Wave-MAE: Wavelet Transform Meets Masked Autoencoder for Multiple Thorax Disease Classification

Yunze Wang*[1], Silin Chen*[3], Tianyang Wang*[2], Zhuo Zhang[4], Jingxin Liu†[2]
[1]University of Edinburgh, Edinburgh, UK
[2]Xi'an Jiaotong - Liverpool University, Suzhou, China
[3]Zhejiang University, Hangzhou, China
[4]National University of Defense Technology, Changsha, China

***Abstract*—We introduce Wave-MAE, a novel wavelet-based masked autoencoder designed for medical image analysis. In the pre-training stage, Wave-MAE leverages wavelet transforms to decompose images into high- and low-frequency feature patches, from which a random subset is selected for input to the encoder. The decoder leverages cross-attention between masked and visible tokens to reconstruct a subset of masked tokens, with a cross-layer interaction mechanism that enables it to effectively utilize feature maps from all encoder layers. Additionally, we propose a two-step tuning strategy to optimize the model for downstream tasks. To the best of our knowledge, Wave-MAE is the first wavelet-based self-supervised approach for medical image analysis. Extensive experiments on three public Chest X-Ray datasets demonstrate that our method significantly improves performance on multiple thorax disease classification tasks.**

***Index Terms*—Self-supervised Learning, Wavelet Transform, Medical Image Classification, Masked Autoencoder, Chest X-Ray**

I. Introduction

Vision Transformers (ViTs) [1] have achieved strong performance in computer vision tasks, thanks to their ability to capture long-range dependencies and complex visual patterns [2], [3]. Despite these advancements, training ViTs effectively requires a large volume of labelled data, which poses significant challenges in the context of medical image analysis [4]. Unlike natural images, which can be easily annotated through crowdsourcing, medical images, such as chest X-rays and CT scans, demand highly specialized knowledge for precise labelling. This is particularly difficult in the classification of multiple thoracic diseases, where precise labelling is crucial but labour-intensive and costly [5].

A promising solution to this problem is the application of Self-Supervised Learning (SSL) techniques, which aim to extract meaningful representations from large amounts of unlabeled data [6]. SSL has emerged as a viable method for addressing the scarcity of labelled medical data by learning latent patterns that can be transferred to downstream tasks [7]–[9]. This approach is especially relevant for medical imaging, where acquiring large labelled datasets is often impractical. In the case of thoracic disease classification, SSL methods enable models to learn robust features that can improve classification performance, even with limited labelled samples [10].

Despite this, existing SSL approaches like Masked Autoencoders (MAE) do not fully leverage the High-Frequency (HF) and Low-Frequency (LF) features of medical images, which are key for representation learning. They also lack optimization for handling out-of-distribution tasks, such as varying pathological features in thoracic disease classification, highlighting the need for specialized SSL techniques [11].

In this study, we propose a novel self-supervised framework called Wave-MAE, which incorporates wavelet transforms to improve representation learning. Unlike vanilla MAE [12], Wave-MAE converts images into HF and LF feature maps during pre-training. A random subset of these maps is input to the encoder, while the decoder uses cross-attention between the fused encoder output and masked tokens to reconstruct both frequency components. This cross-attention mechanism allows for partial token reconstruction, reducing computational cost. We also introduce a cross-layer interaction module that enables the decoder to adaptively utilize outputs from different encoder layers. During fine-tuning, the pre-trained encoder extracts latent representations, and a two-stage tuning strategy mitigates domain shifts between pre-training and downstream tasks. These components together enhance performance on downstream tasks with efficient computation.

The contributions of this paper are as follows:

(1) We propose Wave-MAE, a self-supervised learning framework utilizing wavelet transforms for more efficient and improved representation learning.

(2) We introduce a two-step tuning strategy to boost generalization performance on downstream tasks.

(3) Extensive experiments on three publicly available datasets validate the effectiveness of our framework.

II. Methodology

In this section, we present a wavelet-based self-supervised learning framework for thoracic disease classification. As illustrated in Fig. 1, the proposed framework consists of two main stages: pre-training and two-step tuning. We will elaborate on the details of each stage in the following subsections. Specifically, in Section. II-A, we revisit the wavelet transform. Then, in Section. II-B1, we introduce the details of the Wave-MAE encoder. In Section. II-B2, we discuss the cross-attention decoder and further explain the partial reconstruction and inter-layer interaction mechanisms. Finally, in Section. II-C, we present our improved two-step tuning strategy in detail.

A. Wavelet Transform

We perform a first-level wavelet decomposition of an image into four components: Low-Low (LL), High-Low (HL), Low-

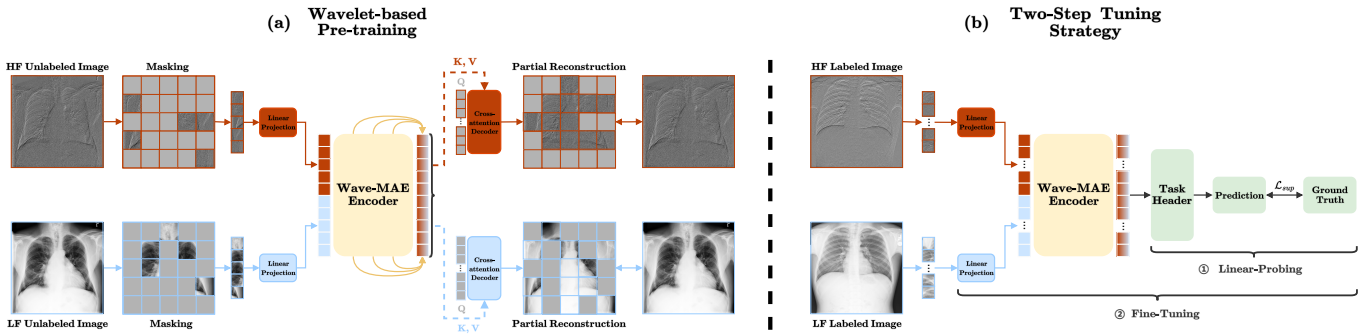

Fig. 1: **(a) Pre-training**: Patches randomly sampled from HF and LF features are linearly projected to a fixed dimension and encoded using a Transformer. The decoder leverages cross-attention between the encoded features (as keys and values) and the masked tokens (as queries) for reconstruction. Benefiting from the use of cross-attention, we are able to: 1) Reconstruct only a subset of masked tokens, thereby enhancing the efficiency of pre-training; and 2) Unlike MAE, which only feeds the final layer feature map to the decoder, we incorporate feature maps from intermediate layers of the encoder. **(b) Two-step Tuning Strategy**: After pre-training on HF and LF features, Wave-MAE can be tuned on downstream tasks. Considering potential domain shifts, we first tune the linear layer while keeping the pre-trained weights fixed, followed by tuning the entire network.

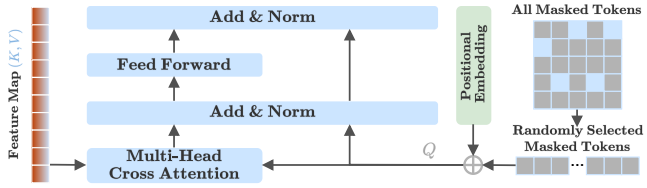

Fig. 2: **Structure of the first cross-attention decoder block**.

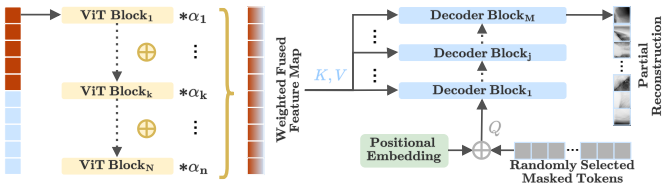

Fig. 3: **Details of the full pre-training pipeline.**

High (LH), and High-High (HH). Inspired by [13], we define the low-frequency image, I_{LF}, as:

$$I_{LF} = LL, \tag{1}$$

and the high-frequency image, I_{HF}, as a combination of the remaining components:

$$I_{HF} = HL + LH + HH. \tag{2}$$

The I_{LF} captures detailed structural information, while I_{HF} highlights edge and boundary features, enhancing sensitivity to subtle features and improving boundary detection. These images are used as both inputs and reconstruction targets, enriching the diversity of representation learning [14].

B. Wavelet-based Pre-training Protocol

1) Wave-MAE Encoder: As shown in the left half of Fig. 1(a), we propose a two-branch encoder to enhance feature diversity, inspired by the multi-modal MAE [15]. We use the random subsets of HF and LF features obtained from the wavelet transform as inputs. The proportion of tokens sampled in each branches is controlled by the Dirichlet distribution $(\lambda_{\text{HF}}, \lambda_{\text{LF}}) \sim \text{Dir}(\alpha)$, where $\lambda_{\text{HF}} + \lambda_{\text{LF}} = 1$. Smaller α values ($\alpha \ll 1$) concentrate tokens in one branch, while larger values ($\alpha \gg 1$) distribute them evenly. We set α to positive infinity ($\alpha = +\infty$) to avoid biasing the sampling towards any certain branch, i.e. HF or LF. Next, we randomly sample the corresponding number of visible tokens from each branch without replacement [12]. For efficiency, we fix the number of visible tokens at 49 in each branch for all our experiments (75% masking ratio of all tokens from a 224×224 image with 16×16 patches).

Separate linear projection layers map sampled HF and LF tokens to Transformer-compatible dimensions. 2D sine-cosine positional encodings are added [16], [17] and we do not add any frequency-specific encodings. The projected tokens are then concatenated and passed through the Transformer. In summary, the Wave-MAE encoder enhances the quality of representation learning by explicitly leveraging HF and LF features during the pre-training stage, which leads to better initialization of the ViT for numerous downstream tasks.

2) Cross-Attention Decoder: As shown in the right half of Figure 1(a), we used different decoders to separately do our best reconstruct the masked tokens in each branch. Building on experiments with CrossMAE [18], we consider that self-attention among masked tokens is unnecessary. Instead, we solely use cross-attention between visible and masked tokens for reconstruction. The details are illustrated in Fig. 2, for each branch, the key and value tokens are taken from the encoder output, with the first decoder block using the sum of masked tokens and their positional embeddings as queries, while later blocks use the previous layer's output as queries.

Utilizing cross-attention instead of self-attention for reconstruction brings two additional benefits [18]: (i) Unlike vanilla MAE, which reconstructs all masked tokens, our method

TABLE 2: Performance comparison of different self-supervised methods on various thoracic conditions in ChestX-Ray14 [19].

Method	SimMIM [20]	DINO [21]	MAE [12]	SimCLRV2 [22]	MoCoV3 [23]	BYOL [24]	PixMIM [25]	**Wave-MAE$^{\triangledown}$**	**Wave-MAE**
Atelectasis	81.39 ± 0.21	81.28 ± 0.23	80.88 ± 0.12	78.14 ± 0.14	79.32 ± 0.11	77.27 ± 0.22	83.04 ± 0.21	83.99 ± 0.20	**84.01 ± 0.41**
Nodule	72.88 ± 0.11	69.53 ± 0.20	71.09 ± 0.18	71.02 ± 0.33	70.18 ± 0.62	70.25 ± 0.24	74.11 ± 0.16	76.42 ± 0.36	**76.77 ± 0.32**
Pneumonia	68.23 ± 0.55	69.11 ± 0.42	70.55 ± 0.21	67.32 ± 0.74	68.74 ± 0.22	66.69 ± 0.34	70.88 ± 0.48	70.87 ± 0.43	**71.33 ± 0.44**
Pneumothorax	84.28 ± 0.38	83.54 ± 0.13	82.94 ± 0.11	82.71 ± 0.55	84.10 ± 0.11	81.16 ± 0.46	85.68 ± 0.49	**86.69 ± 0.37**	86.60 ± 0.15
Infltrate	73.10 ± 0.19	72.17 ± 0.63	73.36 ± 0.10	70.01 ± 0.16	71.93 ± 0.21	71.06 ± 0.29	**75.12 ± 0.77**	74.88 ± 0.11	74.76 ± 0.31
Effusion	80.17 ± 0.34	77.58 ± 0.51	80.48 ± 0.44	77.29 ± 0.42	78.67 ± 0.26	76.39 ± 0.52	81.23 ± 0.89	82.79 ± 0.40	**83.11 ± 0.29**
Cardiomegal	87.49 ± 0.32	86.02 ± 0.17	86.26 ± 0.16	85.48 ± 0.31	87.77 ± 0.18	84.57 ± 0.29	87.64 ± 0.23	**88.61 ± 0.31**	88.46 ± 0.13
Mass	76.12 ± 0.11	76.91 ± 0.19	75.01 ± 0.31	74.62 ± 0.41	74.13 ± 0.41	73.62 ± 0.55	77.95 ± 0.12	78.76 ± 0.58	**79.08 ± 0.22**
Consolidation	78.42 ± 0.34	75.40 ± 0.57	79.75 ± 0.16	74.03 ± 0.17	74.96 ± 0.38	73.41 ± 0.23	78.19 ± 0.78	80.03 ± 0.27	**80.29 ± 0.31**
Pleural Thicken	**80.69 ± 0.11**	79.71 ± 0.49	79.16 ± 0.23	77.45 ± 0.26	79.16 ± 0.21	76.34 ± 0.32	79.08 ± 0.11	79.53 ± 0.21	79.61 ± 0.45
Emphysema	87.13 ± 0.21	86.74 ± 0.41	87.29 ± 0.11	86.23 ± 0.38	87.12 ± 0.11	86.04 ± 0.12	88.42 ± 0.25	89.85 ± 0.51	**90.22 ± 0.24**
Fibrosis	77.16 ± 0.36	76.33 ± 0.42	78.05 ± 0.37	74.93 ± 0.25	76.45 ± 0.26	75.61 ± 0.42	79.51 ± 0.19	81.29 ± 0.14	**81.45 ± 0.27**
Edema	83.72 ± 0.21	82.43 ± 0.54	82.99 ± 0.12	81.36 ± 0.19	82.83 ± 0.47	80.42 ± 0.21	84.22 ± 0.17	**85.04 ± 0.33**	84.39 ± 0.11
Hernia	74.18 ± 0.56	73.11 ± 0.30	73.61 ± 0.25	73.42 ± 0.16	75.44 ± 0.11	72.51 ± 0.71	77.05 ± 0.44	78.93 ± 0.11	**79.24 ± 0.47**
Mean	78.93 ± 0.09	77.85 ± 0.11	78.67 ± 0.06	76.72 ± 0.10	77.91 ± 0.08	76.10 ± 0.10	80.15 ± 0.12	81.26 ± 0.09	**81.38 ± 0.08**

TABLE 1: Hyperparameters for pre-training and fine-tuning.

Hyperparameters	Pre-Training	Fine-Tuning	Linear-Probing
Optimizer	AdamW [26]	AdamW [26]	LARS [27]
Base learning rate	1.5e-4	1e-4	0.1
Weight decay	0.05	1.5e-3	1.5e-3
Optimizer momentum	(0.9,0.95) [28]	(0.9,0.99) [28]	0.9
Batch size	512	256	2048
Learning rate schedule	Cosine decay [29]	Cosine decay [29]	Cosine decay [29]
Training epochs	800	400	100
Warm-up epochs	20	10	10

allows for the reconstruction of only a subset of the masked tokens, significantly enhancing pre-training efficiency. In our experiments, we fixed the number of decoded tokens in each branch to 98 for all experiments (50% prediction ratio of all tokens from a 224×224 image with 16×16 patches). (ii) Rather than only leveraging the latent feature from the final encoder layer, our method incorporates low-level intermediate encoder feature maps, improving representation learning. Each decoder layer combines these feature maps with a weighted mechanism by applying a 1×1 convolution layer to the stacked encoder outputs. These components collaborate together to achieve better feature learning efficiently. The full pre-training pipeline of our Wave-MAE is shown in Fig. 3.

C. Two-step Tuning Protocol

During the fine-tuning stage, the pre-trained Wave-MAE encoder weights are used to initialize the backbone. At this stage, the HF and LF features are concatenated as input, and no masking is applied. For classification, global average pooling is performed on the output from the final backbone layer, which is then passed through a randomly initialized linear layer. The network is optimized using cross-entropy loss between the predictions and ground truth, adjusting the weights of the entire network.

When adapting pre-trained models to specific tasks, two common strategies are used: fine-tuning (FT) and linear probing (LP). FT updates the entire network to adapt the pre-trained features to a specific task, while LP updates only the task head, preserving the pre-trained features. Previous studies on natural images [30] suggest that FT performs better on in-distribution datasets, while LP is more effective on out-of-distribution datasets when pre-trained features are robust.

TABLE 3: Downstream tasks performance comparison of different self-supervised methods pre-trained in ChestX-Ray14.

Method	Pneumonia X-ray [31]	NIH Shenzhen CXR [32]
SimMIM [20]	95.41 ± 0.21	94.23 ± 0.40
DINO [21]	93.68 ± 0.23	92.76 ± 0.14
MAE [12]	95.89 ± 0.14	93.04 ± 0.11
SimCLRV2 [22]	94.13 ± 0.42	91.90 ± 0.36
MoCoV3 [23]	92.08 ± 0.37	92.57 ± 0.20
BYOL [24]	90.75 ± 0.44	92.01 ± 0.10
PixMIM [25]	95.97 ± 0.10	94.25 ± 0.17
Wave-MAE$^{\triangledown}$	96.12 ± 0.11	94.67 ± 0.49
Wave-MAE	**97.24 ± 0.12**	**95.53 ± 0.25**

To address varying levels of domain shift in medical image classification, we propose a two-step tuning strategy: first, apply LP to update only the classification head, followed by FT to update the entire network. This approach enhances performance on both in-domain and out-of-domain datasets. Detailed configurations are listed in TAB. 1.

III. Experiment

A. Experimental Datasets

We utilize three publicly available Chest X-Ray datasets to evaluate our model: ***(1) ChestX-Ray14*** from NIH contains 112,120 labeled frontal-view X-rays from 30,805 patients [19]. ***(2) Pneumonia X-Ray*** includes 5,856 images, with 1583 normal and 4273 pneumonia cases [31]. ***(3) NIH Shenzhen CXR*** has 662 images, with 326 normal and 336 tuberculosis cases [32]. These datasets contain various challenges like class imbalance and domain shift, making them ideal for evaluating the true performance of the proposed framework. We use the official data split method for ChestX-Ray14. For Pneumonia X-ray or NIH Shenzhen CXR, we randomly divide the dataset into 70% for training, 20% for testing, and 10% for validating. Similiar to [33], we use all unlabeled data from the ChestX-Ray14 to pre-train our framework to prevent potential data leakage between the proxy tasks and downstream tasks.

B. Implementation Details

We use the ViT-S/16 version as the backbone for all comparison experiments and try our best to reproduce the other benchmarking methods using the MMSelfSup 1.0 Platform

TABLE 4: Ablation study of the **Wave-MAE** in ChestX-Ray14. We evaluate the performance of the model under different pre-training settings. **Bold** text indicates the best performance, while underlined text denotes the default settings of our framework.

Wavelet Bases	mAUC.(%)
Haar	80.52 ± 0.27
Bior 1.5	81.15 ± 0.18
Db2	**81.38 ± 0.08**

Prediction Ratio	mAUC.(%)
25%	79.25 ± 0.20
50%	81.38 ± 0.08
75%	**81.43 ± 0.10**

# Feature Maps Fused	mAUC.(%)
1	80.56 ± 0.23
6	81.02 ± 0.30
12	**81.38 ± 0.08**

# Decoder Depth	mAUC.(%)
4	80.67 ± 0.10
8	81.11 ± 0.12
12	**81.38 ± 0.08**

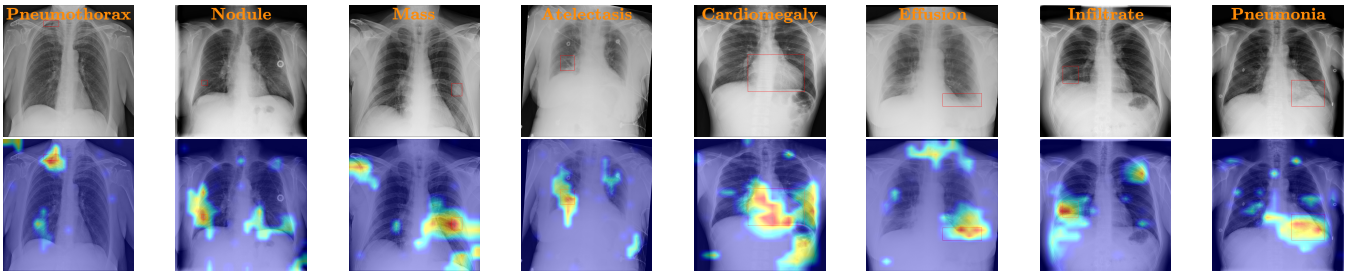

Fig. 4: **A series of Grad-CAM visualizations generated by Wave-MAE on the ChestX-Ray14 dataset. The red boxes mark the lesion locations, while the red areas in the saliency maps highlight the regions most influential to the prediction.**

[34] to ensure fairness. To avoid randomness, all the results are presented as the mean ± standard deviation across three experimental runs. All the image size is cropped and resized to 224 × 224 during the pre-training and fine-tuning stage. As for data augmentation, we apply random resized cropping (0.5, 1.0) and horizontal flipping during the pre-training stage. In the fine-tuning stage, we use RandAug (9, 0.5) [35] and set DropPath [36] rate as 0.2. All the performance will be evaluated based on the Area Under the Receiver Operating Characteristic Curve (AUC). Our implementation is built in PyTorch, and all experiments are conducted on four NVIDIA GeForce RTX 4090 Ti GPUs, each with 64 GB of memory.

C. Quantitative Results

To thoroughly evaluate the effectiveness of the proposed modules, we present the results of **Wave-MAE** along with one variant: **Wave-MAE**$^{\triangledown}$. Specifically, **Wave-MAE** refers to our model pre-trained on the ChestX-Ray14, followed by proposed two-step tuning on downstream tasks. In contrast, **Wave-MAE**$^{\triangledown}$ means only applying standard fine-tuning after pre-training. As shown in TAB. 2 and 3, on the one hand, **Wave-MAE** consistently outperforms other strong self-supervised baselines (such as SimMIM, DINO, MAE, SimCLRv2, MoCoV3, PixMIM, and BYOL) trained on the same backbone in both in-domain and out-of-domain downstream datasets in terms of fine-tuning performance. This demonstrates that **Wave-MAE** effectively leverages intrinsic HF and LF features and synergizes with the other proposed modules to enhance the quality of representation learning, thereby improving performance on downstream tasks. On the other hand, when comparing **Wave-MAE** and **Wave-MAE**$^{\triangledown}$, the differences in performance on the in-domain downstream dataset is negligible. However, in out-of-domain downstream datasets, **Wave-MAE** substantially surpasses **Wave-MAE**$^{\triangledown}$, highlighting the merits of two-step tuning strategy, particularly in mitigating the performance degradation caused by domain shift.

D. Ablation Studies

We further conduct comprehensive ablation studies on wavelet bases, prediction ratios, the number of feature maps fused, and the depth of the decoder, as detailed in TAB. 4. We identify Db2 as the most suitable wavelet transformation method, and to balance performance and efficiency, we select a prediction ratio of 50%. Regarding the decoder input, we observe that relying solely on the final layer (i.e., only one feature map fused) is insufficient, as its performance is significantly lower than incorporating a multi-layer fusion mechanism. The fusion of feature maps across all 12 layers and the use of a 12-layer decoder both result in the best performance, configurations we have adopted in our model.

E. Interpretability Analysis

To facilitate interpretability, we also utilize the Grad-CAM technique [37] to highlight key areas of the input image that contributed to the model's predictions. As shown in Fig. 4, the saliency maps in the ChestX-Ray14 reveal that the regions emphasized by our model align well with the true lesion sites.

IV. Conclusion

In this paper, we introduce Wave-MAE, a novel wavelet-based self-supervised framework specifically designed for medical image classification. Our framework enhances feature representation significantly by leveraging both high- and low-frequency features during the pre-training stage. The decoder, utilizing a cross-layer interaction mechanism, reconstructs a subset of masked tokens by using cross-attention between masked and visible tokens, thereby further enhancing the feature learning efficiently. Moreover, a two-step tuning strategy is proposed to optimize the model for downstream tasks, effectively mitigating domain shifts and improving performance on out-of-domain datasets. Overall, Wave-MAE establishes a new benchmark for self-supervised learning in medical imaging, offering improved generalization and robust performance.

REFERENCES

[1] A. Dosovitskiy, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[2] L. Cai, J. Gao, and D. Zhao, "A review of the application of deep learning in medical image classification and segmentation," *Annals of translational medicine*, vol. 8, no. 11, 2020.

[3] F. Shamshad, S. Khan, S. W. Zamir, M. H. Khan, M. Hayat, F. S. Khan, and H. Fu, "Transformers in medical imaging: A survey," *Medical Image Analysis*, p. 102802, 2023.

[4] A. Parvaiz, M. A. Khalid, R. Zafar, H. Ameer, M. Ali, and M. M. Fraz, "Vision transformers in medical computer vision—a contemplative retrospection," *Engineering Applications of Artificial Intelligence*, vol. 122, p. 106126, 2023.

[5] Y. H. Khor, V. Cottin, A. E. Holland, Y. Inoue, V. M. McDonald, J. Oldham, E. A. Renzoni, A. M. Russell, M. E. Strek, and C. J. Ryerson, "Treatable traits: a comprehensive precision medicine approach in interstitial lung disease," *European Respiratory Journal*, vol. 62, no. 1, 2023.

[6] X. Liu, F. Zhang, Z. Hou, L. Mian, Z. Wang, J. Zhang, and J. Tang, "Self-supervised learning: Generative or contrastive," *IEEE transactions on knowledge and data engineering*, vol. 35, no. 1, pp. 857–876, 2021.

[7] A. Chowdhury, J. Rosenthal, J. Waring, and R. Umeton, "Applying self-supervised learning to medicine: review of the state of the art and medical implementations," in *Informatics*, vol. 8, no. 3. MDPI, 2021, p. 59.

[8] Y. Zhang, M. Li, Z. Ji, W. Fan, S. Yuan, Q. Liu, and Q. Chen, "Twin self-supervision based semi-supervised learning (ts-ssl): Retinal anomaly classification in sd-oct images," *Neurocomputing*, vol. 462, pp. 491–505, 2021.

[9] K. S. Shakya, A. Alavi, J. Porteous, P. K, A. Laddi, and M. Jaiswal, "A critical analysis of deep semi-supervised learning approaches for enhanced medical image classification," *Information*, vol. 15, no. 5, p. 246, 2024.

[10] Y. Wang, H. Chen, Y. Fan, W. Sun, R. Tao, W. Hou, R. Wang, L. Yang, Z. Zhou, L.-Z. Guo *et al.*, "Usb: A unified semi-supervised learning benchmark for classification," *Advances in Neural Information Processing Systems*, vol. 35, pp. 3938–3961, 2022.

[11] Q. Ma, J. Gao, B. Zhan, Y. Guo, J. Zhou, and Y. Wang, "Rethinking safe semi-supervised learning: Transferring the open-set problem to a close-set one," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 16 370–16 379.

[12] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 000–16 009.

[13] Y. Zhou, J. Huang, C. Wang, L. Song, and G. Yang, "Xnet: Wavelet-based low and high frequency fusion networks for fully-and semi-supervised semantic segmentation of biomedical images," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 21 085–21 096.

[14] G. Cincotti, G. Loi, and M. Pappalardo, "Frequency decomposition and compounding of ultrasound medical images with wavelet packets," *IEEE transactions on medical imaging*, vol. 20, no. 8, pp. 764–771, 2001.

[15] R. Bachmann, D. Mizrahi, A. Atanov, and A. Zamir, "Multimae: Multi-modal multi-task masked autoencoders," in *European Conference on Computer Vision*. Springer, 2022, pp. 348–367.

[16] J. Devlin, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[17] A. Vaswani, "Attention is all you need," *Advances in Neural Information Processing Systems*, 2017.

[18] L. Fu, L. Lian, R. Wang, B. Shi, X. Wang, A. Yala, T. Darrell, A. A. Efros, and K. Goldberg, "Rethinking patch dependence for masked autoencoders," *arXiv preprint arXiv:2401.14391*, 2024.

[19] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2097–2106.

[20] Z. Xie, Z. Zhang, Y. Cao, Y. Lin, J. Bao, Z. Yao, Q. Dai, and H. Hu, "Simmim: A simple framework for masked image modeling," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 9653–9663.

[21] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 9650–9660.

[22] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. E. Hinton, "Big self-supervised models are strong semi-supervised learners," *Advances in neural information processing systems*, vol. 33, pp. 22 243–22 255, 2020.

[23] X. Chen, S. Xie, and K. He, "An empirical study of training self-supervised vision transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 9640–9649.

[24] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar *et al.*, "Bootstrap your own latent-a new approach to self-supervised learning," *Advances in neural information processing systems*, vol. 33, pp. 21 271–21 284, 2020.

[25] Y. Liu, S. Zhang, J. Chen, K. Chen, and D. Lin, "Pixmim: Rethinking pixel reconstruction in masked image modeling," *arXiv preprint arXiv:2303.02416 , year=2023*.

[26] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[27] Y. You, I. Gitman, and B. Ginsburg, "Large batch training of convolutional networks," *arXiv preprint arXiv:1708.03888*, 2017.

[28] M. Chen, A. Radford, R. Child, J. Wu, H. Jun, D. Luan, and I. Sutskever, "Generative pretraining from pixels," in *International conference on machine learning*. PMLR, 2020, pp. 1691–1703.

[29] I. Loshchilov and F. Hutter, "Sgdr: Stochastic gradient descent with warm restarts," *arXiv preprint arXiv:1608.03983*, 2016.

[30] A. Kumar, A. Raghunathan, R. Jones, T. Ma, and P. Liang, "Fine-tuning can distort pretrained features and underperform out-of-distribution," *arXiv preprint arXiv:2202.10054*, 2022.

[31] D. Kermany, K. Zhang, M. Goldbaum *et al.*, "Labeled optical coherence tomography (oct) and chest x-ray images for classification," *Mendeley data*, vol. 2, no. 2, p. 651, 2018.

[32] S. Jaeger, S. Candemir, S. Antani, Y.-X. J. Wáng, P.-X. Lu, and G. Thoma, "Two public chest x-ray datasets for computer-aided screening of pulmonary diseases," *Quantitative imaging in medicine and surgery*, vol. 4, no. 6, p. 475, 2014.

[33] H. Yu and Q. Dai, "Self-supervised multi-task learning for medical image analysis," *Pattern Recognition*, vol. 150, p. 110327, 2024.

[34] M. Contributors, "MMSelfSup: Openmmlab self-supervised learning toolbox and benchmark," https://github.com/open-mmlab/mmselfsup, 2021.

[35] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "Randaugment: Practical automated data augmentation with a reduced search space," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2020, pp. 702–703.

[36] G. Huang, Y. Sun, Z. Liu, D. Sedra, and K. Q. Weinberger, "Deep networks with stochastic depth," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*. Springer, 2016, pp. 646–661.

[37] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.