



# Deblurring Masked Autoencoder Is Better Recipe for Ultrasound Image Recognition

Qingbo Kang<sup>1,3</sup>, Jun Gao<sup>1,4</sup>, Kang Li<sup>1,3(✉)</sup>, and Qicheng Lao<sup>2,3(✉)</sup>

<sup>1</sup> West China Biomedical Big Data Center, West China Hospital, Sichuan University, Chengdu, China

[likang@wchscu.cn](mailto:likang@wchscu.cn)

<sup>2</sup> School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing, China

[qicheng.lao@bupt.edu.cn](mailto:qicheng.lao@bupt.edu.cn)

<sup>3</sup> Shanghai Artificial Intelligence Laboratory, Shanghai, China

<sup>4</sup> College of Computer Science, Sichuan University, Chengdu, China

**Abstract.** Masked autoencoder (MAE) has attracted unprecedented attention and achieves remarkable performance in many vision tasks. It reconstructs random masked image patches (known as proxy task) during pretraining and learns meaningful semantic representations that can be transferred to downstream tasks. However, MAE has not been thoroughly explored in ultrasound imaging. In this work, we investigate the potential of MAE for ultrasound image recognition. Motivated by the unique property of ultrasound imaging in high noise-to-signal ratio, we propose a novel deblurring MAE approach that incorporates deblurring into the proxy task during pretraining. The addition of deblurring facilitates the pretraining to better recover the subtle details presented in the ultrasound images, thus improving the performance of the downstream classification task. Our experimental results demonstrate the effectiveness of our deblurring MAE, achieving state-of-the-art performance in ultrasound image classification. Overall, our work highlights the potential of MAE for ultrasound image recognition and presents a novel approach that incorporates deblurring to further improve its effectiveness.

**Keywords:** Image Deblurring · Masked Autoencoders · Self-Supervised Learning · Ultrasound Recognition

## 1 Introduction

Recently, as representative of generative self-supervised learning (SSL) methods, masked autoencoder (MAE) [8] has achieved great success in many vision

Code will be available at: <https://github.com/MembrAI/DeblurringMIM>.

**Supplementary Information** The online version contains supplementary material available at [https://doi.org/10.1007/978-3-031-43907-0\\_34](https://doi.org/10.1007/978-3-031-43907-0_34).

tasks [10, 11, 24]. In general, MAE belongs to the masked image modeling (MIM) paradigm [29], where some parts of the image are randomly masked, and the purpose of pretraining (i.e., proxy or pretext task) is to recover the missing pixels. After the pretraining, the learned image representation can be transferred to downstream tasks for improved performance. With the advent of MAE, many MAE variants have been proposed [7, 22, 25]. Tian *et al.* [22] investigate other image degradation methods during MAE pretraining and find that the optimal practice is enriching masking with spatial misalignment for nature images. Wu *et al.* [25] design a denoising MAE by introducing Gaussian noising into MAE pretraining, showing that their denoising MAE is robust to additive noises.

On the other hand, although numerous work has been proposed for applying MAE to medical imaging across different modalities including pathological images [1, 14, 19], X-rays [26, 31], electrocardiogram [30], immunofluorescence images [15], MRI and CT [4, 27, 31]. However, the majority of them have not fully exploited the characteristics of medical images and instead, focus on vanilla applications [4, 26, 30, 31]. This is especially problematic given the domain gap between medical and natural images, as well as the unique imaging properties associated with each medical imaging modality [16, 18, 20]. Furthermore, as an important and widely used medical imaging modality, ultrasound has not been extensively explored in the context of MAE-based approaches.

Based on the aforementioned analysis, in this paper, we propose a deblurring masked auto-encoder framework, which is specifically designed for ultrasound image recognition. The primary motivation for the deblurring comes from the unique imaging properties of ultrasound, e.g., high noise-to-signal ratio. Compared with nature images, the subtle details within ultrasound are particularly important for downstream analysis (e.g., microcalcifications is an important sign for malignant nodules, which is represented as tiny bright spots in ultrasound [17, 21]). Moreover, the motivation also stems from the findings of our preliminary experiments, which suggest that denoising may not be appropriate for inherently noisy ultrasound images. Therefore, we introduce the opposite direction with a deblurring approach for ultrasound images. Specifically, we first apply blurring operations to the ultrasound images prior to the random masking during pretraining, enabling the model to learn how to de-blur and reconstruct the original image. It should be emphasized that denoising and deblurring are two opposite directions, i.e., denoising first adds noise to the clean image and learns to remove the noise, while deblurring blurs the noisy ultrasound image and learns to sharpen the image. The deblurring facilitates the pretraining in recovering the subtle details within the image, which is crucial for ultrasound image recognition. It should be emphasized that while blurring operation has been shown ineffective for natural images [22], ultrasound images are fundamentally different and may benefit from blurring operation.

Furthermore, to the best of our knowledge, this paper is the first attempt to apply the MAE approach to ultrasound image recognition. Our work also addresses some fundamental concerns that are of great interest to the medical imaging community with the example of ultrasound, such as the importance of

in-domain data pretraining for MAE in ultrasound, as well as the finding that SSL pretraining is consistently better than the supervised pretraining as with nature images. To conclude, our contributions can be summarized as follows:

1. We propose a deblurring MAE framework that is specifically designed for ultrasound images by incorporating a deblurring task into MAE pretraining. This is motivated by the fact that ultrasound images have a high noise-to-signal ratio, and in contrast to denoising for natural images, we demonstrate that deblurring is a better recipe for ultrasound images.
2. We explore the effectiveness of various image blurring methods in our deblurring MAE and find that a simple Gaussian blurring performs the best, showing superior transferability compared with the vanilla MAE.
3. We conduct experiments on more than 10k ultrasound images for pretraining and 4,494 images for downstream thyroid nodule classification. The results demonstrate the effectiveness of the proposed deblurring MAE, achieving state-of-the-art classification performance for ultrasound images.

Note that, as a representative MIM approach, the MAE is adopted to validate our proposed deblurring pretraining in this work, our method can also be seamlessly integrated with other MIM-based approaches such as ConvMAE [7].

## 2 Method

### 2.1 Preliminary: MAE

The MAE pipeline consists of two primary stages: self-supervised pretraining and transferring for downstream tasks. During the self-supervised pretraining, the model is trained to reconstruct masked input image patches using an asymmetric encoder-decoder architecture. The encoder is typically a ViT [6], which compresses the input image into a latent representation, while the decoder is a lightweight Transformer that reconstructs the original image from the latent representation. The loss used during pretraining is the mean squared error (MSE) between the reconstructed and original images. In the transfer stage, the weights of the pre-trained ViT encoder are transferred and used as a feature extractor, to which task-specific heads are appended for learning various downstream tasks. Typically, there are two common practices in the transfer stage: 1) end-to-end fine-tuning which tunes the entire model, and 2) linear probing, which only tunes the task-specific head.

### 2.2 Our Proposed Deblurring MAE

Similar to MAE, our proposed deblurring MAE also contains pretraining and transfer learning for downstream tasks. We employ the same asymmetric encoder-decoder architecture as the original MAE.

**Deblurring MAE Pretraining.** For the pretraining, besides the original masked image modeling task in the MAE, we introduce one additional task, i.e., deblurring, into the pretraining thus making the pretraining as deblurring pretraining. As shown in Fig. 1, the deblurring is achieved by simply inserting an image blurring operation prior to random masking. The pipeline of our deblurring MAE pretraining is illustrated in Eq. 1:

$$x \xrightarrow{\text{Blurring}} x_b \xrightarrow{\text{Masking}} x_b^m \xrightarrow{\text{ViT Encoder}} h \xrightarrow{\text{Decoder}} \hat{x}. \quad (1)$$

Specifically, the original ultrasound image  $x$  is first blurred by a chosen image blurring operation *Blurring* to obtain  $x_b$ . After that, several patches in the blurred image  $x_b$  are randomly masked by the *Masking* operation with a pre-defined ratio to obtain  $x_b^m$ . Next, the masked blurred image  $x_b^m$  is passed as input to the ViT Encoder, which generates a latent representation  $h$ . Finally, the Decoder receives the representation  $h$  and outputs reconstructed image  $\hat{x}$ .

The image blurring operation *Blurring* is a commonly used technique for reducing the sharpness or details of an image, resulting in a smoother, less-detailed appearance. There exist many different methods for image blurring, with most of them involving the averaging of neighboring pixels in some way. In Fig. 1, we provide examples of two representative blurring methods: Gaussian blur and speckle reducing anisotropic diffusion (SRAD) [28].

Gaussian blur involves convolving an input image with a Gaussian kernel  $G(\sigma)$ , which is a two-dimensional Gauss function that represents a normal distribution with standard deviation of  $\sigma$ . Mathematically, Gaussian blur can be defined as follows:

$$x_b = \text{Gaussian}(x, \sigma) = x * G(\sigma) = x * \frac{1}{2\pi\sigma^2} e^{-(u^2+v^2)/2\sigma^2}, \quad (2)$$

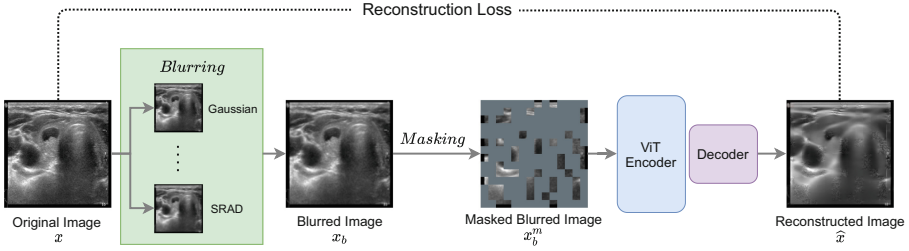
where  $*$  denotes the convolution operation, and  $(u, v)$  represents the coordinates in the kernel. The degree of blurring (i.e., blurriness) in the resulting image is determined by the standard deviation  $\sigma$ .

The SRAD is a nonlinear anisotropic diffusion technique for removing speckled noises, which has been extensively used in medical ultrasound images, due to its edge-sensitivity for speckled images and powerful preservation of useful information. The SRAD operation is implemented by repeating an anisotropic diffusion equation for  $N$  iterations. It can be formally given as:

$$x_b = \text{SRAD}(x, N, t) = x(i, j, 0) + \Delta t * \sum_{k=0}^{N-1} \text{div}(c(i, j, k) \nabla x(i, j, k)), \quad (3)$$

where  $x$  is the original image,  $N$  stands for the number of iterations,  $t$  means time.  $x(i, j, k)$  and  $c(i, j, k)$  represent the image and diffusivity coefficient at iteration  $k$ , respectively.  $\nabla x$  is the gradient of  $x$  and  $\text{div}$  is the divergence operator. The larger  $N$  or  $t$  leads to a blurrier resulting image.

The pixel-wise MSE between the reconstructed image  $\hat{x}$  and the original image  $x$  is utilized as the loss function during pretraining:  $\mathcal{L}_{MSE} = \|\hat{x} - x\|_2$ .



**Fig. 1.** Illustration of our proposed deblurring MAE pretraining.

It should be noted that a key difference from MAE is that we compute the loss across all patches, including the masked ones. This operation is necessary due to the fact that our blurring operation covers the entire image. Through the use of the proposed deblurring MAE pretraining, we aim to leverage both masked image modeling and deblurring in order to learn a robust and effective latent representation that could be successfully applied to a range of downstream tasks.

**Deblurring MAE Transfer.** After the deblurring MAE pretraining, only the pre-trained encoder is transferred to the downstream thyroid nodule classification task. One multi-layer perceptron (MLP) head is appended after the pre-trained encoder. The transfer learning pipeline is shown in Eq. 4:

$$x \xrightarrow{\text{Blurring}} x_b \xrightarrow{\text{ViT Encoder}} h \xrightarrow{\text{MLP}} \hat{y}, \quad (4)$$

It should be noted here that, in order to prevent data distribution shift between pretraining and transfer stages, the original image  $x$  also needs to be blurred before fed into the pre-trained encoder during transfer learning. The cross-entropy loss between ground-truth classification label  $y$  and predicted label  $\hat{y}$  is used as the loss function:  $\mathcal{L}_{CE} = -[y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})]$ .

### 3 Experiments and Results

#### 3.1 Experimental Settings

**Dataset.** All thyroid ultrasound images used in our study for both pretraining and downstream classification were acquired at West China Hospital with ethical approval. We use a total of 10,675 images for pretraining and 4,493 images for the downstream classification. To avoid any potential data leakage, the images used in pretraining were not included in the test set for thyroid nodule classification. The downstream classification dataset contains 2,576 benign and 1,917 malignant cases. We randomly split the dataset into train/validation/test subsets with a 3:1:1 ratio. The classification ground-truth labels were obtained either from the fine-needle aspiration for malignant nodules or clinical diagnosis by senior radiologists for benign nodules.

**Table 1.** Performance comparison of different methods.

	Method	Architecture	Pretraining	ACC (%)	F1 (%)	AUROC (%)
Supervised	ResNet [9]	ResNet-101	-	86.06 ± 0.87	83.18 ± 1.15	91.96 ± 1.47
	ConvNeXt [13]	ConvNeXt-L	ImageNet	87.76 ± 0.66	85.47 ± 0.91	93.22 ± 1.10
	Swin Transformer [12]	Swin-L	ImageNet	87.43 ± 0.68	84.92 ± 0.82	92.83 ± 1.02
	Wang <i>et al.</i> [23]	-	-	87.44 ± 0.75	85.16 ± 0.87	93.11 ± 1.09
	Zhou <i>et al.</i> [32]	-	-	88.15 ± 0.67	86.09 ± 0.74	94.17 ± 1.21
	ViT [6]	ViT-B	-	80.60 ± 1.62	76.98 ± 2.05	83.89 ± 2.89
			ImageNet	86.38 ± 0.74	84.17 ± 0.98	92.69 ± 0.48
SSL	SimCLR [2]	ResNet-50	ImageNet	86.21 ± 0.96	83.81 ± 1.24	92.16 ± 1.08
	MoCo v3 [3]	ViT-B	ImageNet	86.96 ± 0.85	84.48 ± 1.12	92.77 ± 0.67
			Ultrasound	87.08 ± 0.78	84.55 ± 1.04	92.95 ± 0.59
	MAE [8]		ImageNet	87.25 ± 0.51	85.23 ± 0.57	93.71 ± 0.60
			Ultrasound	89.45 ± 0.53	87.54 ± 0.62	95.54 ± 0.46
	Denoising MAE		Ultrasound	80.38 ± 1.37	77.99 ± 1.75	84.38 ± 2.13
	Ours [SRAD]	Ultrasound	90.07 ± 0.47	88.13 ± 0.51	95.87 ± 0.45	
Ours [Gaussian]	Ultrasound	<b>90.19 ± 0.47</b>	<b>88.48 ± 0.50</b>	<b>96.08 ± 0.41</b>		

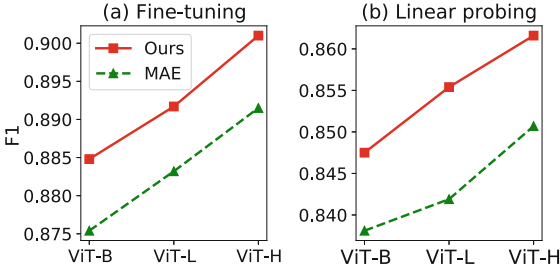
**Fig. 2.** Our deblurring MAE vs. vanilla MAE. Pre-trained with the same ultrasound data.**Table 2.** Ablation study.

Image Blurring		Blurriness	
Method	F1 (%)	$\sigma$	F1 (%)
Gaussian	<b>88.48</b>	0.8	88.03
SRAD	88.13	1.1	<b>88.48</b>
Mean	88.11	1.4	87.56
Median	87.63	1.7	85.89
Motion	78.33	2.0	84.74
Defocus	85.37	2.3	84.62
Baseline	87.54		

**Implementation Details.** We use a mask ratio of 75% during the pretraining. We set the batch size to 256 for both pretraining and end-to-end fine-tuning, and 1024 for linear probing. The epochs of pretraining is 12,000 due to our relatively small data. The full detailed experimental settings are presented in the appendix. We implement our approach based on PyTorch. The image size for both pretraining and transfer learning is  $224 \times 224$ . For classification, we choose the model that performs the best on the validation set as the final model to evaluate on the test set. Three widely used metrics accuracy (ACC), F1-score (F1), and the area under the receiver operating characteristic (AUROC) are utilized for classification performance evaluation.

### 3.2 Results and Comparisons

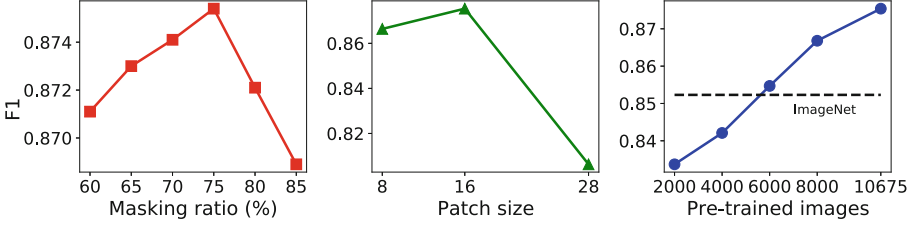
**Our Deblurring MAE vs. Vanilla MAE.** First of all, in order to evaluate the effectiveness of the proposed deblurring MAE for ultrasound images, we compare the transfer learning performance between our deblurring MAE and the

vanilla MAE. Table 1 and Fig. 2 give the classification performance comparison of these two approaches. For our deblurring MAE, we use Gaussian blurring with  $\sigma$  equal to 1.1 as the blurring operation. In Fig. 2, we report the experimental results of three models: ViT-Base (ViT-B), ViT-Large (ViT-L) and ViT-Huge (ViT-H), and two transfer learning paradigms: end-to-end fine-tuning and linear probing. As shown in the figure, both the fine-tuning and linear probing performance of our proposed deblurring MAE is consistently better than that of the vanilla MAE, which indicates the effectiveness of deblurring for enhancing the transferability of learned representations during ultrasound pretraining.

**Comparison with State-of-the-Art Approaches.** Secondly, we also compare our approach with more approaches and the results are listed in Table 1. We implement two variants of our deblurring MAE which differ in blurring operation: the SRAD with  $N$  equals to 40 and  $t$  equals to 0.1, and the Gaussian blur with  $\sigma$  equals to 1.1. We compare with methods based on supervised learning or self-supervised learning. In addition, we still add the denoising MAE for comparison, although it has proved to be ineffective for ultrasound images based on our preliminary experiments. We adopt ViT-B as the architecture for these SSL-based methods except SimCLR [2] which uses ResNet-50, and we use two types of data for pretraining, i.e., ImageNet [5] and ultrasound. The results are based on end-to-end fine-tuning. According to Table 1, we can draw the following conclusions:

**The Deblurring MAE Pretraining Can Improve the Transferability of Learned Representations.** First of all, both the two variants of our proposed approach (Ours [SRAD] and Ours [Gaussian]) obtain much higher classification metrics compared with the MAE pretrained using ultrasound, which indicates the learned representation of our deblurring MAE is more effective than the vanilla MAE when transferred to downstream classification. In addition, Table 1 also shows that the performance of our proposed deblurring MAE with Gaussian blurring achieves state-of-the-art performance in terms of all metrics, surpassing all competing SSL or supervised-based approaches, which further demonstrates the superior performance of our proposed deblurring MAE. It is noteworthy that, as the opposite approach to our deblurring MAE, the denoising MAE obtains worse performance compared with the vanilla MAE, suggesting that adding noise to ultrasound images during MAE pretraining is unfavorable.

**Ultrasound Pretraining is Better than ImageNet Pretraining, Better than Supervised Pretraining.** Table 1 shows that the performance of MAE with ultrasound pretraining is better than the ImageNet pretraining, which underlines the importance of in-domain self-supervised pretraining in MAE. In contrast to MAE, our experiments show that the MoCo v3 [3] achieves only marginal improvement with ultrasound pretraining. Furthermore, the MAE ImageNet pretraining also performs much better than the ImageNet supervised pretraining. These two conclusions are consistent with other works [8, 26].



**Fig. 3.** Hyper-parameter choices for MAE pretraining.

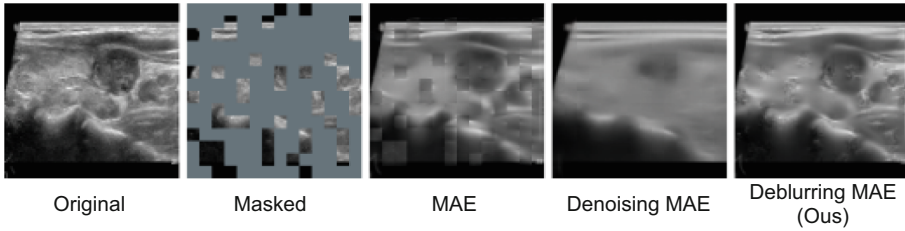
**Ablation Study.** We design two sets of ablation studies, i.e., different image blurring methods, and the degree of blurring (blurriness) used in our deblurring MAE. We adopt the ViT-B as the architecture and end-to-end fine-tuning in transfer learning for the ablation experiments. Table 2 presents the performance results of the ablation study, where the ‘Baseline’ represents the vanilla MAE.

Firstly, besides the Gaussian and SRAD, we also try several other blurring methods that are commonly used in the fields including mean, median, motion and defocus blur. We set the kernel size to 5 in mean, median and motion blur, and the radius of defocusing is set to 5 for defocus blur. The performance results are presented on the left side of Table 2. From this table, we can observe that the Gaussian blurring achieves the best F1. And these six blurring methods are not all beneficial for pretraining, where some of them (motion, defocus) perform even worse than the baseline. Secondly, to investigate the effect of blurriness on the pretraining, we conduct ablation experiments on blurriness based on Gaussian blurring. The right side of Table 2 reports the performance results and we can see that the  $\sigma$  with 1.1 obtains the highest F1. In addition, as the  $\sigma$ , i.e., the blurriness continues to increase, the performance drops rapidly, which indicates that only a limited range of blurriness has a positive effect on the pretraining.

**Hyper-parameter Choices for MAE Pretraining.** We conduct experiments to explore hyper-parameter choices for MAE pretraining based on ViT-B, and the results are presented in Fig. 3. Our findings indicate that a masking ratio of 75% and a patch size of 16 achieve the best transfer performance, which is consistent with MAE for natural images [8]. Additionally, we observed that transfer performance improves with an increase in pre-trained images, surpassing ImageNet transfer only when a substantial amount of pre-trained images is used.

**Visualization.** The comparisons of reconstructed image examples among MAE, denoising MAE, and our proposed deblurring MAE are illustrated in Fig. 4. Although there is no strong evidence that reveals the relationship between reconstruction quality in pretraining and downstream task performance in MAE-based approaches, we can still obtain some insights from the reconstruction quality. As shown in Fig. 4, we can clearly observe that the reconstructed images of the





**Fig. 4.** Comparisons of reconstructed images.

denoising MAE are the smoothest and lost most details among all the three approaches, followed by the vanilla MAE, and our deblurring MAE achieves the best reconstruction quality with much finer details. The comparisons indicate that our deblurring MAE can capture critical details that are beneficial for downstream classification. More comparisons can be found in the appendix.

## 4 Conclusion and Future Work

In this paper, we propose a novel deblurring MAE by incorporating deblurring into the proxy task during MAE pretraining for ultrasound image recognition. The deblurring task is implemented by inserting image blurring operation prior to the random masking during pretraining. The integration of deblurring enables the pretraining pay more attention to recovering the intricate details presented in ultrasound images, which are critical for downstream image classification. We explore the effect of several different image blurring methods and find that Gaussian blurring achieves the best performance and only a limited range of blurriness has a beneficial effect for pretraining. Based on the optimal blurring method and blurriness, our deblurring MAE achieves state-of-the-art performance in the downstream classification of ultrasound images, indicating the effectiveness of incorporating deblurring into MAE pretraining for ultrasound image recognition. However, this work has some limitations. For example, only one downstream task: nodule classification is evaluated in this study. We plan to extend our approach to include more tasks such as segmentation in the future.

**Acknowledgment.** This work was supported by Natural Science Foundation of Sichuan Province under Grant NO. 2022NSFSC1855.

## References

1. An, J., Bai, Y., Chen, H., Gao, Z., Litjens, G.: Masked autoencoders pre-training in multiple instance learning for whole slide image classification. In: *Medical Imaging with Deep Learning* (2022)
2. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: *International Conference on Machine Learning*, pp. 1597–1607. PMLR (2020)

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. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. IEEE (2009)
6. Dosovitskiy, A., et al.: An image is worth  $16 \times 16$  words: transformers for image recognition at scale. arXiv preprint [arXiv:2010.11929](https://arxiv.org/abs/2010.11929) (2020)
7. Gao, P., Ma, T., Li, H., Dai, J., Qiao, Y.: ConvMAE: masked convolution meets masked autoencoders. arXiv preprint [arXiv:2205.03892](https://arxiv.org/abs/2205.03892) (2022)
8. 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)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
10. Ke, L., Danelljan, M., Li, X., Tai, Y.W., Tang, C.K., Yu, F.: Mask transfiner for high-quality instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4412–4421 (2022)
11. Li, Y., Mao, H., Girshick, R., He, K.: Exploring plain vision transformer backbones for object detection. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) ECCV 2022. Lecture Notes in Computer Science, vol. 13669, pp. 280–296. Springer, Cham (2022). [https://doi.org/10.1007/978-3-031-20077-9\\_17](https://doi.org/10.1007/978-3-031-20077-9_17)
12. Liu, Z., et al.: Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022 (2021)
13. 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)
14. Luo, Y., Chen, Z., Gao, X.: Self-distillation augmented masked autoencoders for histopathological image classification. arXiv preprint [arXiv:2203.16983](https://arxiv.org/abs/2203.16983) (2022)
15. Ly, S.T., Lin, B., Vo, H.Q., Maric, D., Roysam, B., Nguyen, H.V.: Student collaboration improves self-supervised learning: dual-loss adaptive masked autoencoder for brain cell image analysis. arXiv preprint [arXiv:2205.05194](https://arxiv.org/abs/2205.05194) (2022)
16. Niu, S., Liu, M., Liu, Y., Wang, J., Song, H.: Distant domain transfer learning for medical imaging. IEEE J. Biomed. Health Inform. **25**(10), 3784–3793 (2021)
17. Park, M., et al.: Sonography of thyroid nodules with peripheral calcifications. J. Clin. Ultrasound **37**(6), 324–328 (2009)
18. Qin, Z., Yi, H., Lao, Q., Li, K.: Medical image understanding with pretrained vision language models: a comprehensive study. arXiv preprint [arXiv:2209.15517](https://arxiv.org/abs/2209.15517) (2022)
19. Quan, H., et al.: Global contrast masked autoencoders are powerful pathological representation learners. arXiv preprint [arXiv:2205.09048](https://arxiv.org/abs/2205.09048) (2022)
20. Raghu, M., Zhang, C., Kleinberg, J., Bengio, S.: Transfusion: understanding transfer learning for medical imaging. In: Advances in Neural Information Processing Systems, vol. 32 (2019)

21. Taki, S., et al.: Thyroid calcifications: sonographic patterns and incidence of cancer. *Clin. Imaging* **28**(5), 368–371 (2004)
22. Tian, Y., et al.: Beyond masking: demystifying token-based pre-training for vision transformers. arXiv preprint [arXiv:2203.14313](https://arxiv.org/abs/2203.14313) (2022)
23. Wang, P., Patel, V.M., Hacihaliloglu, I.: Simultaneous segmentation and classification of bone surfaces from ultrasound using a multi-feature guided CNN. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) *MICCAI 2018*. LNCS, vol. 11073, pp. 134–142. Springer, Cham (2018). [https://doi.org/10.1007/978-3-030-00937-3\\_16](https://doi.org/10.1007/978-3-030-00937-3_16)
24. Wang, X., Zhao, K., Zhang, R., Ding, S., Wang, Y., Shen, W.: ContrastMask: Contrastive learning to segment every thing. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11604–11613 (2022)
25. Wu, Q., Ye, H., Gu, Y., Zhang, H., Wang, L., He, D.: Denoising masked autoencoders are certifiable robust vision learners. arXiv preprint [arXiv:2210.06983](https://arxiv.org/abs/2210.06983) (2022)
26. Xiao, J., Bai, Y., Yuille, A., Zhou, Z.: Delving into masked autoencoders for multi-label thorax disease classification. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3588–3600 (2023)
27. Xu, Z., et al.: Swin MAE: masked autoencoders for small datasets. arXiv preprint [arXiv:2212.13805](https://arxiv.org/abs/2212.13805) (2022)
28. Yu, Y., Acton, S.T.: Speckle reducing anisotropic diffusion. *IEEE Trans. Image Process.* **11**(11), 1260–1270 (2002)
29. Zhang, C., Zhang, C., Song, J., Yi, J.S.K., Zhang, K., Kweon, I.S.: A survey on masked autoencoder for self-supervised learning in vision and beyond. arXiv preprint [arXiv:2208.00173](https://arxiv.org/abs/2208.00173) (2022)
30. Zhang, H., et al.: MaeFE: masked autoencoders family of electrocardiogram for self-supervised pretraining and transfer learning. *IEEE Trans. Instrum. Meas.* **72**, 1–15 (2022)
31. Zhou, L., Liu, H., Bae, J., He, J., Samaras, D., Prasanna, P.: Self pre-training with masked autoencoders for medical image analysis. arXiv preprint [arXiv:2203.05573](https://arxiv.org/abs/2203.05573) (2022)
32. Zhou, Y., et al.: Multi-task learning for segmentation and classification of tumors in 3D automated breast ultrasound images. *Med. Image Anal.* **70**, 101918 (2021)