This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2023.3297173

Progressive Pretraining Network for 3D System Matrix Calibration in Magnetic Particle Imaging

Gen Shi, Lin Yin, Yu An, Guanghui Li, Liwen Zhang, Zhongwei Bian, Ziwei Chen, Haoran Zhang, Hui Hui and Jie Tian, *Fellow, IEEE*

***Abstract*— Magnetic particle imaging (MPI) is an emerging technique for determining magnetic nanoparticle distributions in biological tissues. Although system-matrix (SM)-based image reconstruction offers higher image quality than the X-space-based approach, the SM calibration measurement is time-consuming. Additionally, the SM should be recalibrated if the tracer's characteristics or the magnetic field environment change, and repeated SM measurement further increase the required labor and time. Therefore, fast SM calibration is essential for MPI. Existing calibration methods commonly treat each row of the SM as independent of the others, but the rows are inherently related through the coil channel and frequency index. As these two elements can be regarded as additional multimodal information, we leverage the transformer architecture with a self-attention mechanism to encode them. Although the transformer has shown superiority in multimodal fusion learning across several fields, its high complexity may lead to overfitting when labeled data are scarce. Compared with labeled SM (i.e., full size), low-resolution SM data can be easily obtained, and fully using such data may alleviate overfitting. Accordingly, we propose a pseudo-label-based progressive pretraining strategy to leverage unlabeled data. Our method outperforms existing calibration methods on a public real-world OpenMPI dataset and simulation dataset. Moreover, our method improves the resolution of two in-house MPI scanners without requiring full-size SM measurements. Ablation studies confirm the contributions of modeling SM inter-row relations and the proposed pretraining strategy.**

***Index Terms*— Magnetic particle imaging, system matrix, multimodal data, pretraining strategy.**

This work was supported in part by the National Key Research and Development Program of China under Grant: 2017YFA0700401; the National Natural Science Foundation of China under Grant: 62027901, 81827808, 81227901, 62201570, 82230067; Beijing Natural Science Foundation JQ22023; Guangdong Key Research and Development Program of China (2021B0101420005); CAS Youth Innovation Promotion Association under Grant Y2022055. The authors would like to acknowledge the instrumental and technical support of Multimodal Biomedical Imaging Experimental Platform, Institute of Automation, Chinese Academy of Sciences.(Corresponding author: Hui Hui, Jie Tian)

Gen Shi and Lin Yin contributed equally to this work.

Gen Shi, Yu An, Guanghui Li, Zhongwei Bian, Ziwei Chen, Haoran Zhang and Jie Tian with School of Engineering Medicine and School of Biological Science and Medical Engineering, Beihang University, Beijing, 100191, China, and also with the Key Laboratory of Big DataBased Precision Medicine (Beihang University), Ministry of Industry and Information Technology of China, Beijing, 100191, China (e-mail: {shigen, yuan1989, sy2110120, bianzw, chenziwei, hrzhang}@buaa.edu.cn, tian@ieee.org).

Lin Yin, Liwen Zhang and Hui Hui are with the CAS Key Laboratory of Molecular Imaging, Institute of Automation, Chinese Academy of Sciences, Beijing, 100190, China (e-mail: {yinlin2016, zhangliwen2018, hui.hui}@ia.ac.cn).

I. Introduction

Magnetic particle imaging (MPI) [1], [2] is an emerging medical imaging technique that provides high imaging speed and sensitivity [3]–[5]. MPI uses a tracer and the nonlinear response of magnetic nanoparticles (MNPs) in a magnetic field to determine their distribution. Additionally, new MPI designs are currently being developed [6]. MPI has been widely used in areas such as cell tracing [7], [8], functional neuroimaging [9], [10], and vessel imaging [11].

Two conventional reconstruction methods [12] for MPI are available: X-space- [13] and system-matrix (SM)-based [4] methods. Compared with the X-space-based method, the SM-based method achieves a higher image quality [14], but the SM measurement is time-consuming. For the SM measurement, a delta MNP sample should be repeatedly moved across each voxel in the field of view (FOV), and the corresponding signals are recorded. Each measurement takes approximately 15 h for an MPI system with a small 3D FOV ($30\,\mathrm{mm} \times 30\,\mathrm{mm} \times 30\,\mathrm{mm}$) [15]. Multiple averaging is commonly required to improve the SM measurement quality, significantly increasing the calibration time (averaging ten measurements can take more than 100 h). More importantly, the SM should be recalibrated when changes to the tracer's properties or magnetic field environment occur. Frequent SM recalibration results in excessive labor and time costs. Therefore, fast SM calibration is an area of research interest for MPI. Several compressed sensing (CS)- [16], [17] and deep learning-based methods [18], [19] have recently been proposed to reduce the SM calibration time. However, despite the success of existing studies on SM calibration, as reviewed in Section II, there is much room for improvement. In this study, we devise SM calibration improvements in two aspects:

1) Introduction of coil channel and frequency index to model SM inter-row relations. Existing methods often treat an SM row as an independent data point. This modeling approach neglects the SM integrity and the relationships between frequency components. In fact, the SM frequency components are not entirely independent. For example, each frequency component contains two additional information elements: the coil channel (i.e., the receiving coil obtaining a specific frequency component) and the frequency index. These elements can be regarded as additional multimodal

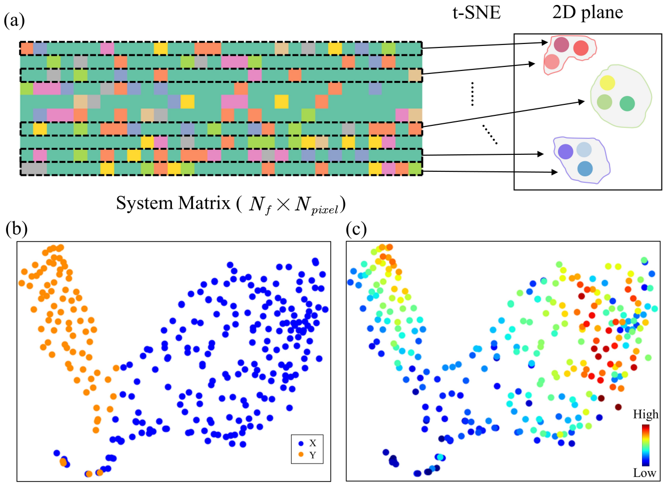

Fig. 1. Visualization of t-distributed stochastic neighbor embedding from SM rows. (a) The illustration of SM dimension reduction by the embedding method. (b), (c) show the visualization results for OpenMPI calibration dataset 5. Each point represents one SM row, and the color indicates its receiving coils in (b) and frequency in (c).

TABLE I
SYMBOLS AND INTERPRETATIONS.

Symbol	Interpretation
s_i^L	measured low-resolution SM component
s_i^H	measured high-resolution SM component
$\hat{s}_i^H$	predicted high-resolution SM component
h, w, d	the 3D shape of s_i^L, $h \times w \times d = N_{pixel}^L$
H, W, D	the 3D shape of s_i^H, $H \times W \times D = N_{pixel}^H$
$\mathfrak{p}_i, \mathfrak{f}_i$	coil channel and frequency index of s_i^L
e_i^p, e_i^f	embeddings of $\mathfrak{p}_i$ and $\mathfrak{f}_i$
x_i^L	the output of transformer encoder
x_i^H	the output of upsampling module with x_i^L as input
$\hat{x}_i^H$	the output of sucessive convolution operations
$\tilde{x}_i^H$	the output of skip connection with s_i^L as input
$z_i^{(l)}$,	the hidden output of l-th layer in encoder
p	patch size in SM component sequencing process
$\mathcal{N}_{ds}$	downsampling point set
F	hidden representation dimension in the encoder
C', C	hidden channels in the encoder and decoder, respectively
Φ_θ	trainable parameters in the proposed model

information. Consider a result on OpenMPI data (calibration 5) for illustrating the influence of the two elements. The dimension of each SM row is reduced using t-distributed stochastic neighbor embedding (t-SNE), as shown in Fig. 1(a), and the visualization results are shown in Fig. 1(b), (c). SM rows in the same receiving coil or with a close frequency index are usually clustered. Because the fusion of multimodal information can improve the model performance [20], [21], we integrate the coil channel and frequency index as multimodal information into a model to improve the SM calibration accuracy.

2) Use of unlabeled SM data through progressive pretraining. Deep learning methods have achieved great success for fast SM calibration [18], [19]. However, existing supervised models are limited because they require a large, labeled dataset (high-resolution SM). Insufficient labeled data may cause overfitting and poor performance. Because unlabeled SM data (low-resolution SM) can be obtained relatively fast and affect model performance, we use unlabeled data to increase the SM calibration accuracy.

Driven by the abovementioned analysis, we propose a progressive pretraining transformer-based network called ProTSM to handle multimodal information for fast 3D SM calibration. Because transformer has shown superiority in multimodal fusion learning across many fields [22], [23], we use the self-attention mechanism to integrate coil information. In particular, the coil information is interpreted as tokens by embedding layers and interacts with the SM row through the transformer's self-attention. Additionally, to prevent overfitting owing to the high complexity of the transformer, we propose a pseudo-label-based progressive pretraining strategy that uses unlabeled data. The proposed ProTSM was evaluated on real-world and simulation datasets for 3D SM calibration, and it notably outperformed similar methods.

Our main contributions of the proposed work are summarized as follows:

- We firstly take the coil channel and frequency index into consideration for SM calibration. Our visualization analysis shows that frequency components are not independent, and we explicitly model their relationships using the transformer to improve the calibration.
- We propose using unlabeled data with a progressive pretraining strategy. We generate pseudo-labels for the isolated unlabeled pretraining dataset. These data are used to train our model, which is then finetuned on accurately labeled data. Our results show that pretraining accelerates model convergence and improves the SM calibration performance.
- We propose a transformer-based 3D SM calibration framework. ProTSM is evaluated on real-world and simulation datasets and outperforms the state-of-the-art methods. Additionally, the proposed ProTSM is embedded into two in-house MPI systems to generate high-resolution images without requiring a full-size SM measurement.

II. RELATED WORK

Interpolation-based methods are straightforward and easy to implement for super-resolution SM calibration. The performance of bicubic and nearest-neighbor interpolation has been investigated in SM calibration [24]. Simple linear interpolation can help resolve high-resolution structures. Additionally, CS-based methods have admirable performance in super-resolution SM calibration. Knopp and Weber [25] first used CS to speed up SM calibration. They sparsified the SM using certain basis transformations, such as discrete Fourier and cosine transforms. Accordingly, many CS-based variant methods have been developed [16], [17], [26]–[28]. For example, Ilbey et al. [27] proposed a coded calibration scene method, which places multiple MNP samples inside the FOV in each MPI scan instead of using a single MNP sample, as in conventional methods. This operation increases the signal-to-noise ratio and significantly improves the conventional CS calibration.

MPI reconstruction [29], [30] and SM calibration [18], [19], [31], [32] have both demonstrated the efficacy of deep

This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2023.3297173

learning. For the MPI image reconstruction area, Gungor et al. [33] proposed a deep equilibrium-based model using learned data consistency. This method demonstrated excellent generalization and quick imaging. Similarly, deep learning-based methods for the SM calibration area can benefit from measured high-resolution SMs and integrate prior knowledge of SM calibration through training. Many deep learning models have been proposed for the SM calibration. For example, 3dSMRnet was the first model based on a convolutional neural network (CNN) for 3D SM calibration [18]. This model improved both SM calibration and image reconstruction.

The transformer architecture has recently emerged for diverse computer vision applications [34], [35]. Despite the success of CNN, long-range dependencies are not adequately modeled. The transformer architecture has also been applied to SM calibration. Gungor et al. [36] introduced a CNN-transformer hybrid model (TranSMS) for 2D SM calibration. TranSMS contains one CNN and one transformer branch for feature extraction. The fusion feature maps are then upsampled, and a high-resolution SM is generated through a data consistency module. This model shows a performance improvement compared with CNN-based methods.

Because the SM frequency components are inherently related, we can model these relationships using the multimodal information of coil channel and frequency index. Several studies have shown that multimodal information fusion improves model performance [20], [21], which is encouraging for SM calibration. In our previous conference paper [37], we preliminarily demonstrated feasibility of utilizing multimodal information using transformer. In this study, on the basis of introducing multimodal information, we propose a novel pretraining strategy to prevent potential overfitting caused by the high complexity of the transformer architecture. We also provide more extensive experiments and in-depth discussions to confirm the contribution of coil information and the effectiveness of our pretraining strategy. Overall, this study offers valuable insights and a comprehensive evaluation of our proposed method, which may advance the current researches on fast SM calibration.

III. PROPOSED PROTSM

The architecture of the proposed ProTSM is shown in Fig. 2(a). The transformer encodes the low-resolution SM and the multimodal tokens of the coil channel and frequency index. Then, the encoded hidden representation is upsampled and followed by successive convolution blocks to predict the high-resolution SM components. The adopted notation is listed in Table I, and details of the proposed model are provided in the following subsections.

A. Problem Formulation

Let $u \in \mathcal{C}^{N_f \times 1}$ and $S \in \mathcal{C}^{N_f \times N^H_{pixel}}$ be the measured voltage signals in an MPI scan and SM, respectively, where N_f and N^H_{pixel} denote the total number of frequency components and pixel number of high-resolution SM, respectively. Image reconstruction aims to solve the MNP concentration $c \in \mathcal{R}^{N^H_{pixel} \times 1}$ in $u = Sc$. The measurement of full-size (high-resolution) S is generally time-consuming. Therefore, a small size (low-resolution) SM, $S^L \in \mathcal{R}^{N_f \times N^L_{pixel}}$, is measured in an attempt to recover the full-size SM, $S^H \in \mathcal{R}^{N_f \times N^H_{pixel}}$.

Each row of S^L is considered a low-resolution 3D image with two channels (real and imaginary channels) $s^L_i \in \mathcal{R}^{2 \times h \times w \times d}$ with $h \times w \times d = N^L_{pixel}$. Additionally, each SM component, s^L_i, comprises a coil channel $\mathfrak{p}_i$ and frequency index $\mathfrak{f}_i$. $\mathfrak{p}_i \in \{0, 1, 2\}$ is a discrete variable, which denotes the spatial coil related to s^L_i. $\mathfrak{f}_i$ is the frequency index of s^L_i, and the range of values of $\mathfrak{f}_i$ depends on the MPI system and filtered frequency components (e.g., $50\,\mathrm{kHz}{\sim}500\,\mathrm{kHz}$ in the OpenMPI dataset). $\mathfrak{p}_i$ and $\mathfrak{f}_i$ are auxiliary and multimodal data related to s^L_i. The goal is to recover $s^H_i \in \mathcal{R}^{2 \times H \times W \times D}$ using a deep learning model, $f(\cdot)$, with parameters Φ_θ (i.e., $\hat{s}^H_i = f(s^L_i, \mathfrak{p}_i, \mathfrak{f}_i | \Phi_\theta)$).

B. Progressive Pretraining Strategy

The flowchart of the proposed pretraining strategy and finetuning process is shown in Figs. 2(b) and 2(c), respectively. We first collect a large unlabeled dataset, $\{\mathcal{S}^{un}\}$, and obtain pseudo-labels $\{\mathcal{Y}\}$ using a super-resolution model. This model can be a simple linear model (trilinear interpolation) or a trained deep learning model. The proposed model is then pretrained on this large dataset and optimized using pseudo-labels as follows:

$$\min_{\Phi_\theta} \mathcal{L}(s^{un}_i, y_i, \Phi_\theta) = \|y_i - f(s^{un}_i, \mathfrak{p}_i, \mathfrak{f}_i | \Phi_\theta)\|_1, \quad (1)$$

$$\Phi_\theta = \Phi_\theta - \eta \cdot \nabla \mathcal{L}(s^{un}_i, y_i, \Phi_\theta), \quad (2)$$

where s^{un}_i and y_i are pretraining data represented by $\{\mathcal{S}^{un}\}$ and $\{\mathcal{Y}\}$, respectively, and η is the learning rate. Following pretraining, the model has better initial weight parameters Φ^{pre}_θ than those obtained through random initialization. The model is then finetuned on an accurately labeled dataset, $\{\mathcal{S}^L\}$, $\{\mathcal{S}^H\}$ starting with pretraining initialization and a smaller learning rate.

$$\Phi_\theta = \Phi_\theta - \eta \cdot \nabla \|s^H_i - f(s^L_i, \mathfrak{p}_i, \mathfrak{f}_i | \Phi^{pre}_\theta)\|_1, \quad (3)$$

The proposed pretraining strategy achieves the following improvements while fully using low-resolution SM data:

1) The pretrained model performs a weak super-resolution SM calibration, which improves the performance of the SM calibration and serves as a suitable initialization for optimization through supervised learning.
2) Compared with supervised methods, our model leverages low-resolution SM data. Hence, the risk of overfitting owing to limited SM data is mitigated.
3) Compared with training from scratch, finetuning simply optimizes our model from a weak to a more refined one, hastening the training convergence.

C. Transformer Encoder with Coil Embedding

1) Embedding of Coil Channel and Frequency Index: . Because $\mathfrak{p}_i$ and $\mathfrak{f}_i$ are single numeric variables, we project them

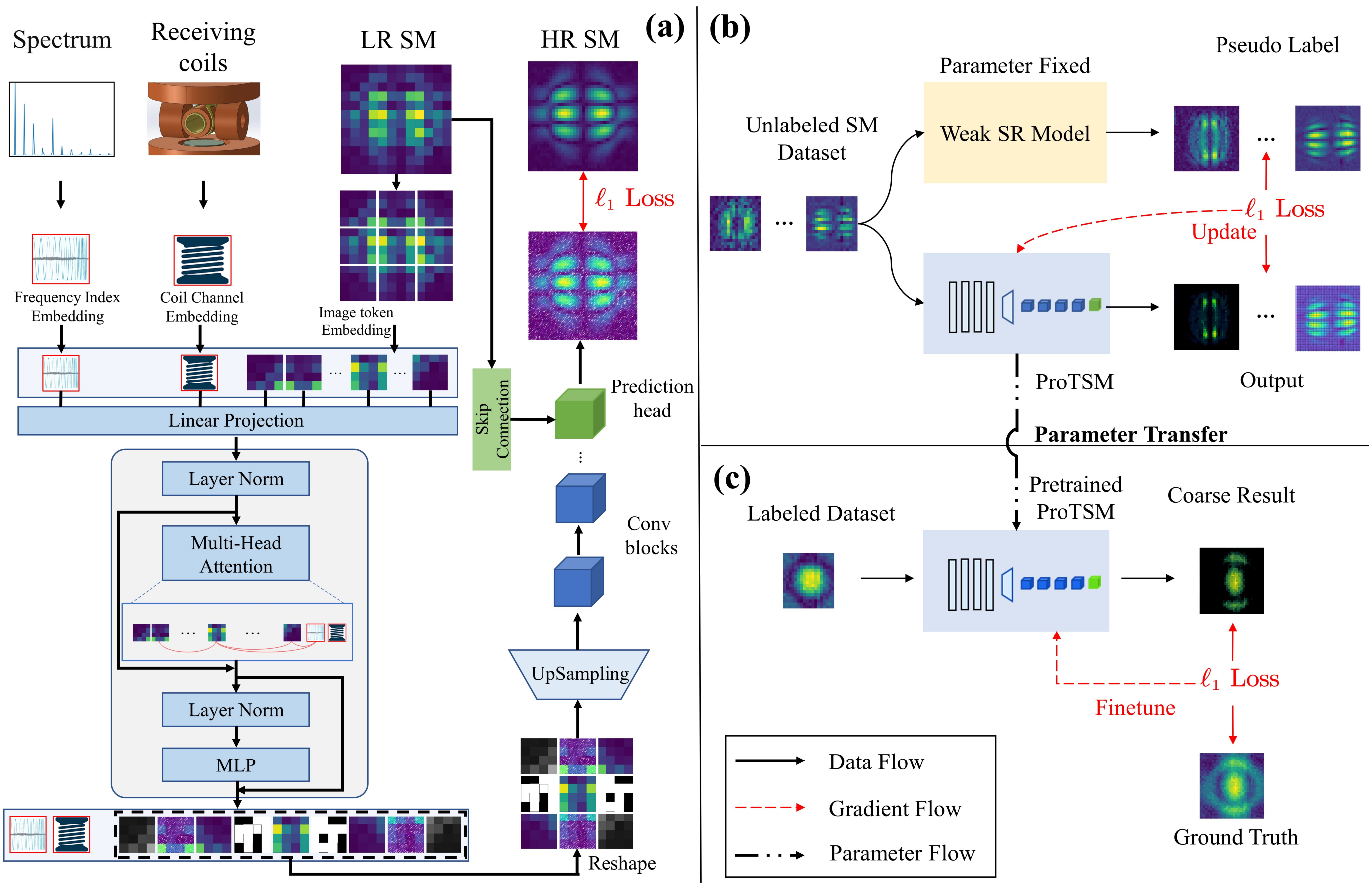

Fig. 2. (a) The overall framework of the proposed method. (b) The illustration of our proposed pseudo-label-based pretraining strategy. (c) The finetune process after pretraining.

onto a vector space for computation. We use the following linear and embedding layers for projection:

$$e_i^p = \mathrm{EmbeddingLayer}(\mathfrak{p}_i) \in \mathcal{R}^{1\times F}, \tag{4}$$

$$e_i^f = \mathrm{LinearLayer}(\mathfrak{f}_i) \in \mathcal{R}^{1\times F}, \tag{5}$$

where F denotes the latent representation dimension.

2) SM Component Sequencing: . To handle the 3D image s_i^L as the input for the transformer encoder, we first reshape it as 1D sequence tokens $s_i^L \rightarrow x_i \in \mathcal{R}^{(\frac{h}{p}\cdot\frac{w}{p}\cdot\frac{d}{p})\times(C\cdot p^3)}$. Then, a linear layer projects the tokens into latent space $z_i = W_p x_i + b_p$, where $z_i \in \mathcal{R}^{(\frac{h}{p}\cdot\frac{w}{p}\cdot\frac{d}{p})\times F}$ and W_p and b_p are trainable parameters.

Before feeding z_i into the transformer encoder, e_i^p and e_i^f are added to z_i, with e_i^p and e_i^f serving as global tokens used in self-attention calculations with other image tokens. Therefore, the final input is constructed as $z_i = [W_p x_i + b_p; e_i^p; e_i^f] \in \mathcal{R}^{(N+2)\times F}$, where $N = \frac{h}{p}\cdot\frac{w}{p}\cdot\frac{d}{p}$.

3) Transformer Encoding: . Following an existing method [34], we add absolute position embeddings $e_{pos} \in \mathcal{R}^{(N+2)\times F}$ to label the patch position, i.e., $z_i = [W_p x_i + b_p; e_i^p; e_i^f] + e_{pos}$. Compared with its relative counterpart, absolute position encoding explicitly indicates the spatial location relationship between image tokens, likely supporting dense prediction (e.g., super-resolution reconstruction).

The transformer encoder contains two modules: multi-head self-attention MSA and multilayer perceptron MLP. Encoding can be expressed as follows:

$$z_i^{(l)'} = \mathrm{MSA}(\mathrm{LayerNorm}(z_i^{(l-1)})) + z_i^{(l-1)}, \tag{6}$$

$$z_i^{(l)} = \mathrm{MLP}(\mathrm{LayerNorm}(z_i^{(l)'})) + z_i^{(l)'}, \tag{7}$$

where $z_i^{(l)'}$ and $z_i^{(l)}$ are the hidden result and the output of layer l, respectively. $\mathrm{MSA}(\cdot)$ is the key operation of the transformer and can be expressed as

$$\mathrm{MSA}(z_i) = \mathop{\Big\|}_{h=1}^{H} \frac{Q^h(z_i)\cdot K^h(z_i)^{\mathrm{T}}}{\sqrt{d}} V^h(z_i), \tag{8}$$

where $\|$ and H are the concatenation operation and number of heads, respectively; $Q(\cdot)$, $K(\cdot)$, and $V(\cdot)$ are linear transformation operations, with $Q(z_i) = Wq \cdot z_i$; and d denotes the number of dimensions in this head.

The information from e_i^p and e_i^f is encoded into s_i^L using the multi-head self-attention module. Additionally, each s^L has the same encoding parameters of $\mathfrak{p}$ and $\mathfrak{f}$. If two SM components have the same coil channel or similar frequency index, their e^p and e^f are the same or similar, respectively. Thus, we establish the relationship between the SM components using the coil channel and frequency index.

This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2023.3297173

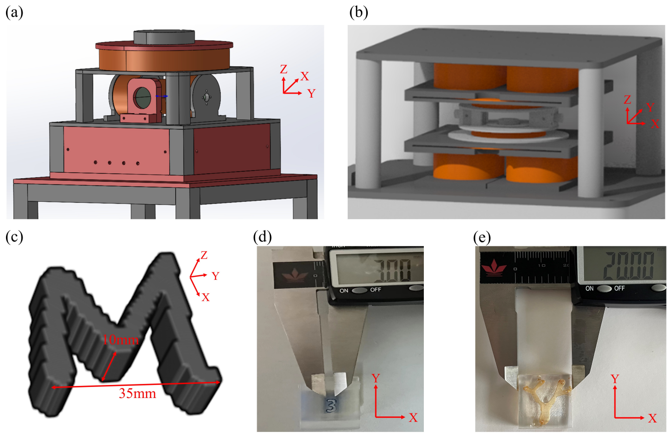

Fig. 3. (a),(b) show the 3D schematic diagrams of the field-free point (a) and field-free line (b) scanners. (c) The numerical phantom "M" used in simulation dataset. (d), (e) show the phantoms used in field-free point scanner (d) and field-free line (e) scanner for 2D imaging.

D. Decoder

The decoder contains upsampling and convolution blocks. First, it upsamples the output of the transformer encoder before generating high-resolution frequency components through successive 3D convolution blocks.

Considering that e_i^p and e_i^f are encoded into image tokens, they are not involved in SM construction during decoding. Let $z_i^L \in \mathcal{R}^{N\times F}$ be the output of the transformer encoder without coil tokens. We reshape z_i^L into a 3D image $z_i^L \rightarrow x_i^L \in \mathcal{R}^{C'\times h\times w\times d}$ and upsample x_i^L to obtain a high-resolution feature map through 3D pixel shuffling as follows:

$$x_i^H = \mathrm{UpSampling}(x_i^L) \in \mathcal{R}^{\frac{C'}{r^3}\times H\times W\times D}, \tag{9}$$

where x_i^H and r denote the hidden representation after upsampling and the downsampling ratio, respectively. The subsequent convolution operations produce the feature map for the prediction header (i.e., $1\times1\times1$ kernel convolution operation):

$$\hat{x}_i^H = \mathrm{Conv3D}(x_i^H) \in \mathcal{R}^{C\times H\times W\times D}. \tag{10}$$

E. Skip Connection

To alleviate the potential vanishing gradient problem in the deep network, we add a skip connection to our model. In particular, we directly upsample the original 3D SM components and extract shallow feature map $\tilde{x}_i^H$ as follows:

$$\tilde{x}_i^H = \mathrm{Conv3D}(\mathrm{UpSampling}(s_i^L)) \in \mathcal{R}^{C\times H\times W\times D}. \tag{11}$$

Finally, we aggregate $\tilde{x}_i^H$ and $\hat{x}_i^H$ to predict the high-resolution component $\hat{s}_i^H$ using the prediction header as follows:

$$\hat{s}_i^H = \mathrm{Conv3D}_{1\times1\times1}(\hat{x}_i^H + \tilde{x}_i^H) \in \mathcal{R}^{2\times H\times W\times D}. \tag{12}$$

IV. Datasets and Experimental Setup

A. Datasets

1) Evaluation Datasets: We evaluated the proposed ProTSM on two datasets:

TABLE II
SM Calibration Results on OpenMPI and Simulation Datasets.

Dataset	OpenMPI		Simulation	
Ratio	2×	4×	2×	4×
Method	nRMSE	nRMSE	nRMSE	nRMSE
Bicubic	5.44%	8.91%	7.27%	18.23%
Trilinear	5.27%	6.80%	6.95%	17.83%
CS	4.40%	7.70%	11.82%	21.68%
SRCNN	3.55%	5.18%	2.22%	5.22%
VolumeNet	3.79%	5.90%	3.22%	6.91%
3dSMRnet	4.02%	4.86%	1.01%	2.75%
MetaBlock	3.60%	4.51%	0.93%	2.81%
IDL	3.37%	4.56%	0.99%	2.74%
ProTSM	**3.08%**	**4.10%**	**0.72%**	**2.70%**

- **OpenMPI dataset**. OpenMPI is the first open-source MPI dataset [38]. It contains SM calibration and phantom measurements from multiple MNPs. Similar to [18], we used SM calibration experiment 7 with Synomag-D MNPs (Micromod GmbH, Germany) to construct the training set and evaluated the model performance on calibration experiment 6 with Perimag MNPs (Micromod GmbH, Germany). This setting was intended to evaluate the generalization ability for different MNP types. In both training and test sets, we only preserved the SM rows with a signal-to-noise ratio of $SNR > 3$, leaving 4129 and 3290 for training and test sets, respectively.
- **Simulation dataset**. We rewrote a 3D version for simulating SMs based on a code[1] and [39]. The FOV size was $40\,\mathrm{mm} \times 40\,\mathrm{mm} \times 40\,\mathrm{mm}$ and the grid size was $40\times40\times40$. The sampling frequency was $1\,\mathrm{MHz}$. The drive frequencies along the X, Y, and Z axes were $24.51\,\mathrm{kHz}$, $26.04\,\mathrm{kHz}$, and $25.25\,\mathrm{kHz}$, respectively. The MNP temperature was $300\,\mathrm{K}$, and the Boltzmann constant k_B was set as 1.38×10^{-23}. We evaluated the model generalization performance for different MNP diameters and selection field gradients. In particular, the training set included three 3D SMs (gradients of $0.5\,\mathrm{T/m}$, $1\,\mathrm{T/m}$, and $5\,\mathrm{T/m}$). The MNP diameter was $25\,\mathrm{nm}$. For the testing set, the SM gradient and the MNP diameter were $1\,\mathrm{T/m}$ and $12.5\,\mathrm{nm}$, respectively. The remaining data for training and test sets are 3933 and 1311, respectively. The phantom used for imaging is shown in 3(c)

2) Pretraining Dataset: We obtained low-resolution SM data from OpenMPI calibration experiments 7, 8, and 9. In particular, we extracted $20\times20\times20$ and $10\times10\times10$ SM samples for downsampling ratios of 2 and 4, respectively. This pretraining dataset contains 14596 samples. Then, we obtained pseudo-labels using the super-resolution CNN (SRCNN) [40] model trained on the OpenMPI training set.

3) In-House Datasets for Generalization Ability Evaluation: We evaluated the proposed ProTSM trained on the OpenMPI dataset using two in-house MPI systems: field-free point (FFP) and field-free line (FFL) scanners. The 3D model schematic diagrams for the two scanners are shown in Figs. 3(a) and 3(b), respectively. For the FFP scanner, the selection field gradient was $\{-1.7, -1.7, 3.4\}\mathrm{T/m}$ along the axes X, Y, Z. The excitation frequency along the X axis was $25\,\mathrm{kHz}$ and

[1] https://github.com/OS-MPI/Educational_Simulations

This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2023.3297173

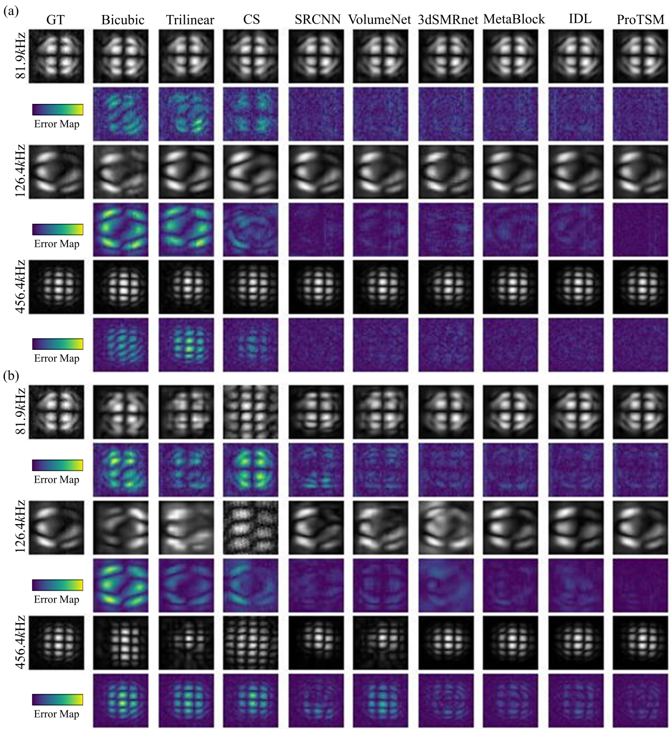

Fig. 4. The visualization results of three reconstructed SM rows (center slice) for downsampling ratio 2 (a) and 4 (b), respectively.

the driving frequency along the Y axis was $20\,\mathrm{Hz}$. A Cartesian trajectory was used to scan the FOV. The sampling frequency was $2.5\,\mathrm{MHz}$. The FOV of the MPI scanner was $20\,\mathrm{mm} \times 20\,\mathrm{mm}$. A delta sample ($2\,\mathrm{mm} \times 2\,\mathrm{mm}$) filled with Perimag MNPs was used to measure the low-resolution SM with a grid size of 10×10. For image reconstruction, the frequency components were selected using the formula $f = m_x f_x + m_y f_y$. In this study, $m_x \in [1, 13]$, $m_y \in [-7, 7]$ and only frequency components with $f < 330\,\mathrm{kHz}$ are used. Finally, 195 frequency components were preserved. This FFP instrument uses active compensation techniques to minimize the influence of excitation feed-through, and the base frequency signal was unfiltered. The phantom used for imaging is shown in Fig. 3(d). For the FFL scanner, the selection field gradient was $0.6\ \mathrm{T/m}$ along the X, and the drive frequency was 2.51 kHz. For 2D imaging, the object to be imaged rotates along the Z-axis in the FOV. The sampling frequency was $1\,\mathrm{MHz}$. The FFL scanner was rotated along the XY plane from 0 to 180° with increments of 12° (15 measured angles). We measured a square grid of 9×9 for the SM with a delta sample ($3 \times 3\mathrm{mm}^2$). The second through thirteenth frequency components for each angle (totaling $15 \times 12 = 180$ frequency components) were used for image reconstruction. The phantom used for imaging is shown in Fig. 3(e). We stacked the replicated 2D frequency components along the Z axis to create 3D data. Then, the predicted 3D high-resolution data (i.e., grid size of $40 \times 40 \times 40$) were averaged along the Z axis for 2D image reconstruction. We did not measure the high-resolution SM but conducted a qualitative analysis of the reconstructed image.

B. Implementation Details

The proposed ProTSM contains four transformer layers and four 3D convolutions per upsampling block. In this study, the hidden representation dimension F was 1024. The number of heads was eight, each of which had 128 dimensions (denoted by d) per head. The number of channels, C, for the convolutions was 64. For pretraining, the batch size was 50 and the learning rate was 5×10^{-4}. We pretrained the model for 50 epochs. For finetuning, the batch size was eight and the learning rate was 1×10^{-3} (half the learning rate for the encoder). We first trained the model for ten epochs using linear warmup and then for 50 and 100 epochs using a constant learning rate for downsampling ratios of 2 and 4, respectively. We conducted two experiments using different downsampling ratios (2 and 4) on each dataset. The model contained two upsampling blocks for a downsampling ratio of 4. The patch size was set to two and one for downsampling ratios of 2 and 4, respectively. For image reconstruction based on the calibrated SM, we used the kaczmarzReg algorithm[2] with parameter $\lambda = 0.75$ over three iterations.

C. Baselines and Evaluation Metrics

- **Bicubic Interpolation** [24]. Bicubic interpolation is a common super-resolution reconstruction method. However, because it can only process 2D images, we applied bicubic interpolation twice to perform a 3D upsampling. In particular, we first upsampled the SM component along the XY and then along the Z axis.
- **Trilinear Interpolation** [41]. Trilinear interpolation calculates the values of points in a cube based on the values of its vertices.
- **CS** [27]. CS assumes that the SM components are sparse after applying the discrete cosine transform DCT. We obtained the low-resolution data through Poisson disc sampling and optimized the following problem: $\min_{\hat{s}_i^H} \|\mathrm{DCT}(\hat{s}_i^H)\|_1$ subject to $\mathrm{Poisson}(\hat{s}_i^H) = s_i^L$.
- **SRCNN** [40]. SRCNN is the first CNN-based super-resolution reconstruction model. It first upsamples low-resolution images through bilinear interpolation before reconstructing high-resolution images using three convolutions.
- **VolumeNet** [42]. VolumeNet is a CNN-based super-resolution model designed for 3D medical images. It contains several parallel branches for multiscale feature extraction. The features are aggregated to generate a high-resolution image through voxel shuffling.
- **3dSMRnet** [18]. 3dSMRnet is a state-of-the-art method for super-resolution 3D SM calibration. It leverages residual-in-residual dense blocks to extract features from low-resolution SM components. Then, it upsamples the feature maps and reconstructs high-resolution SM components using 3D convolutions. We executed the open-source code at the website[3].

[2] https://github.com/MagneticParticleImaging/MDF/tree/master/python
[3] https://github.com/Ivo-B/3dSMRnet

This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2023.3297173

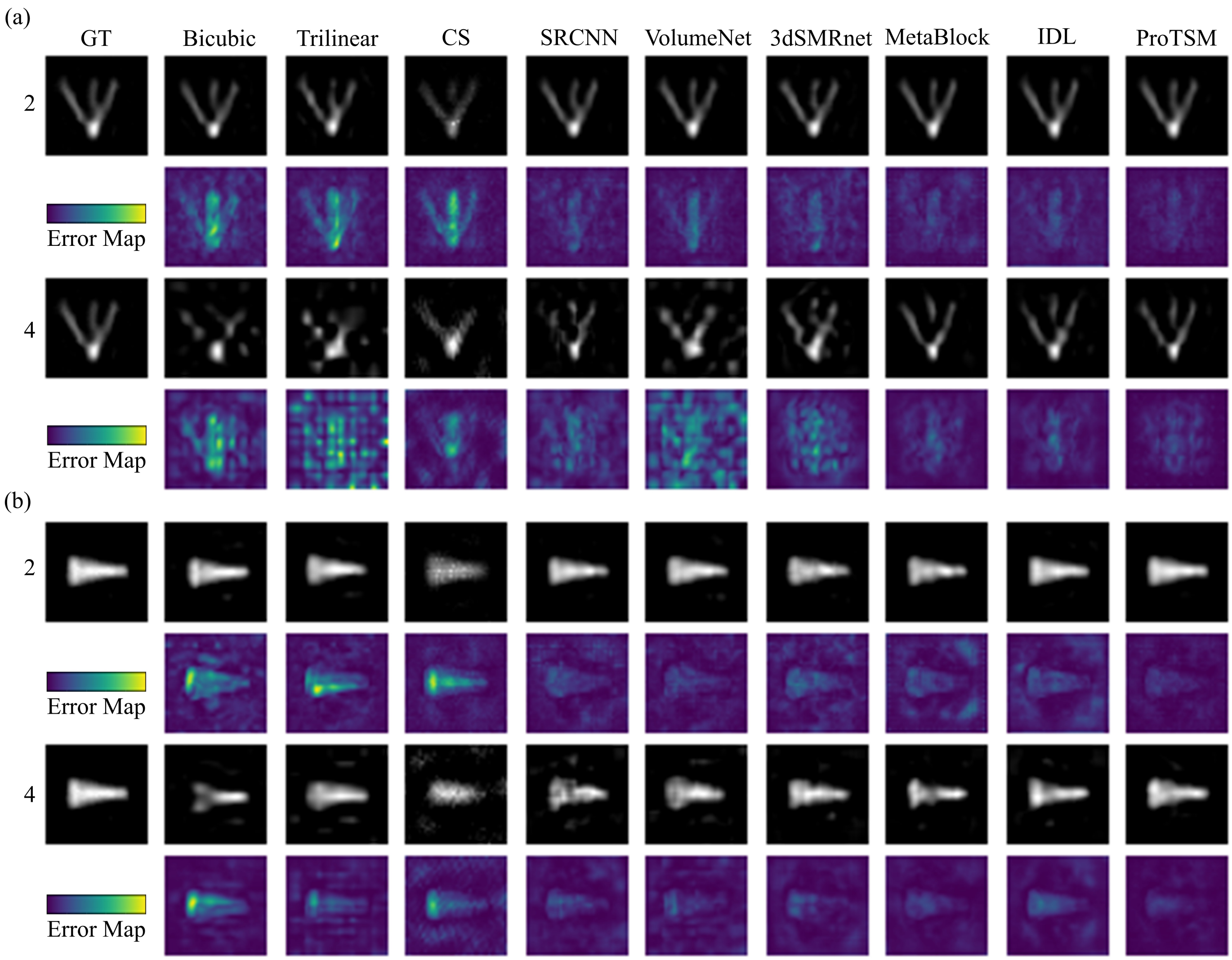

Fig. 5. The image reconstruction result for resolution and shape phantom in OpenMPI dataset. The first row shows the reconstructed image, and the second row shows the corresponding 3D error map that is averaged in Z-axis. Number "2" and "4" indicate the downsampling ratio. GT image is reconstructed by the measured full-size SM.

In addition to the above-mentioned baseline models, we present two competitive baselines that use coil information:

- **MetaBlock** [43]. MetaBlock uses an attention-based mechanism to enhance image features using non-image data (such as age and gender). In this study, the frequency index and coil channel represent the non-image data.
- **IDL** [44]. IDL proposes a multistage interactive fusion strategy to convolve image and non-imaging data. Instead of simple concatenation of multimodal data, this model uses channel-wise multiplication at each feature map downsampling level.

In our 2D experiments, we select the same baseline models as the recent work [36], and the extra methods are listed below:

- **VDSR** [45]. VDSR uses a very deep CNN-based neural network model for super-resolution tasks. This model learns the residual between the low- and high-resolution images to address the gradient vanishing and explosion problem.
- **TranSMS** [36]. TranSMS is the most recent state-of-the-art model for 2D SM calibration. This model proposes a two-branch architecture with a convolutional branch and a transformer branch. The transformer branch contains a novel transformer block with a convolution-based patch embeded method.

For each experiment, both the baseline models and our proposed model required the same number of calibration measurements. For the SM calibration, we obtained the normalized root-mean-square error (nRMSE) as the evaluation metric, as in [18]:

$$\mathrm{nRMSE}(\hat{s}_i^H) = \frac{\|\hat{s}_i^H - s_i^H\|_F}{\max(|s_i^H|) - \min(|s_i^H|)}, \tag{13}$$

where $\|\cdot\|_F$ denotes the Frobenius norm, $|\cdot|$ denotes the complex modulus, and $\hat{s}_i^H$ and s_i^H are converted into the complex format for evaluation.

To evaluate a reconstructed image, we calculated the peak signal-to-noise ratio (PSNR), structure similarity index measure (SSIM), and nRMSE.

TABLE III
IMAGE RECONSTRUCTION RESULTS BASED ON CALIBRATED SM ON OPENMPI DATASET.

Phantom	Resolution						Shape					
ratio		2×			4×			2×			4×	
Method	nRMSE↓	PNSR↑	SSIM↑	nRMSE↓	PNSR↑	SSIM↑	nRMSE↓	PNSR↑	SSIM↑	nRMSE↓	PNSR↑	SSIM↑
Bicubic	2.15%	33.34	0.8155	5.51%	25.18	0.3360	4.11%	27.73	0.6269	7.93%	22.01	0.4357
Trilinear	2.11%	33.50	0.8456	7.14%	22.93	0.2133	3.68%	28.69	0.7250	5.46%	25.25	0.4568
CS	2.08%	33.64	0.8324	3.94%	28.08	0.6107	3.75%	28.53	0.7162	6.95%	23.15	0.3618
SRCNN	1.15%	38.82	0.8995	3.55%	29.00	0.4498	2.58%	31.76	0.7689	4.25%	27.42	0.5754
VolumeNet	1.28%	37.89	0.9177	4.30%	27.32	0.3984	2.12%	33.46	0.8110	4.32%	27.29	0.5216
3dSMRnet	1.32%	37.56	0.8660	3.63%	28.81	0.4687	2.87%	30.85	0.7205	3.93%	28.11	0.5706
MetaBlock	1.07%	39.45	0.9112	2.279%	32.85	0.7196	2.49%	32.09	0.7796	4.40%	27.13	0.6036
IDL	1.06%	37.47	0.8914	2.276%	32.86	0.6908	2.63%	31.61	0.7182	3.39%	29.38	0.6651
ProTSM	**0.86%**	**41.43**	**0.9410**	**2.13%**	**33.43**	**0.7376**	**1.60%**	**35.90**	**0.8763**	**2.64%**	**31.57**	**0.7540**

TABLE IV
IMAGE RECONSTRUCTION RESULTS BASED ON CALIBRATED SM IN SIMULATION DATASET.

Phantom	M					
Ratio		2×			4×	
Method	nRMSE↓	PNSR↑	SSIM↑	nRMSE↓	PNSR↑	SSIM↑
Bicubic	3.47%	29.19	0.9285	8.88%	21.04	0.7194
Trilinear	3.33%	29.62	0.9310	8.57%	21.34	0.7306
CS	4.49%	26.96	0.9010	9.99%	20.01	0.6713
SRCNN	1.74%	35.21	0.9613	2.13%	33.42	0.9561
VolumeNet	1.46%	36.71	0.9736	2.60%	31.71	0.9552
3dSMRnet	1.32%	37.61	0.9772	1.66%	35.60	0.9605
MetaBlock	1.30%	37.74	0.9767	1.67%	35.54	0.9628
IDL	1.43%	36.87	0.9742	1.78%	35.00	0.9629
ProTSM	**1.22%**	**38.25**	**0.9804**	**1.49%**	**36.53**	**0.9742**

TABLE V
2D SM CALIBRATION RESULTS COMPARED WITH SOTA METHODS IN OPENMPI DATASET.

Ratio	2×	4×	8×
Method	nRMSE	nRMSE	nRMSE
Bicubic	4.55%	18.13%	52.02%
Bicubic (str.)	16.86%	47.41%	92.08%
CS	8.81%	51.48%	101.31%
SRCNN	50.88%	62.81%	106.76%
VDSR	3.34%	11.83%	113.81%
2d-SMRnet	6.86%	17.22%	78.88%
TranSMS	3.32%	10.66%	114.45%
ProTSM	**3.13%**	**9.88%**	**49.98%**

V. RESULTS

A. SM Calibration

Table II lists the 3D SM calibration results for the two datasets. The proposed ProTSM is highly superior to the other evaluated methods on the OpenMPI dataset in terms of nRMSE (3.08% and 4.10% for downsampling ratios of 2 and 4, respectively), with an improvement of approximately 15% over the best single modal-based method. Additionally, the proposed ProTSM achieves a relative improvement of approximately 9.5% compared with other multimodal-based methods. ProTSM also performs the best on the simulation dataset, with nRMSE values of 0.72% and 2.70% for downsampling ratios of 2 and 4, respectively.

Fig. 4 shows the center slice of the reconstructed 3D SM data for a qualitative evaluation. Overall, the deep learning models, such as SRCNN, VolumeNet and 3dSMRnet outperform other methods for the two downsampling ratios. The CS- and interpolation-based methods cannot use prior knowledge from the existing high-resolution SM data. Consequently, they are unable to provide satisfactory calibration accuracy. For a large downsampling ratio (Fig. 4(b)), the proposed ProTSM produces the best SM recovery results.

B. Evaluation of Image Reconstruction

We evaluated the image reconstruction performance using a super-resolution calibrated SM. For image reconstruction, we selected the phantom shape and resolution from the OpenMPI dataset. Additionally, we simulated numerical phantom M (see 3(c)) in the simulation dataset. The corresponding reconstruction results are listed in Tables III and IV.

The results of image reconstruction and SM calibration are consistent. The proposed ProTSM achieves the best performance for the three metrics (nRMSE, PSNR, and SSIM) on both OpenMPI and simulation datasets. On the OpenMPI dataset, ProTSM outperforms single-modal-based methods at high downsampling ratios. The PSNR of ProTSM and the best single-modal-based model are 35.90 and 33.46 (7.29% improvement) for phantom shape, respectively, for a downsampling ratio of 2. The PSNR of ProTSM and the best single-modal-based model are 31.57 and 28.11 (12.3% improvement) for a ratio of 4. A similar trend is observed for phantom shape. Our proposed ProTSM still performs better than the two multimodal methods. On the simulation dataset, ProTSM also performs better (PSNR of 38.25 and 36.53 for downsampling ratios of 2 and 4, respectively). Therefore, ProTSM consistently outperforms the other evaluated methods.

Fig. 5 shows two reconstructed images for qualitative evaluation. The figure shows the center slice of 3D images and the 3D error map averaged along the Z axis for phantom resolution. All methods provide an acceptable image quality in the center slice for a downsampling ratio of 2. However, the error maps show how poorly the interpolation-based methods perform with 3D images. When the downsampling ratio is 4, the baseline models reconstruct low-quality images polluted with noise and artifacts. Conversely, the proposed ProTSM provides a better image quality. These qualitative results demonstrate that ProTSM is robust even with a high downsampling ratio.

C. Comparisons with State-of-the-Art 2D Methods

For comparison with the TranSMS state-of-the-art model for 2D SM calibration, we adapted the proposed ProTSM

This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2023.3297173

TABLE VI

2D SM CALIBRATION AND IMAGE RECONSTRUCTION RESULTS OF THE 4 REPRESENTATIVE METHODS IN OPENMPI DATASET. THE METRIC NRMSE IS USED TO ASSESS SM RECOVERY AND METRICS PSNR, SSIM ARE USED TO ASSESS IMAGE QUALITY RECONSTRUCTED BY THE SM.

Ratio		4×			8×	
Method	nRMSE	PNSR	SSIM	nRMSE	PNSR	SSIM
Bicubic	47.45%	28.95	0.7684	68.21%	20.60	0.2266
SRCNN	28.96%	36.32	0.9253	71.58%	24.75	0.3229
TranSMS	24.70%	42.80	0.9716	81.95%	24.80	0.4220
ProTSM	23.84%	44.78	0.9848	65.65%	29.47	0.5250

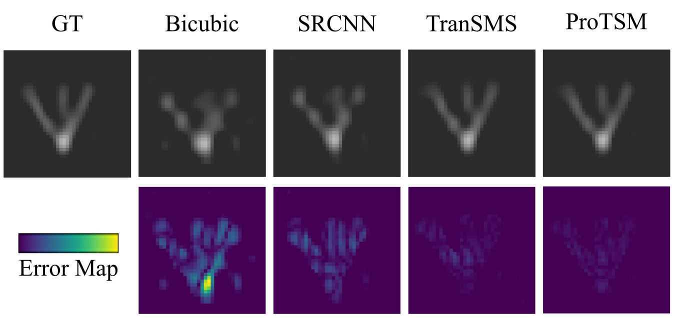

Fig. 6. The 2D image reconstruction results of four representative methods for resolution Phantom in OpenMPI dataset at ratio 4.

to handle 2D data. We first conducted the same experiment using the same dataset used in [36]. We directly referenced the study's results, and the 2D SM calibration comparison results are listed in Table V. ProTSM performs similarly to TranSMS for a small downsampling ratio of 2 and 4. ProTSM outperforms TranSMS for a high downsampling ratio. However, the SM calibration results of all methods are insufficient for a downsampling ratio of 8, which may mean that the metric nRMSE is insignificant.

Four representative methods—bicubic, SRCNN, TranSMS, and ProTSM—were selected for its validation, and another experiment (OpenMPI calibration 7 for training and calibration 6 for test) was conducted in 2D settings. Table VI and Fig. 6 show the results. For ratio 4, ProTSM and TranSMS continue to perform better in terms of SM calibration and image reconstruction. Although bicubic achieves a better metric nRMSE for SM recovery for ratio 8, the metrics of the reconstructed image are lower. All calibrated SMs fail to reconstruct a satisfactory image; therefore, metric nRMSE may not be able to assess the model's performance in such a scenario.

D. Application to In-House MPI Systems

We applied the proposed ProTSM to in-house MPI systems to improve the quality of the image reconstruction. We estimated high-resolution SM from a measured low-resolution SM, and reconstructed images using the measured SM and estimated high-resolution SMs. The corresponding results are shown in Fig. 7. The reconstructed images from two phantoms are shown. We measured the phantom resolution using two parallel cylindrical tubes filled with Perimag MNPs with 3 mm distance using the FFP scanner (top of Fig. 7) and the phantom vessel using the FFL scanner (bottom of Fig. 7). The boundaries of the reconstructed images appear mixed for the measured low-resolution SM, whereas the image reconstructed using the high-resolution SM shows better quality for phantom resolution. For the phantom vessel, the reconstructed image using the low-resolution SM does not distinguish the vascular bifurcation in the upper-right region, whereas the image generated by the calibrated SM clearly shows that structure. The evaluation results of ProTSM embedded in in-house FFP and FFL scanners validate our proposal.

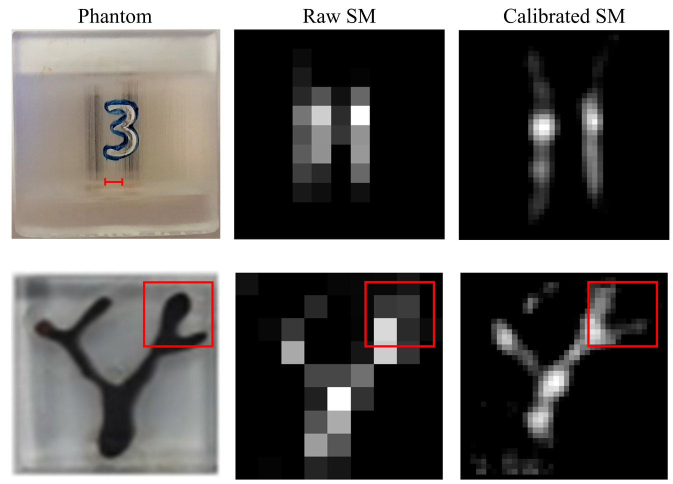

Fig. 7. The reconstructed image with the raw measured low-resolution SM and predicted high-resolution SM for two in-house MPI instruments. The first and second rows show the image reconstruction results of FFP (resolution phantom of two parallel cylindrical tubes with $3\,\mathrm{mm}$ distance) and FFL (vessel phantom) instruments, respectively.

E. Ablation Studies

We also investigated the impact of three main design components in the proposed ProTSM: pretraining strategy, modeling of coil information, and transformer architecture. Three ablation models [ProTSM-scratch (ProTSM without pretraining strategy), ProTSM-w/o coil information (ProTSM without coil information and pretraining strategy), and ProTSM-CNN (replace the transformer layer with equal number of convolution layer for ProTSM-w/o coil information)] were evaluated on the public OpenMPI dataset. In Section IV-B, the other experiment settings remain unchanged. Both SM calibration and image reconstruction tasks were conducted, and the corresponding results are shown in Tables VII and VIII and Fig. 8.

Regarding the pretraining strategy, the nRMSE values of ProTSM without pretraining (ProTSM-scratch) are 3.29% and 4.33% for downsampling ratios of 2 and 4, respectively. This demonstrates a performance decline of approximately 6%. ProTSM w/o coil information refers to ProTSM results that do not consider the coil channel and frequency index. The corresponding nRMSE metrics for downsampling ratios of 2 and 4 are 3.44% and 4.45%, respectively. Finally, to investigate the impact of the transformer, the encoder was replaced with a CNN. The performance is comparable to that of the CNN-based models (VolumeNet and 3dSMRnet) without the transformer. Therefore, super-resolution SM calibration benefits from the transformer, as discussed in [36].

This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2023.3297173

TABLE VII
THE ABLATION RESULTS IN OPENMPI DATASET FOR SM CALIBRATION. THE NUMBER INDICATES THE NRMSE METRIC.

Method	2×	4×
ProTSM	3.08%	4.10%
ProTSM-scratch	3.29%	4.33%
ProTSM-w/o coil information	3.44%	4.45%
ProTSM-CNN	3.75%	4.54%

TABLE VIII
THE ABLATION RESULTS IN OPENMPI DATASET FOR IMAGE RECONSTRUTION. THE DOWNSAMPLING RATIO IS 4.

Phantom	Resolution		Shape	
Method	PSNR	SSIM	PSNR	SSIM
ProTSM	31.57	0.7540	33.43	0.7376
ProTSM-scratch	29.82	0.6834	31.88	0.6593
ProTSM-w/o coil information	27.35	0.6665	31.11	0.6448
ProTSM-CNN	26.07	0.6584	30.59	0.6092

Additionally, in Fig. 8(a), we show the image reconstruction and error map results using the calibrated SMs for downsampling ratio 4. The ProTSM-scratch-reconstructed image contains more artifacts around it. Additionally, ProTSM without coil information generates a distorted image, and ProTSM-CNN shows low image quality.

We further highlight the effectiveness of the proposed pretraining strategy. The training loss and test nRMSE variations for ProTSM training with and without pretraining are shown in Fig. 8(b). Compared with training starting from scratch, finetuning provides a lower loss during training. Furthermore, the test nRMSE indicates that finetuning has better performance and stability. These results confirm the importance and contribution of the proposed pretraining strategy.

F. Visualization Results

To demonstrate an intuitive comprehension, this section visualizes hidden features from the transformer layer. Particularly, we averaged the feature maps using the token dimension after obtaining them through the final transformer layer. We used t-SNE to visualize the representations in Fig. 9(a). ProTSM-rand. init denotes the ProTSM model without training (i.e., with randomly initialized model parameters). The low-resolution SM rows are mixed distributed before training, and they are clustered closer together through the frequency index after training. This demonstrates that the calibration may help the low-resolution SM rows regain the coil-related properties.

Additionally, three examples of test set data demonstrate the impact of coil information. The performance of ProTSM-w/o coil information and ProTSM-scratch is compared, and the attention map is calculated using the frequency index and coil channel as seeds. The attention mask is averaged using the two tokens, and the top 25% activation areas are preserved. The results are shown in Fig. 9(b). The attention mask covers relatively important areas, and the coil information may help the ProTSM-scartch perform better. The above results show that the SM calibration task may benefit from the coil information.

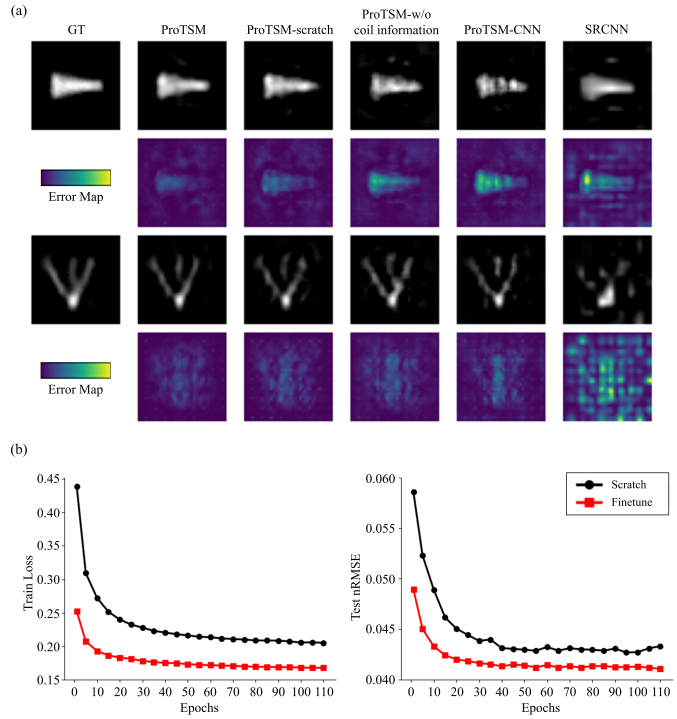

Fig. 8. (a) Reconstructed image based on SMs predicted by ProTSM variant models. (b) Variation in train loss and test nRMSE for finetune mode and train from scratch mode during training epochs.

VI. DISCUSSION

To accelerate the 3D SM calibration for MPI, we propose a transformer-based method to model the relationship between SM rows and a pretraining strategy to use unlabeled data. The estimated time of high-resolution SM in the OpenMPI dataset is shown in Table IX. The measurement time cost is estimated using [38]. Measurement and CS methods take a lot of time to recover SM. Interpolation-based methods have notably shorten the calibration time, while the quality of recovered SM is not satisfactory especially in high downsampling ratio. The deep learning-based approaches reduce the calibration time to the hundred-second level. Considering that the SM calibration is not required to be real-time, the proposed method, just like other deep learning-based approaches, has efficiently saved time and labor costs compared with the measurement. Moreover, in light of the quality of the recovered SM, our proposed method may also strike a more desirable balance between SM recovery prediction accuracy and calibration time.

Existing methods conceptualize SM calibration as a super-resolution task in natural images, but the calibration accuracy of the SM frequency components is higher than the reconstruction accuracy of natural images. Additionally, the spatial size of the SM rows ($32 \times 32 \times 32$) is significantly smaller than that of natural and medical images (e.g., $256 \times 256 \times 128$). In large images, the relationship between distant pixels is relatively weak, while the SM's compact size promotes stronger relationships between its elements. Considering the high level of accuracy required and the strong relationship between elements, SM calibration may benefit from modeling long-range dependencies than natural image reconstruction.

This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2023.3297173

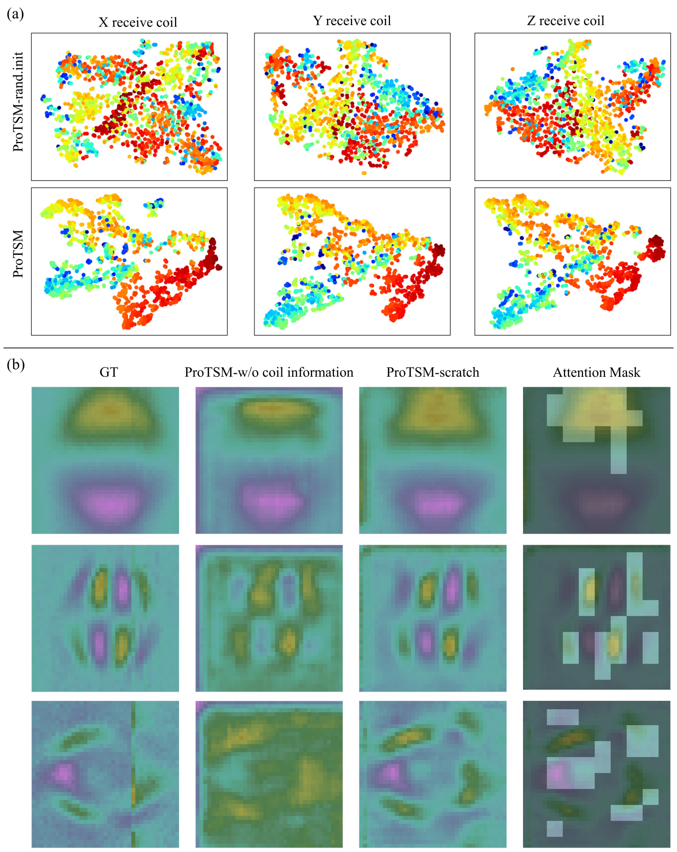

Fig. 9. (a) The t-SNE visualization of SM rows generated from the model. The color of the points represents the frequency index. ProTSM-rand. init indicates the ProTSM model without training. (b) Qualitative visualizations of the ProTSM-scratch and ProTSM-w/o coil information for three representative SM rows. The attention mask indicates the most attentive areas with the coil information as seed.

This may explain the notable contribution of transformer architecture to SM calibration.

To prevent overfitting owing to the high complexity of the transformer, we introduce a pretraining strategy that leverages low-resolution SM data. A low-resolution SM is easily collected during the development of an MPI system. We may measure the small SM repeatedly throughout system development to verify its performance. However, we should not measure the full-size SM because it is inaccurate after system upgrade. Hence, massive low-resolution SM data can be collected during the development process and used for SM calibration.

Despite the success of previous SM recovery studies [16]–[18], [36], they may have overlooked the potential benefit of the hardware information (e.g., coil information in this study). Numerous studies have shown the importance of multimodal data fusion learning [46], [47], e.g., non-image data in medical image analysis. However, the effectiveness of multimodal data (i.e., frequency index and coil channel) in the MPI area has not been evaluated. This study introduces previously overlooked hardware information and validates its effectiveness for SM recovery.

One limitation of our study is that the robustness of the proposed method has not been validated in vivo imaging. Several phantoms were imaged in vitro for image reconstruction task assessment, and we only assessed the performance using nRMSE, PSNR and SSIM. These metrics evaluate the overall quality of the reconstructed image, but may be insufficient in assessing the specific image details, especially in vivo imaging. Different nanoparticle behaviors have been observed between in vitro and in vivo settings because tracers' signals will change when they interact with biological tissue [48], [49]. Therefore, higher metric (PSNR and SSIM) may not guarantee better performance in vivo imaging especially for clinical applications. The solution to this problem remains an open debate. We intend to develop better metric to discuss the potential solution to this problem, and validate the effectiveness of our proposed method in vivo settings in our future research.

TABLE IX
THE COMPARISON OF ESTIMATED TIME (SECONDS) FOR HIGH-RESOLUTION SM IN OPENMPI DATASET.

Method	2×	4×
Measurement	124621.28	423971.82
Bicubic	0.81	0.51
Trilinear	0.75	0.50
CS	54984.28	19620.41
SRCNN	22.22	21.93
3dSMRnet	107.92	22.53
VolumeNet	7.83	6.19
IDL	30.26	17.90
MetaBlock	92.04	39.98
ProTSM	58.34	62.59

There are two future research directions to improve the current study:

1) Better utilization of multimodal information. We use the coil channel and frequency index for SM calibration, but the integrated method may not be optimal. Hence, multimodal information should be fully used to model the relationships between SM rows and improve the calibration accuracy. For example, graph convolutional networks [50], [51] may better model the relationships using graphs. Therefore, developing SM calibration using such networks may be a direction worth exploring.

2) More powerful pretraining strategy. We introduce a pseudo-label-based pretraining strategy to use available unlabeled data. A more enhanced pretraining strategy should be explored and analyzed. For example, more accurate and transferable pseudo-labels should be generated for different downstream datasets. Additionally, self-supervised pretraining has demonstrated its effectiveness on medical data [52], [53]. The fusion of such pretraining strategies may further improve the SM calibration.

VII. CONCLUSION

We proposed a transformer-based model for fast 3D SM calibration that uses multimodal information. Additionally, we proposed a pretraining strategy to fully use available unlabeled SM data. Our results on the OpenMPI and simulation datasets demonstrated that our ProTSM outperforms other methods. Moreover, the results for in-house MPI systems indicated the applicability and generalization ability of ProTSM.

This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2023.3297173

References

[1] N. Panagiotopoulos, R. L. Duschka, M. Ahlborg, G. Bringout, C. Debbeler, M. Graeser, C. Kaethner, K. Lüdtke-Buzug, H. Medimagh, J. Stelzner *et al.*, "Magnetic particle imaging: current developments and future directions," *International journal of nanomedicine*, vol. 10, p. 3097, 2015.

[2] T. Knopp and T. M. Buzug, *Magnetic particle imaging: an introduction to imaging principles and scanner instrumentation*. Springer Science & Business Media, 2012.

[3] J. Weizenecker, J. Borgert, and B. Gleich, "A simulation study on the resolution and sensitivity of magnetic particle imaging," *Physics in Medicine & Biology*, vol. 52, no. 21, p. 6363, 2007.

[4] T. Knopp, T. F. Sattel, S. Biederer, J. Rahmer, J. Weizenecker, B. Gleich, J. Borgert, and T. M. Buzug, "Model-based reconstruction for magnetic particle imaging," *IEEE Transactions on Medical Imaging*, vol. 29, no. 1, pp. 12–18, 2009.

[5] W. Tong, H. Hui, W. Shang, Y. Zhang, F. Tian, Q. Ma, X. Yang, J. Tian, and Y. Chen, "Highly sensitive magnetic particle imaging of vulnerable atherosclerotic plaque with active myeloperoxidase-targeted nanoparticles," *Theranostics*, vol. 11, no. 2, p. 506, 2021.

[6] G. Jia, L. Huang, Z. Wang, X. Liang, Y. Zhang, Y. Zhang, Q. Miao, K. Hu, T. Li, Y. Wang *et al.*, "Gradient-based pulsed excitation and relaxation encoding in magnetic particle imaging," *IEEE Transactions on Medical Imaging*, vol. 41, no. 12, pp. 3725–3733, 2022.

[7] X. Y. Zhou, Z. W. Tay, P. Chandrasekharan, Y. Y. Elaine, D. W. Hensley, R. Orendorff, K. E. Jeffris, D. Mai, B. Zheng, P. W. Goodwill *et al.*, "Magnetic particle imaging for radiation-free, sensitive and high-contrast vascular imaging and cell tracking," *Current opinion in chemical biology*, vol. 45, pp. 131–138, 2018.

[8] R. Kuo, P. Chandrasekharan, B. Fung, and S. Conolly, "In vivo therapeutic cell tracking using magnetic particle imaging," *International Journal on Magnetic Particle Imaging*, vol. 8, no. 1 Suppl 1, 2022.

[9] L. C. Wu, Y. Zhang, G. Steinberg, H. Qu, S. Huang, M. Cheng, T. Bliss, F. Du, J. Rao, G. Song *et al.*, "A review of magnetic particle imaging and perspectives on neuroimaging," *American Journal of Neuroradiology*, vol. 40, no. 2, pp. 206–212, 2019.

[10] C. Z. Cooley, J. B. Mandeville, E. E. Mason, E. T. Mandeville, and L. L. Wald, "Rodent cerebral blood volume (cbv) changes during hypercapnia observed using magnetic particle imaging (mpi) detection," *NeuroImage*, vol. 178, pp. 713–720, 2018.

[11] A. C. Bakenecker, M. Ahlborg, C. Debbeler, C. Kaethner, T. M. Buzug, and K. Lüdtke-Buzug, "Magnetic particle imaging in vascular medicine," *Innovative surgical sciences*, vol. 3, no. 3, pp. 179–192, 2018.

[12] M. Grüttner, T. Knopp, J. Franke, M. Heidenreich, J. Rahmer, A. Halkola, C. Kaethner, J. Borgert, and T. M. Buzug, "On the formulation of the image reconstruction problem in magnetic particle imaging," *Biomedizinische Technik/Biomedical Engineering*, vol. 58, no. 6, pp. 583–591, 2013.

[13] P. W. Goodwill and S. M. Conolly, "Multidimensional x-space magnetic particle imaging," *IEEE transactions on medical imaging*, vol. 30, no. 9, pp. 1581–1590, 2011.

[14] L. Yin, W. Li, Y. Du, K. Wang, Z. Liu, H. Hui, and J. Tian, "Recent developments of the reconstruction in magnetic particle imaging," *Visual computing for industry, biomedicine, and art*, vol. 5, no. 1, pp. 1–13, 2022.

[15] T. Knopp, N. Gdaniec, and M. Möddel, "Magnetic particle imaging: from proof of principle to preclinical applications," *Physics in Medicine & Biology*, vol. 62, no. 14, p. R124, 2017.

[16] A. von Gladiß, M. Ahlborg, T. Knopp, and T. M. Buzug, "Compressed sensing of the system matrix and sparse reconstruction of the particle concentration in magnetic particle imaging," *IEEE Transactions on Magnetics*, vol. 51, no. 2, pp. 1–4, 2015.

[17] M. Grosser, M. Möddel, and T. Knopp, "Using low-rank tensors for the recovery of mpi system matrices," *IEEE Transactions on Computational Imaging*, vol. 6, pp. 1389–1402, 2020.

[18] I. M. Baltruschat, P. Szwargulski, F. Griese, M. Grosser, R. Werner, and T. Knopp, "3d-smrnet: Achieving a new quality of mpi system matrix recovery by deep learning," in *MICCAI*. Springer, 2020, pp. 74–82.

[19] A. Güngör, B. Askin, D. A. Soydan, C. B. Top, and T. Cukur, "Deep learned super resolution of system matrices for magnetic particle imaging," in *EMBC*. IEEE, 2021, pp. 3749–3752.

[20] G. Holste, S. C. Partridge, H. Rahbar, D. Biswas, C. I. Lee, and A. M. Alessio, "End-to-end learning of fused image and non-image features for improved breast cancer classification from mri," in *ICCV*, 2021, pp. 3294–3303.

[21] C. Cui, H. Yang, Y. Wang, S. Zhao, Z. Asad, L. A. Coburn, K. T. Wilson, B. A. Landman, and Y. Huo, "Deep multi-modal fusion of image and non-image data in disease diagnosis and prognosis: a review," *arXiv preprint arXiv:2203.15588*, 2022.

[22] J. Huang, J. Tao, B. Liu, Z. Lian, and M. Niu, "Multimodal transformer fusion for continuous emotion recognition," in *ICASSP*. IEEE, 2020, pp. 3507–3511.

[23] P. Xu, X. Zhu, and D. A. Clifton, "Multimodal learning with transformers: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[24] A. Güngör and C. B. Top, "Super-resolving reconstruction technique for mpi," *International Journal on Magnetic Particle Imaging*, vol. 6, no. 2 Suppl 1, pp. 1–3, 2020.

[25] T. Knopp and A. Weber, "Sparse reconstruction of the magnetic particle imaging system matrix," *IEEE Transactions on Medical Imaging*, vol. 32, no. 8, pp. 1473–1480, 2013.

[26] M. Grosser, M. Boberg, M. Bahe, and T. Knopp, "Enhanced compressed sensing recovery of multi-patch system matrices in mpi," *International Journal on Magnetic Particle Imaging*, vol. 6, no. 2 Suppl 1, 2020.

[27] S. Ilbey, C. B. Top, A. Güngör, T. Çukur, E. U. Saritas, and H. E. Güven, "Fast system calibration with coded calibration scenes for magnetic particle imaging," *IEEE Transactions on Medical Imaging*, vol. 38, no. 9, pp. 2070–2080, 2019.

[28] M. Maass, M. Ahlborg, A. Bakenecker, F. Katzberg, H. Phan, T. M. Buzug, and A. Mertins, "A trajectory study for obtaining mpi system matrices in a compressed-sensing framework," *International Journal on Magnetic Particle Imaging*, vol. 3, no. 2, 2017.

[29] X. Wu, B. He, P. Gao, P. Zhang, Y. Shang, L. Zhang, J. Zhong, J. Jiang, H. Hui, and J. Tian, "Pgnet: Projection generative network for sparse-view reconstruction of projection-based magnetic particle imaging," *Medical Physics*, 2022.

[30] Y. Shang, J. Liu, L. Zhang, X. Wu, P. Zhang, L. Yin, H. Hui, and J. Tian, "Deep learning for improving the spatial resolution of magnetic particle imaging," *Physics in Medicine & Biology*, 2022.

[31] F. Schrank, D. Pantke, and V. Schulz, "Deep learning mpi super-resolution by implicit representation of the system matrix," *International Journal on Magnetic Particle Imaging*, vol. 8, no. 1 Suppl 1, 2022.

[32] L. Yin, H. Guo, P. Zhang, Y. Li, H. Hui, Y. Du, and J. Tian, "System matrix recovery based on deep image prior in magnetic particle imaging," *Physics in Medicine & Biology*, 2022.

[33] A. Güngör, B. Askin, D. A. Soydan, C. B. Top, E. U. Saritas, and T. Çukur, "Deq-mpi: A deep equilibrium reconstruction with learned consistency for magnetic particle imaging," *arXiv preprint arXiv:2212.13233*, 2022.

[34] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[35] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *ICCV*, 2021, pp. 10 012–10 022.

[36] A. Güngör, B. Askin, D. A. Soydan, E. U. Saritas, C. B. Top, and T. Çukur, "Transms: Transformers for super-resolution calibration in magnetic particle imaging," *IEEE Transactions on Medical Imaging*, 2022.

[37] G. Shi, L. Zhang, H. Hui, and J. Tian, "3d system matrix calibration by using coil information and transformer," *International Journal on Magnetic Particle Imaging IJMPI*, vol. 9, no. 1 Suppl 1, 2023.

[38] T. Knopp, P. Szwargulski, F. Griese, and M. Gräser, "Openmpidata: An initiative for freely accessible magnetic particle imaging data," *Data in brief*, vol. 28, p. 104971, 2020.

[39] Y. Shen, C. Hu, P. Zhang, J. Tian, and H. Hui, "A novel software framework for magnetic particle imaging reconstruction," *International Journal of Imaging Systems and Technology*, 2022.

[40] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 2, pp. 295–307, 2015.

[41] J. M. Kasson, S. I. Nin, W. Plouffe, and J. L. Hafner, "Performing color space conversions with three-dimensional linear interpolation," *Journal of Electronic Imaging*, vol. 4, no. 3, pp. 226–250, 1995.

[42] Y. Li, Y. Iwamoto, L. Lin, R. Xu, R. Tong, and Y.-W. Chen, "Volumenet: a lightweight parallel network for super-resolution of mr and ct volumetric data," *IEEE Transactions on Image Processing*, vol. 30, pp. 4840–4854, 2021.

[43] A. G. Pacheco and R. A. Krohling, "An attention-based mechanism to combine images and metadata in deep learning models applied

to skin cancer classification," *IEEE journal of biomedical and health informatics*, vol. 25, no. 9, pp. 3554–3563, 2021.

[44] H. Duanmu, P. B. Huang, S. Brahmavar, S. Lin, T. Ren, J. Kong, F. Wang, and T. Q. Duong, "Prediction of pathological complete response to neoadjuvant chemotherapy in breast cancer using deep learning with integrative imaging, molecular and demographic data," in *MICCA*. Springer, 2020, pp. 242–252.

[45] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *CVPR*, 2016, pp. 1646–1654.

[46] P. Rajpurkar, E. Chen, O. Banerjee, and E. J. Topol, "Ai in health and medicine," *Nature medicine*, vol. 28, no. 1, pp. 31–38, 2022.

[47] S. Parisot, S. I. Ktena, E. Ferrante, M. Lee, R. Guerrero, B. Glocker, and D. Rueckert, "Disease prediction using graph convolutional networks: application to autism spectrum disorder and alzheimer's disease," *Medical image analysis*, vol. 48, pp. 117–130, 2018.

[48] M. G. Kaul, T. Mummert, C. Jung, J. Salamon, A. P. Khandhar, R. M. Ferguson, S. J. Kemp, H. Ittrich, K. M. Krishnan, G. Adam *et al.*, "In vitro and in vivo comparison of a tailored magnetic particle imaging blood pool tracer with resovist," *Physics in Medicine & Biology*, vol. 62, no. 9, p. 3454, 2017.

[49] J. Dieckhoff, M. Kaul, T. Mummert, C. Jung, J. Salamon, G. Adam, T. Knopp, F. Ludwig, C. Balceris, and H. Ittrich, "In vivo liver visualizations with magnetic particle imaging based on the calibration measurement approach," *Physics in Medicine & Biology*, vol. 62, no. 9, p. 3470, 2017.

[50] G. Shi, Y. Zhu, J. K. Liu, and X. Li, "Hegcl: Advance self-supervised learning in heterogeneous graph-level representation," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–12, 2023.

[51] G. Shi, Y. Zhu, Z. Chen, J. Liu, and X. Li, "Are non-image data really necessary for disease prediction with graph convolutional networks?" *IEEE Transactions on Cognitive and Developmental Systems*, 2022.

[52] R. Krishnan, P. Rajpurkar, and E. J. Topol, "Self-supervised learning in medicine and healthcare," *Nature Biomedical Engineering*, pp. 1–7, 2022.

[53] L. Chen, P. Bentley, K. Mori, K. Misawa, M. Fujiwara, and D. Rueckert, "Self-supervised learning for medical image analysis using image context restoration," *Medical image analysis*, vol. 58, p. 101539, 2019.