Wavefield MAE: leveraging large vision models for ultrasonic wavefield pattern analysis

Jiaxing Ye†, Takumi Kobayashi†, Nobuyuki Toyama†
†National Institute of Advanced Industrial Science and Technology, Japan

***Abstract*—This paper delves into the adaptation of large vision models, initially pretrained on extensive natural image datasets, for the analysis of physical signals. Inspired by their remarkable generalization and transferability across diverse computer vision tasks, we leverage the large vision models for analyzing ultrasonic wavefield images captured via physics-based imaging, which exhibit different characteristics from natural images. To bridge the substantial gap between the original domain of vast natural images and the target wavefield patterns, we introduce the wavefield MAE model featured with a two-stage adaptation process: self-supervised learning using a reference wavefield image dataset, followed by finetuning to the wavefield classification task. The proposed scheme progressively refines (visual) feature representation toward wavefield patterns by consolidating tripartite information of 1) natural image patterns in pretraining, 2) Simulated wavefield data that mirror physics principles of wave propagation, and 3) real wavefield data of ultrasonic testing (UT). Comprehensive experimental comparisons have validated the proposed feature adaptation scheme. Notably, the proposed scheme is not confined to ultrasonic wavefield analysis but has a broad reach in general topics of adapting the pretrained models for specific tasks with limited data availability.**

***Index Terms*—foundation model, machine learning, pattern recognition, ultrasonics, wavefield image**

I. Introduction

In recent years, the computer vision research field has witnessed a paradigm shift with an emerging flow in that the models are pretrained on a large-scale dataset and then adapted, so-called *finetuned*, to downstream tasks [1], [3], [19], [30]. With the support of large-scale datasets and massive computing resources, the pre-trained *vision* model can provide favorable feature representations to facilitate downstream tasks, such as classification [30], detection [6], and semantic segmentation [7]. These successes imply that extensive pretraining hones visual features of generalization capability. Thereby, the subsequent finetuning phase effectively refines the model for a specific task guided by a small-scale in-domain dataset, without the need to tune the features thoroughly [8].

Furthermore, advancements in vision models have unlocked new opportunities across a wide array of scientific and engineering fields [2]. This study seeks to investigate the utility of the advanced vision models for physics signal analysis of ultrasonic wavefield images, which refer to the spatial and temporal distributions of ultrasonic waves as they propagate through a medium [5]. Ultrasonic wavefield images are widely used in non-destructive testing and structural health monitoring to detect and characterize defects in materials [9], [25]. The analysis of ultrasonic wavefield images is challenging, as it requires to unravel intricate visual patterns that are often influenced by the underlying factors of physical structures, specimen materials, measurement devices, and testing environment [13]. Conventional signal processing-based approaches thus encounter difficulties when modeling complex and diversified wavefield patterns [10]. We hypothesize that large vision models, pretrained on vast image datasets, have the potential to render feature representations useful even for describing physical signal patterns, such as spatial frequency and phase, which are fundamental for characterizing ultrasonic wavefields. It, however, is difficult to directly apply the large vision models to ultrasonic wavefield analysis due to the stark differences in data characteristics between natural images and physics-based images. For example, natural images vary greatly in shapes, textures, angles, and lighting conditions, while ultrasonic wavefield images primarily exhibit spectral patterns caused by the scattering of ultrasonic waves through the medium. This significant difference calls for an efficient feature representation learning to fill the gap between the two visual domains.

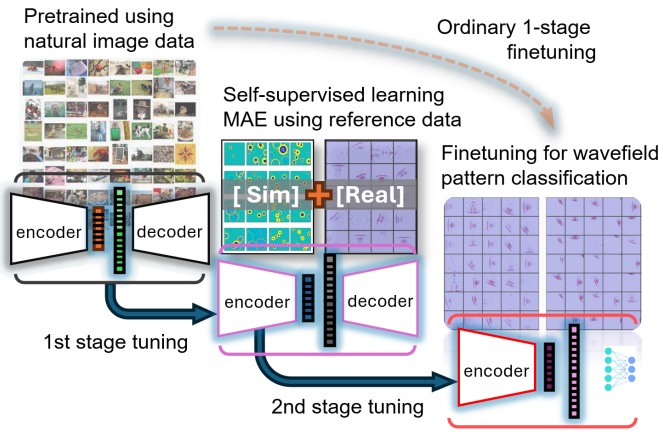

Fig. 1. The proposed wavefield MAE model architecture with two-stage progressive feature adaptation. At the 1st stage, the self-supervised learning (SSL) with masked autoencoder (MAE) tailors the pre-trained large vision model to represent ultrasonic wavefield patterns, followed by the 2nd-stage finetuning on a wavefield pattern classification task.

We propose a strategy to progressively adapt a large vision model pretrained on vast natural images for ultrasonic wavefield analysis. The overall process is illustrated in Fig. 1. Unlike the commonly applied one-stage finetuning [1], [19], this study introduces a *two-stage* adaptation scheme to augment the model's ability to specialize physical wavefield

patterns which are distinct from natural images as discussed above. In detail, our approach leverages a *reference ultrasonic wavefield dataset* to facilitate progressive feature adaptation, which is composed of both *simulated* and real measurement data without any supervision. It provides a collection of visual patterns that mirror the fundamental physical principles of ultrasonic wave propagation and convey extensive information regarding the real measurement conditions, such as characteristics of the device used, the specimen under examination, and the testing environment. By applying self-supervised learning (SSL) through a masked autoencoder (MAE) to this reference wavefield data, we are able to adapt the visual features to suit the specific context of practical ultrasonic wavefield inspection *without annotation*. Then, the model is finally finetuned on a *labeled real* dataset in a task-specific way. In this work, we cope with wavefield pattern classification task to differentiate defective patterns from defect-free ultrasonic inspection wavefield patterns. Our contributions of this paper are as follows:

- We propose a two-stage adaptation scheme to leverage the large vision model for physical pattern analysis of ultrasonic wavefield. The proposed scheme successfully incorporates three data sources to foster effective adaptation; natural images in pretraining, *physics simulation* data in the reference dataset and real measurement data of ultrasonic testing.
- A key process in the two-stage adaptation is the MAE-based SSL at the 1st-stage tuning to take advantage of both simulation and real measurement data of ultrasonic wavefield images without any annotation. It fills the gap between natural images in pre-training and the objective wavefields for the task.
- Extensive experimental comparison validates that the proposed general-to-specific adaptation scheme yields superior performance through comprehensive ablation studies to highlight the critical factors for performing task adaptation.

II. Related work

A. Large-scale pretraining with self-supervised learning

As a pivotal technique driving AI progress today, self-supervised learning (SSL) effectively trains deep models without supervised annotation [16], being categorized under unsupervised learning which strives to automatically uncover the hidden structures in the input data. It is a powerful tool that efficiently utilizes large datasets to learn an feature representation transferrable to downstream tasks; for image-based tasks, the most common "upstream" dataset is ImageNet [21] as demonstrated in recent works addressing the vision tasks [26]. Contrastive learning (CL) was the first SSL model that achieved comparable or even better accuracy than supervised pre-training [27], [28]. Recently, another trend has emerged focusing on masked image modeling (MIM) [19], [29], which effectively takes advantage of vision transformers (ViTs) [30] on vision tasks to achieve favorable generalization performance. MIM using masked auto-encoder (MAE) [19] imposes a pretext task to predict masked parts of its input from other unmasked segments solely using the input data. Due to its superior performance, we adopt the MAE model for the self-supervised learning process for the 1st stage tuning.

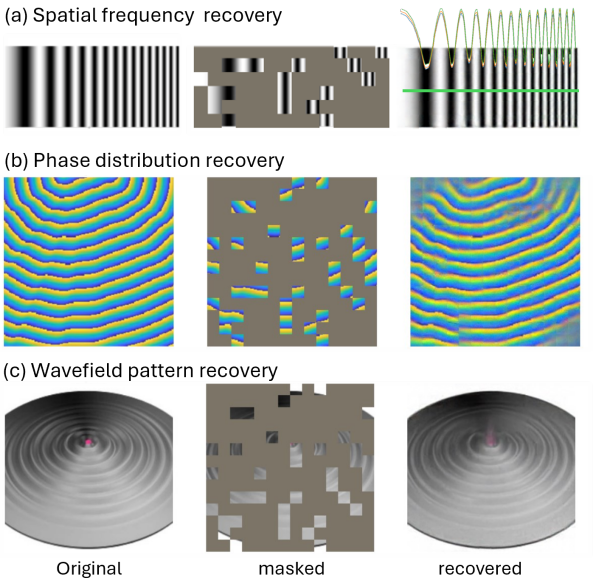

Fig. 2. Zero-shot recovery of physics signal (with 80% pixels discarded) using pretrained model on ImageNet (MAE-ViT-Large)

B. Task-specific fine-tuning

To harness the knowledge acquired during pretraining for a specific task, various finetuning methods have been extensively investigated, including direct finetuning [11], finetuning after linear probing [22], side-tuning [23] , using different learning rates for different layers [24]. It, however, remains elusive how features are adapted during finetuning under different settings.

Despite the extensive research, most efforts have concentrated on directly (one-stage) tuning the pretrained model to the target task, with little exploration of utilizing both labeled and unlabeled data in the target domain. The review reveals that we are navigating a comparatively untapped direction.

III. Wavefield MAE model

A. Vision model for physical signal recovery: a zero-shot test

Large vision models have demonstrated remarkable generalization and adaptability across various tasks [11], [19], inspiring us to explore their potential in analyzing physical patterns. To verify the feasibility of the idea, we preliminarily carried out a zero-shot test by directly applying a pretrained model to a MIM task; it visually evaluates reconstruction performance on (heavily) masked input images of physical signals. In this preliminary study, we tested three representative physics signals, spatial chirp signal, phase map, and wavefield image as shown in Fig. 2. We utilized an ImageNet-pretrained model, MAE-ViT-large [19], to reconstruct the entire image by accessing only random 20% of the pixels. The results in Fig. 2 reveal that the large vision model can mostly restore the original data even in this zero-shot setting; in Fig. 2-(a), we depicted the 1-D slit of the data series which shows successful recovery of spectral chirp structure. Furthermore, we evaluated the model's ability to recover the wavefield images, which are of key interest in this study; the results are illustrated in Fig. 3. The reconstructed wavefields are visually convincing on both

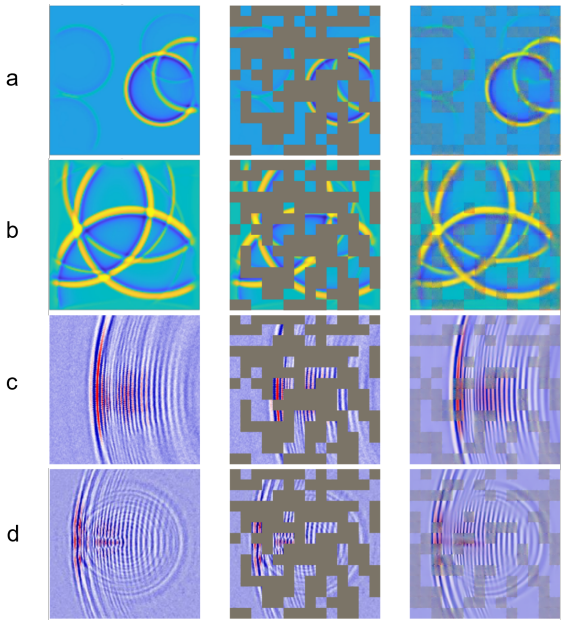

Fig. 3. Zero-shot recovery of ultrasonic wavefield images (with 70% pixels discarded) using pretrained model on ImageNet (MAE-ViT-Large [19]). [a] and [b] denote simulation data. [c] and [d] present the real ultrasonic wavefield images with/without defect presence, respectively (left to right: original image, masked image, and reconstructed image)

simulation data in Fig. 3-(a,b) and real wavefield images in Fig. 3-(c,d), though some degradation is found. These results support our hypothesis that pretrained large vision models potentially describe physical signals governed by frequency and phase. However, especially for the defective wavefield case in Fig. 3-(d), the left-half flaw-induced circular reflection is missing in the recovered image, suggesting that the pretrained model requires additional adjustments to accurately represent the geometric structure in 2D plane wave scattering. These results underscore the validity of our approach and motivate us to conduct further investigation.

B. Wavefield MAE model with two-stage feature adaptation

Given an off-the-shelf pretrained model, e.g., ViT-base [19], we progressively tune the large model toward the target task in the following two-stage manner.

The 1st stage tuning is designated to roughly align the general visual feature representations of the pretrained model to the target context of ultrasonic wavefield characterization. For that purpuse, we apply SSL using MAE on a *reference* dataset which is composed of both *simulated* and real ultrasonic wavefield data as detailed in Sec. III-C *without annotation*. MAE-SSL trains the model by reconstructing the obscured image regions (Fig. 3) from visible areas without any external supervision, which thus derives effective and robust feature representations to capture the underlying semantic spatial structures of the scattering waves. Since the knowledge regarding a target task is not yet embedded at this stage, the tuned features exhibit sufficient generalization applicable to various downstream tasks in the target domain, such as defective pattern classification [11], [19]. In addition, MAE does not require any annotations, greatly easing the integration of simulated data with real-world observations at this 1st-stage feature learning. It is a particularly valuable property in the literature of analyzing physical signals with scarce annotation.

TABLE I
ULTRASONIC WAVEFIELD IMAGE DATASETS.

(a) Two types of data

Data type	*Simulated* data	*Real* data
Data source	k-Wave toolbox [20]	USimgAIST [17]
image size	$224 \times 224 \times 3$	$224 \times 224 \times 3$
# of samples	16,000	7,004 (17 specimens)
Class label	✘	✔ (defect/non-defect)

(b) Datasets in two-stage adaptation

Dataset	Reference at 1st stage	Target at 2nd stage	
		Training	Test
# of samples	16,000 (*simulated*) + 2,000 (*real*)	∼6,592 (*real*) [16 specimens]	∼416 (*real*) [1 specimen]
Class label	✘	✔ (defect/non-defect)	

The 2nd stage tuning focuses on deploying the refined feature representations acquired from the 1st stage tuning for the target task of wavefield pattern classification. We add a classification head on top of the encoder of the MAE model and then finetune it in a standard one-stage finetuning setting. In our case of binary classification, a binary cross-entropy loss is adopted on the *real* dataset of USimgAIST [17] with class labels (defective/non-defective) in a supervised learning fashion.

These two-stage adaptation enables us to tailor the feature representation to the target wavefield analysis while filling the gap between natural images in pretraining dataset, e.g., ImageNet, and the target physical wavefield patterns.

C. Ultrasonic wavefield image dataset for defect identification

As described above, the two-stage adaptation demands two data types, simulated and real ones, which are respectively detailed as follows (Tab. I-(a)); in this work, we employ defect identification of ultrasonic wavefield as the target task.

Simulated data is generated based on physical models. In principle, Ultrasonic Testing (UT) employs an electrical pulser to generate an ultrasonic signal that travels through a specimen; once a defect is encountered, part of the wave energy will reflect back to the surface [10], [25]. Defect-induced patterns are visible in ultrasonic wavefield images through distinct flaw-induced scattering echoes that overlay the propagating waves as shown in Fig. 3-(d). We use the k-Wave toolbox [20] to generate a simulated wavefield dataset with 16,000 images showcasing ultrasonic wave scattering in a 2-D homogeneous medium. Covering a broad spectrum of parameter settings, the dataset presents variations in specimen characteristics, ultrasonic wave and transducer properties, and testing condition of temperatures. Illustrative examples are displayed in Fig. 3-(a,b).

Real data is given in the public USimgAIST dataset [17] which is obtained by using the ultrasonic wave propagation imaging system [25]; it contains 7004 images from 17 spec-

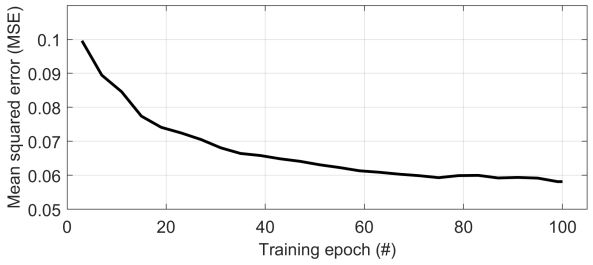

Fig. 4. Training loss of the MAE self-supervised learning process.

imens with varying defect sizes and depths, which are split into 3389 defective and 3615 non-defective images.

To facilitate the 1st stage tuning via MAE-SSL, the reference wavefield dataset is constructed by using all the simulated images and a 30% random subset of USimgAIST training images excluding labels as shown in Tab. I-(b). Notably, the simulated and real data play complementary roles. The simulated data helps the model grasp the physical patterns of wave dispersion, whereas the real data render contextual information about the measurement device used, the material characteristic of tested specimens, and interference from internal/boundary conditions, as well as the experiment environment. The two types of data encapsulate the diverse facets of the ultrasonic wavefield phenomenon that further facilitate the feature alignment from natural images to wavefield images. Then, the 2nd stage tuning characterizes the USimgAIST dataset with class labels to finetune the model for wavefield pattern classification in realistic scenarios; for example, we employ leave-one-specimen-out scheme over 17 specimens in USimgAIST for evaluation as detailed in Tab. I-(b).

IV. Experiments and Results

We apply the proposed method to a wavefield pattern classification on the USimgAIST dataset. The performance is evaluated based on four metrics of accuracy, recall rate, precision and F-score.

A. Implementation detail

We apply ImageNet-pretrained ViT-base [19] to a backbone feature extractor and tune it in the following settings.

In the *1st stage adaptation* using the MAE-SSL was carried out by using 8×A100 GPUs over 100 epochs including 20-epoch warm up with the batch size of 2048 samples of which 80% pixels are masked, and the learning rate of 1e-4. We present the training loss curve in Fig. 4.

For the *2nd stage tuning*, we train the model on a single A100 GPU over 10 epochs including 2-epoch warm up with the batch size of 256 and the learning rate of 5e-4; the other settings are the same as the ViT paper [30] such as for the loss function, optimizer, and learning rate decay scheduler. We monitored the training process and observed that the model achieved convergence over the 10 epochs without overfitting.

TABLE II
Performance comparison in tuning strategies

Strategy	**a**	**b**	**c**	**d**
Natural image	✘	✘	✔	✔
Reference (*sim+real*)	✘	✔	✘	✔
1) Leave-one-specimen-out				
Accuracy(%)	92.97	93.03	97.12	**97.66**
Recall (%)	98.04	98.11	98.71	**99.26**
Precision (%)	92.02	91.76	97.10	**97.34**
F1-score (%)	94.81	94.73	97.84	**98.24**
2) One-specimen-shot				
Accuracy(%)	52.59	82.74	89.54	**91.34**
Recall (%)	100.00	89.24	91.47	**95.16**
Precision (%)	52.58	91.12	88.95	**89.11**
F1-score (%)	68.93	88.23	90.19	**92.04**

B. Comparison in finetuning strategies

To evaluate the proposed model, we carried out experimental comparisons across four settings (Tab. II) to finetune vision models for wavefield pattern classification; without pretraining on natural images, we can apply (**a**) training the classification model from scratch with random weights initialization, and (**b**) pretraining the model via MAE-SSL on the reference wavefield dataset (Tab. I-(b)) followed by finetuning for classification, while the ImageNet-pretrained model can be applied to (**c**) one-stage (standard) finetuning for wavefield classification and (**d**) the proposed two-stage tuning scheme that incorporate both simulation and real data to progressively learn efficient feature representation for discerning defective ultrasonic wavefields.

We evaluate performances of those approaches thoroughly in the two types of cross-validation protocols which differ in the number of training samples; the leave-one-specimen-out scheme described in Sec. III-C and Tab. I-(b) and the one-specimen-shot learning which only uses wavefield images of one specimen for training in a opposite manner to the leave-one-out protocol.

Tab. II shows performance results of the above-mentioned four strategies on the two evaluation schemes. First, we compare the strategy of training from scratch (Tab. II**a**) and the standard ImageNet-pretraining (Tab. II**c**). The result is well aligned with our expectation (Sec. III-A) that the pretrained model achieves transferable visual features through massive natural images that could facilitate downstream tasks even for analyzing physical wavefield signals. On the other hand, training from scratch has difficulty especially in dealing with small-scale data; the strategy (**a**) significantly degrades performance on the one-shot scheme.

Then, we compare the models pretrained on natural images only (Tab. II**c**) and using the reference wavefield dataset (Tab. II**b**), which are two representative ways of tackling small-scale classification. The results showed that pretraining requires larger datasets and more diverse visual patterns; our reference wavefield dataset contains only 18,000 images while ImageNet consists of 1M natural images. It also aligns with the literatures [11], [30]. On the other hand, the proposed

TABLE III
THROUGHOUT PERFORMANCE COMPARISON WITH OTHER METHODS.

	AlexNet [14]	ResNet [18]	DenseNet [12]	USresNet [15]	ST-Net [9]	ST-Net-(2+1)D [9]	ViT [30]	Ours
Accuracy (%)	86.70	96.14	97.62	95.58	96.29	96.67	92.97	**97.66**
Recall (%)	94.80	91.54	95.68	96.85	97.39	96.51	98.04	**99.26**
Precision (%)	67.35	94.46	96.37	94.64	96.51	97.09	92.02	**97.34**
F-score (%)	77.51	92.20	95.85	95.50	97.09	97.20	94.81	**98.24**

strategy (Tab. II**d**) combining those two datasets exhibits superior performance, showcasing its favorable capability in the progressive feature adaptation for describing wavefield images. In particular, it works well on the scenario of scarce training samples in the one-shot scheme, while the ImageNet-pretrained model (**c**) considerably degrades performance in that scenario.

C. Performance comparison

Tab. III displays a comprehensive performance comparison among diverse methods in the leave-one-specimen-out evaluation protocol. Our tuned ViT is compared to CNNs [12], [14], [18] and ViT [30] trained from scratch as well as the other recent methods [9], [15] designated for the wavefield analysis. The proposed approach outperformed these methods, demonstrating that the proposed feature adaptation scheme effectively leverages a pretrained vision model for physical signal analysis of ultrasonic wavefields.

V. CONCLUSION

In this study, we proposed an efficient two-stage adaptation scheme to leverage pretrained large vision model for ultrasonic wavefield pattern analysis. In contrast to one-stage finetuning, we introduced an intermediate tuning stage by means of MAE-SSL on a reference dataset composed of simulated and real wavefield data. It contributes to progressively adapting the visual feature representations pretrained on natural images toward the target context of ultrasonic wavefield patterns. Comprehensive experiments validated efficacy of the proposed scheme on a wavefield pattern classification task, implying applicability of the proposed approach to the adaption of large models for specific tasks with confined data availability.

REFERENCES

[1] Khan S, Naseer M, Hayat M, et al. "Transformers in vision: A survey," ACM computing surveys (CSUR), 2022, 54(10s): 1-41.

[2] Wang, Hanchen, et al. "Scientific discovery in the age of artificial intelligence." Nature, 2023, 620.7972: 47-60.

[3] Chen, Shoufa, et al. "Adaptformer: Adapting vision transformers for scalable visual recognition," Advances in Neural Information Processing Systems, 2022, 35: 16664-16678.

[4] Liu, Ze, et al. "Swin transformer: Hierarchical vision transformer using shifted windows," Proceedings of the IEEE/CVF international conference on computer vision. 2021.

[5] Michaels, J.E. "Ultrasonic Wavefield Imaging," In: Ida, N., Meyendorf, N. (eds) Handbook of Advanced Nondestructive Evaluation. Springer, Cham. 2019

[6] Li, Yanghao, et al. "Exploring plain vision transformer backbones for object detection," European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022.

[7] Cheng, Bowen, et al. "Masked-attention mask transformer for universal image segmentation," Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.

[8] Pinto, André Susano, et al. "Tuning computer vision models with task rewards," International Conference on Machine Learning. PMLR, 2023.

[9] Ye J, Toyama N. "Automatic defect detection for ultrasonic wave propagation imaging method using spatio-temporal convolution neural networks," Structural Health Monitoring, 2022, 21(6): 2750-2767.

[10] Michaels J E. "Ultrasonic wavefield imaging: Research tool or emerging NDE method?" AIP Conference Proceedings. 2017, 1806(1).

[11] Steiner, Andreas, et al. "How to train your vit? data, augmentation, and regularization in vision transformers," Transactions on Machine Learning Research (TMLR), 2022. 5.

[12] Huang G, Liu Z, et al. "Densely connected convolutional networks" Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 4700-4708.

[13] Herzfeld, Karl F., and Theodore A. Litovitz. "Absorption and dispersion of ultrasonic waves," 2013, Vol. 7. Academic Press

[14] Krizhevsky A, Sutskever I and Hinton GE. "Imagenet classification with deep convolutional neural networks," Advances in neural information processing systems.25: 1097-1105. 2012

[15] Ye J, Ito S, Toyama N. "Computerized ultrasonic imaging inspection: From shallow to deep learning," Sensors, 2018, 18(11): 3820.

[16] Liu, Xiao, et al. "Self-supervised learning: Generative or contrastive." IEEE transactions on knowledge and data engineering, 2021, 35.1: 857-876.

[17] Ye J, Toyama N. "Benchmarking deep learning models for automatic ultrasonic imaging inspection," IEEE Access, 2021, 9: 36986-36994.

[18] He K, Zhang X, Ren S et al. "Deep residual learning for image recognition," In Proceedings of the IEEE conference on computer vision and pattern recognition. 2015, pp. 770-778.

[19] He, Kaiming, et al. "Masked autoencoders are scalable vision learners." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.

[20] Treeby B E, Jaros J, Rendell A P, et al. "Modeling nonlinear ultrasound propagation in heterogeneous media with power law absorption using ak-space pseudospectral method," The Journal of the Acoustical Society of America, 2012, 131(6): 4324-4336.

[21] Deng, Jia, et al. "Imagenet: A large-scale hierarchical image database." IEEE conference on computer vision and pattern recognition. 2009.

[22] Kumar, Ananya, et al. "Fine-Tuning can Distort Pretrained Features and Underperform Out-of-Distribution." International Conference on Learning Representations. 2022.

[23] Zhang J O, Sax A, Zamir A, et al. "Side-tuning: a baseline for network adaptation via additive side networks" 16th European Conference on Computer Vision, Proceedings, 2020: 698-714.

[24] Tianyi Zhang, Felix Wu, Arzoo Katiyar, Kilian Q Weinberger, and Yoav Artzi. "Revisiting few-sample bert fine-tuning," In International Conference on Learning Representations, 2020b.

[25] Yashiro, S., Takatsubo, J., Miyauchi, H., Toyama, N. "A novel technique for visualizing ultrasonic waves in general solid media by pulsed laser scan," NDT E Int. 2008, 41, 137-144

[26] Goyal, Sachin, et al. "Finetune like you pretrain: Improved finetuning of zero-shot vision models," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

[27] He K, Fan H, Wu Y, et al. "Momentum contrast for unsupervised visual representation learning", Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 9729-9738.

[28] Caron, Mathilde, et al. "Emerging properties in self-supervised vision transformers." Proceedings of the IEEE/CVF international conference on computer vision. 2021.

[29] Bao, H., Dong, L., and Wei, F. "Beit: Bert pre-training of image transformers," ArXiv, abs/2106.08254, 2021.

[30] Dosovitskiy, et al. "An image is worth 16x16 words: Transformers for image recognition at scale," In Int. Conf. Learn. Represent., 2020.