D^{2}MAE: Diffusional Deblurring MAE for Ultrasound Image Pre-training

Qingbo Kang[1], Jun Gao[1,4], Hongkai Zhao[2], Zhu He[2], Kang Li[1,3] ⋆, and Qicheng Lao[2] ⋆⋆

[1] West China Biomedical Big Data Center, West China Hospital, Sichuan University
[2] School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing, China
[3] Shanghai Artificial Intelligence Laboratory, Shanghai, China
[4] Stork Healthcare, Chengdu, China

Abstract. Recent advances in generative self-supervised learning, particularly Masked Autoencoders (MAE), have shown significant promise in medical image pre-training. However, ultrasound poses unique challenges due to its intrinsic low signal-to-noise ratio. While previous studies have enhanced MAE with deblurring for improved performance, their static deblurring strategy fails to consider domain discrepancies arising from variations in ultrasound imaging. To overcome these limitations, we propose D^{2}MAE—a **D**iffusional **D**eblurring-enhanced **MAE** framework that seamlessly integrates a diffusional deblurring objective into MAE pre-training, simultaneously optimizing both deblurring and masked image reconstruction within a unified framework. Furthermore, we introduce an optimal blurriness-aware fine-tuning strategy that dynamically adjusts blurriness through an optimal blurriness search procedure, effectively accommodating the inherent domain discrepancies in ultrasound images. Extensive experiments across multiple ultrasound datasets, including thyroid, pancreas, and ovary, demonstrate that D^{2}MAE outperforms state-of-the-art methods, significantly enhancing generalizability and diagnostic performance across diverse ultrasound tasks. Our results establish D^{2}MAE as a superior approach for ultrasound imaging pre-training, paving the way for improved ultrasound image analysis. The code and pre-trained models are publicly available on GitHub.

Keywords: Self-Supervised Learning · Masked Autoencoders · Diffusional Deblurring · Ultrasound Pre-training.

1 Introduction

Recent advances in generative self-supervised learning (SSL), particularly Masked Autoencoders (MAE) [10], have demonstrated significant potential in medical

⋆ Corresponding author: likang@wchscu.cn
⋆⋆ Corresponding author: qicheng.lao@bupt.edu.cn

imaging pre-training [9,11]. By reconstructing masked regions from visible contexts, methods like MAE have achieved remarkable success across various medical imaging modalities [29,15,4,14,22,16], enabling the extraction of generalized features without manual annotations [2,10,26].

Ultrasound imaging, a cornerstone of medical diagnostics, poses unique challenges due to its inherent low signal-to-noise ratio [1]. While prior work [15] has enhanced MAE with deblurring to yield promising performance in thyroid ultrasound, their static and inflexible deblurring strategy–i.e., assuming uniform blurriness across all ultrasound images–fails to account for domain discrepancies arising from heterogeneous imaging characteristics across ultrasound devices [19,13,8], scanning protocols [8], and post-processing algorithms [7], hereby limiting diagnostic generalizability across clinical settings [19,7,13,8].

Concurrently, denoising diffusion models, exemplified by denoising diffusion probabilistic models (DDPMs) [12], have emerged as a dominant paradigm in image generation through progressive noise addition and iterative denoising [5]. Although originally designed for image synthesis, diffusion models share a common generative foundation with MAE, and recent studies have begun to explore their integration with generative SSL paradigm [24,25]. For instance, Wei *et al.* [24] introduced DiffMAE, which conditions diffusion models on masked inputs to enhance representation learning, showing benefits for natural image pre-training. However, this fusion remains largely unexplored for medical imaging.

To address these challenges, we propose D^{2}MAE—a **D**iffusional **D**eblurring-enhanced **MAE** framework that seamlessly integrates the strengths of diffusion models and MAE, specifically curated for ultrasound image pre-training. Unlike prior methods such as DeblurrMAE [15] and DiffMAE [24], D^{2}MAE embeds a diffusional deblurring objective into MAE pre-training, enabling joint optimization of progressive deblurring and masked image reconstruction through a unified learning framework. Furthermore, we introduce an optimal blurriness-aware fine-tuning strategy that employs an optimal blurriness search (OBS) procedure to automatically adjust the blurriness during downstream fine-tuning, aligning it with the unique imaging characteristics and anatomical variations in ultrasound images. Our contributions are summarized as follows:

1. We propose D^{2}MAE, a novel generative SSL framework that integrates a diffusional deblurring process with MAE to jointly optimize progressive deblurring and masked image reconstruction.
2. We introduce an optimal blurriness-aware fine-tuning strategy, which employs an optimal blurriness search (OBS) to dynamically select the most appropriate deblurring level for downstream tasks.
3. We demonstrate that D^{2}MAE significantly outperforms state-of-the-art methods in ultrasound tasks across multiple anatomical organs—including thyroid, pancreas, and ovary—with extensive experiments validating its superiority as a pre-training approach for ultrasound images.

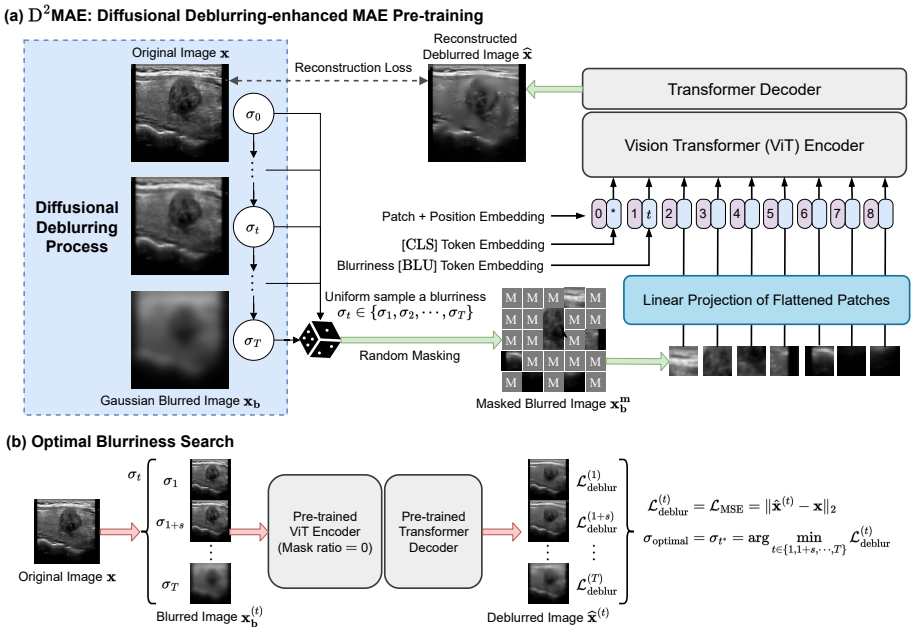

Fig. 1: Overview of our proposed D²MAE. (a) D²MAE pre-training integrates a diffusional deblurring process with MAE's mask-reconstruction pre-training [10]. (b) D²MAE fine-tuning incorporates an optimal blurriness search procedure prior to standard MAE fine-tuning, enabling automatic blurriness adjustment.

2 Method

2.1 Overview

We propose D²MAE, a novel SSL framework that seamlessly integrates a diffusional deblurring process with MAE. As illustrated in Figure 1, our approach comprises two primary components: (a) diffusional deblurring-enhanced MAE pre-training, which simultaneously optimizes progressive deblurring and masked image reconstruction in a unified framework, and (b) optimal blurriness-aware fine-tuning, wherein an OBS procedure is performed prior to standard MAE fine-tuning. The following sections detail these components.

2.2 D²MAE: Pre-training

During the pre-training, D²MAE jointly optimizes progressive deblurring and masked image reconstruction. As depicted in Figure 1(a), given an original ultrasound image $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$, a blurriness level σ_t is randomly sampled from a predefined set $\{\sigma_1, \sigma_2, \cdots, \sigma_T\}$ (with $\sigma_1 < \sigma_2 < \cdots < \sigma_T$ and a fixed interval, we adopt a fixed-step σ schedule to ensure full blur coverage, enable efficient

search, and reduce variance during training), where the time step t being uniformly selected from $\{1, \ldots, T\}$, i.e., $t \sim \mathcal{U}\{1, \ldots, T\}$. A Gaussian blur operation is then applied to $\mathbf{x}$:

$$\mathbf{x_b} = \mathcal{G}_{\sigma_t}(\mathbf{x}), \tag{1}$$

where $\mathcal{G}_{\sigma_t}$ denotes Gaussian blurring with standard deviation σ_t. As shown in Figure 1(a), σ_0 corresponds to the original image (no blur), while σ_T represents the maximum blurriness. Following MAE [10], we randomly mask 75% of the patches in blurred image $\mathbf{x_b}$, yielding the masked input $\mathbf{x_b^m}$.

To explicitly encode the blurriness step information, we employ sinusoidal blur embeddings inspired by DDPM's timestep encoding [12]. A blurriness token $\mathrm{BLU}(t) \in \mathbb{R}^D$, is introduced and inserted after the CLS token, forming the input sequence:

$$\mathbf{z}_0 = [\mathrm{CLS},\ \mathrm{BLU}(t),\ \mathbf{E}_1, \ldots, \mathbf{E}_N] + \mathbf{P}, \tag{2}$$

where $\mathbf{E}_i \in \mathbb{R}^D$ denotes the embeddings of unmasked patches from $\mathbf{x_b^m}$, and $\mathbf{P} \in \mathbb{R}^{(N+2)\times D}$ represents the position embeddings. The blur token is a 768-dimensional vector, adding only 0.0009% parameters to the ViT-Base model. This design enables the cross-attention mechanism to dynamically modulate image representations based on the blurriness.

The input sequence $\mathbf{z}_0$ is then processed by the MAE architecture, which comprises a Vision Transformer (ViT) encoder [6] and a Transformer decoder [23]. The decoder reconstructs the deblurred image $\hat{\mathbf{x}}$, and the pre-training loss is computed as the mean squared error between $\hat{\mathbf{x}}$ and the original image $\mathbf{x}$:

$$\mathcal{L}_{pretrain} = \|\hat{\mathbf{x}} - \mathbf{x}\|_2. \tag{3}$$

2.3 D^{2}MAE: Downstream Fine-tuning

The overall downstream fine-tuning pipeline of D^{2}MAE is summarized in Algorithm 1. Unlike standard MAE fine-tuning, which directly fine-tunes the pre-trained encoder on downstream tasks, our approach incorporates an OBS procedure prior to fine-tuning, which aims to accommodate the varying degrees of blurriness in ultrasound images and select the most suitable level of deblurring. It is important to note that the OBS procedure is executed only once (not at every fine-tuning epoch), resulting in negligible computational overhead.

As illustrated in Figure 1(b), for each image $\mathbf{x}$, we first generate a set of blurred variants $\{\mathbf{x}_b^{(t)}\}_{t=1,\,1+s,\ldots,T}$ by applying a Gaussian blur with varying standard deviations σ_t, where $\mathbf{x}_b^{(t)} = \mathcal{G}_{\sigma_t}(\mathbf{x})$ and s denotes the search step. The pre-trained model P_θ processes each blurred image to yield a deblurred output:

$$\hat{\mathbf{x}}^{(t)} = P_\theta(\mathbf{x}_b^{(t)}, t, \text{mask ratio} = 0), \tag{4}$$

and the deblurring loss is computed as the mean squared error between $\hat{\mathbf{x}}^{(t)}$ and the original image $\mathbf{x}$. We then select the optimal blur level by identifying

$$t^* = \arg\min_{t \in \{1,\,1+s,\,\ldots,\,T\}} \|\hat{\mathbf{x}}^{(t)} - \mathbf{x}\|_2, \quad \sigma_{\text{optimal}} = \sigma_{t^*}. \tag{5}$$

Algorithm 1 D^{2}MAE: Downstream Fine-tuning

Input: Pre-trained model P_θ, fine-tuning model f_θ, dataset $\{\mathbf{X}, \mathbf{Y}\}$, number of classes C, optimal blurriness search step s
for each epoch **do**
  Sample a batch $\{\mathbf{x}, \mathbf{y}\}$ from $\{\mathbf{X}, \mathbf{Y}\}$
  for $t = 1$ **to** T, **step** s **do** ▷ Optimal Blurriness Search (OBS)
    Apply Gaussian blur: $\mathbf{x_b}^{(t)} \leftarrow \mathcal{G}_{\sigma_t}(\mathbf{x})$
    Generate deblurred image: $\hat{\mathbf{x}}^{(t)} \leftarrow P_\theta(\mathbf{x_b}^{(t)}, t, \text{mask ratio} = 0)$
    Compute loss: $\mathcal{L}_{\text{deblur}}^{(t)} \leftarrow \|\hat{\mathbf{x}}^{(t)} - \mathbf{x}\|_2$
  end for
  Select optimal level: $\sigma_{\text{optimal}} = \sigma_{t^*}$, where
    $t^* = \arg\min_{t \in \{1, 1+s, \ldots, T\}} \mathcal{L}_{\text{deblur}}^{(t)}$ ▷ Optimal blurriness level
  Apply optimal blurriness: $\mathbf{x_b} \leftarrow \mathcal{G}_{\sigma_{t^*}}(\mathbf{x})$
  Predict label probabilities: $\hat{\mathbf{y}} \leftarrow f_\theta(\mathbf{x_b}, t^*)$
  Compute cross-entropy loss: $\mathcal{L}_{ft} \leftarrow -\sum_{c=1}^{C} \mathbf{y}_c \log \hat{\mathbf{y}}_c$
  Update f_θ with $\mathcal{L}_{ft}$
end for

Finally, the corresponding blurred images $\mathbf{x}_b^{(t^*)}$ are utilized as input for downstream fine-tuning of the supervised model f_θ.

3 Experiments and Results

3.1 Experimental Settings

Data To evaluate the effectiveness of our proposed D^{2}MAE, we conducted comprehensive experiments on three distinct organs: thyroid, pancreas, and ovary. Table 1 summarizes the datasets used for both pre-training and downstream tasks, including data sources, image counts, and relevant characteristics. For each organ, a representative downstream task was selected: (1) thyroid nodule diagnosis using the GE4K dataset [15], which categorizes nodules as benign or malignant; (2) pancreatic cancer diagnosis using the LEPset dataset [17]; and (3) ovarian tumor diagnosis using the MMOTU dataset [27], which comprises eight diagnostic categories. Dataset splits followed established protocols: a 3:1:1 train/validation/test ratio for GE4K [15], stratified five-fold cross-validation for LEPset [17], and the original splitting protocol for MMOTU [27]. All pre-training and fine-tuning datasets are strictly disjoint, ensuring genuine transfer evaluation without data leakage.

Implementation Details For D^{2}MAE pre-training, the blurriness set is defined as $0.1, 0.2, \cdots, 1.1$, and the OBS step s is set to 2. We implement our approach in PyTorch, using a batch size of 1024 for pre-training and 32 for downstream transfer. Given the limited pre-training data, the number of pre-training epochs was set to 8000, 6000, and 16000 for thyroid, pancreas, and

Table 1: Summary of pre-training and downstream datasets.

Organ	Pre-training			Downstream		
	Source	Images	Total	Dataset	Classes	Images
Thyroid	GE4K [15]	10,675	10,675	GE4K [15]	2	4,494
Pancreas	RadImageNet [21]	21,639	29,639	LEPset [17]	2	3,500
	LEPset [17]	8,000				
Ovary	RadImageNet [21]	3,593	3,593	MMOTU [27]	8	1,469

Table 2: Performance comparison across three datasets using F1 (%). Methods are categorized into supervised pre-training (SP) and self-supervised pre-training (SSP). The best results are bolded (statistical significance: $P < 0.05$), with the second-best are underscored. 'US' denotes pre-training with ultrasound images.

Category	Method	Architecture	Pre-training	GE4K	LEPset	MMOTU
SP	ViT [6]	ViT-B	ImageNet	84.17±0.98	84.34±1.29	68.02±2.56
	Swin Transformer [18]	Swin-L	ImageNet	84.92±0.82	82.66±1.88	68.58±2.02
	ConvNeXt [20]	ConvNeXt-L	ImageNet	85.47±0.91	83.97±1.66	69.06±1.98
	Zhou *et al.* [28]	-	-	86.09±0.74	-	-
SSP	MoCo v3 [3]	ViT-B	ImageNet	84.48±1.12	81.64±1.28	68.87±1.74
		ViT-B	US	84.55±1.04	83.38±1.01	71.68±1.62
	MAE [10]	ViT-B	ImageNet	85.23±0.57	81.71±0.83	68.58±1.96
		ViT-B	US	87.54±0.62	<u>84.83±1.12</u>	71.58±1.81
	USFM [14]	ViT-B	US3M [14]	<u>88.87±0.56</u>	84.02±1.28	70.94±1.54
	DSMT-Net [17]	ViT-B	US	-	82.20±0.90	-
	DiffMAE [24]	ViT-B	US	85.59±0.95	83.43±1.37	70.47±1.79
	DeblurrMAE [15]	ViT-B	US	88.48±0.50	84.61±0.95	<u>72.33±1.96</u>
	D^2MAE (Ours)	ViT-B	US	**89.84±0.27**	**87.41±0.78**	**75.70±1.16**
P-value (best vs. second best)				0.007	<0.001	<0.001

ovary, respectively. All other pre-training and fine-tuning settings followed those described in MAE [10]. The pre-training was conducted on 8 Nvidia V100 GPUs, while downstream fine-tuning was performed on a single V100 GPU. Downstream performance is evaluated using accuracy (ACC), F1-score (F1), and area under the receiver operating characteristic curve (AUROC). The reported F1 is the macro-average of class-wise F1 scores. Error bars denote 95% confidence intervals (95% CI), and statistical significance is assessed via two-sided t-tests.

3.2 Results

Table 2 and Figure 2 present a comprehensive evaluation of downstream classification performance across three ultrasound datasets. We compare D^2MAE against several methods based on two paradigms: supervised pre-training (SP) and self-supervised pre-training (SSP). Notably, we include two MAE variants relevant to our work: DiffMAE [24], which integrates diffusion models with MAE, and DeblurrMAE [15], which enhances MAE with static deblurring.

As shown in Table 2, D^2MAE consistently achieves superior F1 scores across all three datasets, outperforming the second-best methods with statistical signifi-

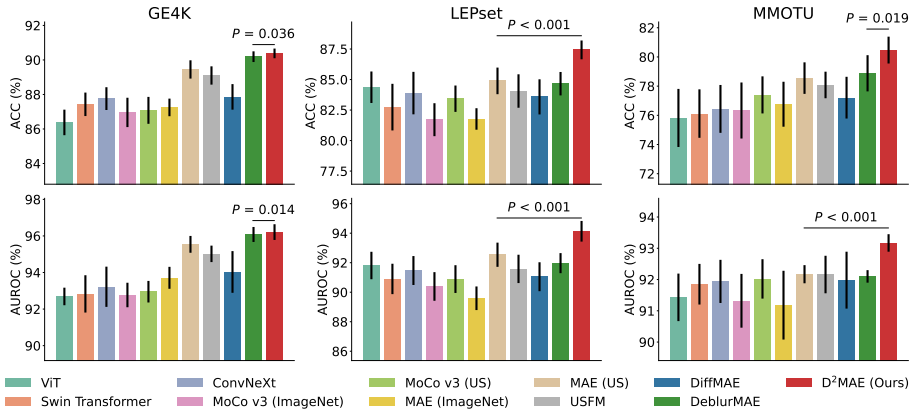

Fig. 2: Performance comparison using ACC (%) and AUROC (%).

Table 3: Ablation study results

(a) Blurriness range

Range	F1 (%)
$0.1 \rightarrow 0.6$	89.68
$0.1 \rightarrow 1.1$	**89.84**
$0.1 \rightarrow 1.6$	89.43

(b) Blurriness embedding

Embedding	F1 (%)
w/o embedding	87.66
Addition	88.57
Extra token	**89.84**

(c) OBS settings

Setting	F1 (%)
No blur	87.95
Fixed($\sigma = 0.5$)	88.36
With OBS	**89.84**

(d) Search step

Step s	F1 (%)
$s = 1$	**89.94**
$s = 2$	89.84
$s = 3$	89.05

cance (all $P < 0.05$). Specifically, D^{2}MAE attains a state-of-the-art F1 of 89.84% for thyroid ultrasound on GE4K (improving by 0.97%), 87.41% for pancreas ultrasound on LEPset (improving by 2.58%), and 75.70% for ovarian ultrasound on MMOTU (improving by 3.37%).

Furthermore, Figure 2 demonstrates that D^{2}MAE outperforms competing methods in terms of both ACC and AUROC. Specifically, D^{2}MAE achieves ACCs of 90.38% on GE4K, 87.43% on LEPset, and 80.47% on MMOTU, along with AUROCs of 96.21% on GE4K, 94.13% on LEPset, and 93.17% on MMOTU. All improvements are statistically significant (all $P < 0.05$).

Collectively, these results underscore the superiority and generalizability of our diffusional deblurring-enhanced MAE pre-training approach across diverse anatomical organs, affirming the efficacy of our targeted ultrasound image pre-training strategy.

3.3 Ablation Studies

Table 3 presents the ablation results, evaluating key design choices in D^{2}MAE across four factors:

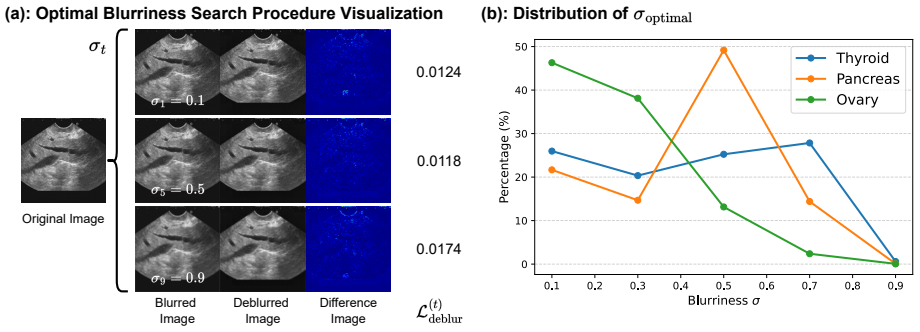

Fig. 3: (a) Visualization of the optimal blurriness search procedure. The original image is sourced from the pancreas LEPset [17]. The 'Difference Image' represents the absolute difference between the deblurred image and the original image. In this example, the optimal blurring level is 0.5, corresponding to the minimum deblurring loss $\mathcal{L}^{(t)}_{\text{deblur}}$. (b) Distributions of σ_{optimal} across three distinct organs.

(a) Blurriness range: We experiment with different ranges during pre-training: $0.1 \rightarrow 0.6$, $0.1 \rightarrow 1.1$, and $0.1 \rightarrow 1.6$, each with a fixed interval of 0.1. The range $0.1 \rightarrow 1.1$ achieves the best performance.
(b) Blurriness embedding: We test three strategies for incorporating blurriness information: (i) without embedding, (ii) adding the blurriness embedding to the image data, and (iii) treating the blurriness embedding as an extra token. The extra token strategy performs best.
(c) OBS settings: We compare three settings before fine-tuning: no blur, a fixed blurriness level ($\sigma = 0.5$), and the full OBS procedure. The full OBS configuration achieves the highest performance.
(d) Search step (s): The OBS process uses a search step to reduce computational overhead. While $s = 1$ gives the highest F1 score, the performance drop with $s = 2$ is marginal, and $s = 2$ provides efficiency gains. We adopt $s = 2$ in our final implementation.

Overall, our ablation results demonstrate that a blurriness range of $0.1 \rightarrow 1.1$, the extra token strategy for blurriness embedding, and the full OBS procedure with a search step of $s = 2$ together yield the best fine-tuning performance. These findings validate our design choices and underscore the effectiveness of the proposed D^{2}MAE framework for ultrasound image pre-training.

3.4 Visualizations

Figure 3(a) illustrates the OBS procedure employed during the optimal blurriness-aware fine-tuning of D^{2}MAE (see Figure 1(b)). This procedure searches for the deblurring level most suitable for fine-tuning, specifically, the blurring level that minimizes the deblurring loss, denoted as $\mathcal{L}^{(t)}_{\text{deblur}}$. In this example of pancreas

ultrasound image, the selected optimal blurring level is 0.5 ($\sigma_{\text{optimal}} = 0.5$). For simplicity, the search step is set to $s = 4$ in this visualization.

Figure 3(b) shows the distributions of σ_{optimal} obtained through the OBS procedure across three distinct organs. Due to inherent domain discrepancies in ultrasound imaging, each organ exhibits a unique optimal blurriness distribution, emphasizing the need for our dynamic blurriness adjustment approach.

4 Conclusion

In this paper, we introduce D^{2}MAE, a novel self-supervised pre-training framework that integrates a diffusional deblurring process with masked image reconstruction paradigm of MAE for ultrasound image pre-training. Unlike prior methods such as DeblurrMAE and DiffMAE, D^{2}MAE integrates a progressive, modality-specific deblurring process with a unified semantic learning objective tailored to ultrasound. By adaptively adjusting the deblurring level during both pre-training and fine-tuning, our approach mitigates the challenges posed by domain discrepancies in ultrasound imaging. Comprehensive experiments conducted on datasets from three distinct organs: thyroid, pancreas, and ovary demonstrate that D^{2}MAE outperforms current state-of-the-art methods in terms of F1, accuracy, and AUROC, highlighting its superior generalizability. These findings highlight the potential of integrating diffusional deblurring with MAE to enhance ultrasound image pre-training.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Avola, D., Cinque, L., Fagioli, A., Foresti, G., Mecca, A.: Ultrasound medical imaging techniques: a survey. ACM Computing Surveys (CSUR) **54**(3), 1–38 (2021)
2. Bao, H., Dong, L., Piao, S., Wei, F.: Beit: Bert pre-training of image transformers. arXiv preprint arXiv:2106.08254 (2021)
3. Chen, X., Xie, S., He, K.: An empirical study of training self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9640–9649 (2021)
4. Chen, Z., Agarwal, D., Aggarwal, K., Safta, W., Balan, M.M., Brown, K.: Masked image modeling advances 3d medical image analysis. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1970–1980 (2023)
5. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021)
6. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)

7. Gao, J., Lao, Q., Kang, Q., Liu, P., Zhang, L., Li, K.: Unsupervised cross-disease domain adaptation by lesion scale matching. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 660–670. Springer (2022)
8. Gao, J., Lao, Q., Liu, P., Yi, H., Kang, Q., Jiang, Z., Wu, X., Li, K., Chen, Y., Zhang, L.: Anatomically guided cross-domain repair and screening for ultrasound fetal biometry. IEEE Journal of Biomedical and Health Informatics **27**(10), 4914–4925 (2023)
9. Gui, J., Chen, T., Zhang, J., Cao, Q., Sun, Z., Luo, H., Tao, D.: A survey on self-supervised learning: Algorithms, applications, and future trends. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)
10. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16000–16009 (2022)
11. He, Y., Huang, F., Jiang, X., Nie, Y., Wang, M., Wang, J., Chen, H.: Foundation model for advancing healthcare: Challenges, opportunities, and future directions. arXiv preprint arXiv:2404.03264 (2024)
12. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
13. Huang, Y., Yang, X., Huang, X., Zhou, X., Chi, H., Dou, H., Hu, X., Wang, J., Deng, X., Ni, D.: Fourier test-time adaptation with multi-level consistency for robust classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (2023)
14. Jiao, J., Zhou, J., Li, X., Xia, M., Huang, Y., Huang, L., Wang, N., Zhang, X., Zhou, S., Wang, Y., et al.: Usfm: A universal ultrasound foundation model generalized to tasks and organs towards label efficient image analysis. Medical Image Analysis **96**, 103202 (2024)
15. Kang, Q., Gao, J., Li, K., Lao, Q.: Deblurring masked autoencoder is better recipe for ultrasound image recognition. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 352–362. Springer (2023)
16. Kang, Q., Lao, Q., Gao, J., Liu, J., Yi, H., Ma, B., Zhang, X., Li, K.: Deblurring masked image modeling for ultrasound image analysis. Medical Image Analysis p. 103256 (2024)
17. Li, J., Zhang, P., Wang, T., Zhu, L., Liu, R., Yang, X., Wang, K., Shen, D., Sheng, B.: Dsmt-net: Dual self-supervised multi-operator transformation for multi-source endoscopic ultrasound diagnosis. IEEE Transactions on Medical Imaging **43**(1), 64–75 (2023)
18. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
19. Liu, Z., Huang, X., Yang, X., Gao, R., Li, R., Zhang, Y., Huang, Y., Zhou, G., Xiong, Y., Frangi, A.F., Ni, D.: Generalize ultrasound image segmentation via instant and plug & play style transfer. In: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI). pp. 419–423 (2021)
20. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11976–11986 (2022)
21. Mei, X., Liu, Z., Robson, P.M., Marinelli, B., Huang, M., Doshi, A., Jacobi, A., Cao, C., Link, K.E., Yang, T., et al.: Radimagenet: an open radiologic deep learning

research dataset for effective transfer learning. Radiology: Artificial Intelligence **4**(5), e210315 (2022)
22. Quan, H., Li, X., Chen, W., Bai, Q., Zou, M., Yang, R., Zheng, T., Qi, R., Gao, X., Cui, X.: Global contrast-masked autoencoders are powerful pathological representation learners. Pattern Recognition **156**, 110745 (2024)
23. Vaswani, A.: Attention is all you need. Advances in Neural Information Processing Systems (2017)
24. Wei, C., Mangalam, K., Huang, P.Y., Li, Y., Fan, H., Xu, H., Wang, H., Xie, C., Yuille, A., Feichtenhofer, C.: Diffusion models as masked autoencoders. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16284–16294 (2023)
25. Xiang, W., Yang, H., Huang, D., Wang, Y.: Denoising diffusion autoencoders are unified self-supervised learners. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15802–15812 (2023)
26. Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., Hu, H.: Simmim: A simple framework for masked image modeling. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9653–9663 (2022)
27. Zhao, Q., Lyu, S., Bai, W., Cai, L., Liu, B., Wu, M., Sang, X., Yang, M., Chen, L.: A multi-modality ovarian tumor ultrasound image dataset for unsupervised cross-domain semantic segmentation. arXiv preprint arXiv:2207.06799 (2022)
28. Zhou, Y., Chen, H., Li, Y., Liu, Q., Xu, X., Wang, S., Yap, P.T., Shen, D.: Multi-task learning for segmentation and classification of tumors in 3d automated breast ultrasound images. Medical Image Analysis **70**, 101918 (2021)
29. Zhou, Y., Chia, M.A., Wagner, S.K., Ayhan, M.S., Williamson, D.J., Struyven, R.R., Liu, T., Xu, M., Lozano, M.G., Woodward-Court, P., et al.: A foundation model for generalizable disease detection from retinal images. Nature **622**(7981), 156–163 (2023)