Kim, U. J. Netz, J. Beuthan, and A. H. Hielscher, "Optimal source-modulation frequencies for transport-theory-based optical tomography of small-tissue volumes," *Optics express*, vol. 16, no. 22, pp. 18 082–18 101, 2008.

[44] X. Intes and B. Chance, "Multi-frequency diffuse optical tomography," *Journal of Modern Optics*, vol. 52, no. 15, pp. 2139–2159, 2005.

[45] V. Mudeng, W. Nisa, and S. S. Suprapto, "Computational image reconstruction for multi-frequency diffuse optical tomography," *Journal of King Saud University-Computer and Information Sciences*, 2021.

[46] M. B. Applegate, C. A. Gómez, and D. Roblyer, "Frequency selection in frequency domain diffuse optical spectroscopy," in *Optical Tomography and Spectroscopy of Tissue XIV*, vol. 11639. International Society for Optics and Photonics, 2021, p. 116390N.

[47] C. Chen, V. C. Kavuri, X. Wang, R. Li, H. Liu, and J. Huang, "Multi-frequency diffuse optical tomography for cancer detection," in *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2015, pp. 67–70.

[48] A. Zimek, E. Schubert, and H.-P. Kriegel, "A survey on unsupervised outlier detection in high-dimensional numerical data," *Statistical Analysis and Data Mining: The ASA Data Science Journal*, vol. 5, no. 5, pp. 363–387, 2012.

[49] A. Pifferi, A. Dalla Mora, L. Di Sieno, E. Ferocino, A. Tosi, E. Conca *et al.*, "Solus: an innovative multimodal imaging system to improve breast cancer diagnosis through diffuse optics and ultrasounds," in *Optical Tomography and Spectroscopy of Tissue XIV*, vol. 11639. International Society for Optics and Photonics, 2021, p. 116390C.

[50] G. Di Sciacca, G. Maffeis, A. Farina, A. Dalla Mora, A. Pifferi, P. Taroni *et al.*, "Evaluation of a pipeline for simulation, reconstruction, and classification in ultrasound-aided diffuse optical tomography of breast tumors," *Journal of biomedical optics*, vol. 27, no. 3, p. 036003, 2022.

[51] M. Althobaiti, H. Vavadi, and Q. Zhu, "Diffuse optical tomography reconstruction method using ultrasound images as prior for regularization matrix," *Journal of biomedical optics*, vol. 22, no. 2, pp. 026 002–026 002, 2017.

[52] D. Liu, Y. Zhang, L. Bai, P. Zhang, and F. Gao, "Combining two-layer semi-three-dimensional reconstruction and multi-wavelength image fusion for functional diffuse optical tomography," *IEEE Transactions on Computational Imaging*, vol. 7, pp. 1055–1068, 2021.

[53] H. O. Kazanci and O. Oral, "Frequency shifting model for diffuse optical tomography," *Optical and Quantum Electronics*, vol. 53, no. 11, pp. 1–6, 2021.

[54] G. A. Perkins, A. T. Eggebrecht, and H. Dehghani, "Multi-modulated frequency domain high density diffuse optical tomography," *Biomedical Optics Express*, vol. 13, no. 10, pp. 5275–5294, 2022.

[55] Y. Zhang, D. Sidibé, O. Morel, and F. Mériaudeau, "Deep multimodal fusion for semantic image segmentation: A survey," *Image and Vision Computing*, vol. 105, p. 104042, 2021.

[56] W. Guo, J. Wang, and S. Wang, "Deep multimodal representation learning: A survey," *IEEE Access*, vol. 7, pp. 63 373–63 394, 2019.

[57] G. Patel and J. Dolz, "Weakly supervised segmentation with cross-modality equivariant constraints." *Medical Image Analysis*, p. 102374, 2022.

[58] A. Nagrani, S. Yang, A. Arnab, A. Jansen, C. Schmid, and C. Sun, "Attention bottlenecks for multimodal fusion," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[59] R. J. Chen, M. Y. Lu, J. Wang, D. F. Williamson, S. J. Rodig, N. I. Lindeman *et al.*, "Pathomic fusion: an integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis," *IEEE Transactions on Medical Imaging*, 2020.

[60] A. Bozic, P. Palafox, J. Thies, A. Dai, and M. Nießner, "Transformerfusion: Monocular rgb scene reconstruction using transformers," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[61] A. Bardes, J. Ponce, and Y. LeCun, "VICReg: Variance-invariance-covariance regularization for self-supervised learning," *arXiv preprint arXiv:2105.04906*, 2021.

[62] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, "Barlow twins: Self-supervised learning via redundancy reduction," in *International Conference on Machine Learning*. PMLR, 2021, pp. 12 310–12 320.

[63] N. Bansal, X. Chen, and Z. Wang, "Can we gain more from orthogonality regularizations in training deep networks?" *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[64] L. Huang, X. Liu, B. Lang, A. W. Yu, Y. Wang, and B. Li, "Orthogonal weight normalization: Solution to optimization over multiple dependent stiefel manifolds in deep neural networks," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[65] K. Ranasinghe, M. Naseer, M. Hayat, S. Khan, and F. S. Khan, "Orthogonal projection loss," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 12 333–12 343.

[66] N. Braman, J. W. Gordon, E. T. Goossens, C. Willis, M. C. Stumpe, and J. Venkataraman, "Deep orthogonal fusion: Multimodal prognostic biomarker discovery integrating radiology, pathology, genomic, and clinical data," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2021, pp. 667–677.

[67] X. Zhen and S. Li, "Towards direct medical image analysis without segmentation," *arXiv preprint arXiv:1510.06375*, 2015.

[68] D. Wu, K. Kim, B. Dong, and Q. Li, "End-to-end abnormality detection in medical imaging," 2018.

[69] M. A. Hussain, G. Hamarneh, T. W. O'Connell, M. F. Mohammed, and R. Abugharbieh, "Segmentation-free estimation of kidney volumes in ct with dual regression forests," in *International Workshop on Machine Learning in Medical Imaging*. Springer, 2016, pp. 156–163.

[70] S. A. Taghanaki, N. Duggan, H. Ma, X. Hou, A. Celler, F. Benard *et al.*, "Segmentation-free direct tumor volume and metabolic activity estimation from pet scans," *Computerized Medical Imaging and Graphics*, vol. 63, pp. 52–66, 2018.

[71] K. Abhishek, J. Kawahara, and G. Hamarneh, "Predicting the clinical management of skin lesions using deep learning," *Scientific reports*, vol. 11, no. 1, pp. 1–14, 2021.

[72] A. B. Konovalov, E. A. Genina, and A. N. Bashkatov, "Diffuse optical mammotomography: state-of-the-art and prospects," *Journal of Biomedical Photonics & Engineering*, vol. 2, no. 2, pp. 020 202–1, 2016.

[73] Z. et al, "A comprehensive survey on transfer learning," *Proceedings of the IEEE*, vol. 109, no. 1, pp. 43–76, 2020.

[74] X. Yu, J. Wang, Q.-Q. Hong, R. Teku, S.-H. Wang, and Y.-D. Zhang, "Transfer learning for medical images analyses: A survey," *Neurocomputing*, vol. 489, pp. 230–254, 2022.