Physics-Guided Prune-then-finetune of Vision Transformers for Wavefield Pattern Analysis

Jiaxing Ye[1], Takumi Kobayashi[1,2]

1. National Institute of Advanced Industrial Science and Technology
2. University of Tsukuba, Japan

{jiaxing.you, takumi.kobayashi}@aist.go.jp

Contents overview:

1. Background: Model scaling and pretrain-finetune paradigm

2. Application: Ultrasonic wavefields pattern recognition

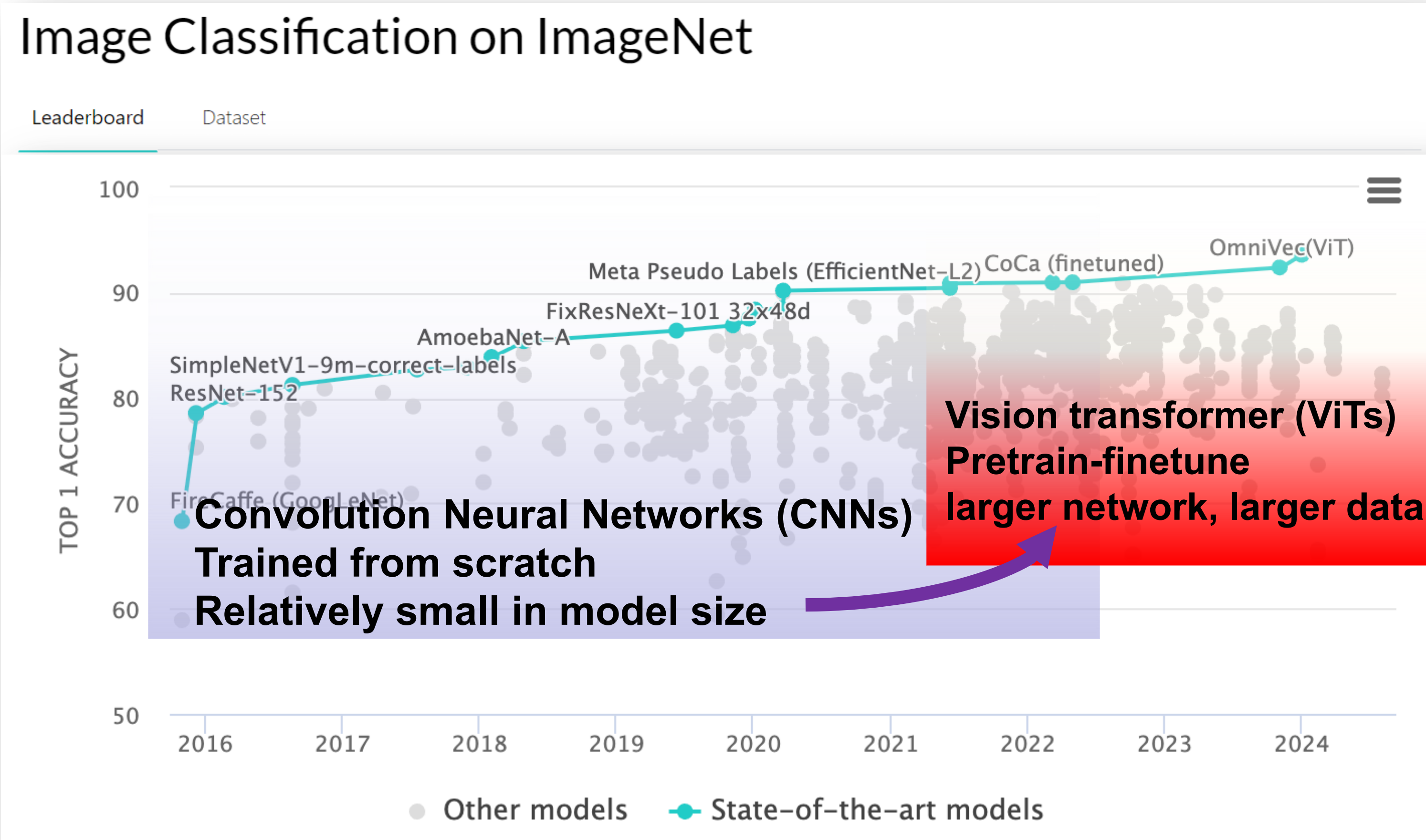
3. Method: prune-then-finetune of ViTs for physics data

4. Evaluation results and conclusions

1. Background: Model scaling and pretrain-finetune paradigm

Paradigm shift since 2022:

more capable model, pretrain-finetune fashion, scaling up



1. Background: Model scaling and pretrain-finetune paradigm

Core techniques/findings fueled the Paradigm shift

1. **Vision transformer (ViTs)**
2. **Pretrain-finetune**
3. **Scaling law for vision**

Vision transformer backbone

DOSOVITSKIY, Alexey, et.al. ***An image is worth 16x16 words: Transformers for image recognition at scale.*** ***ICLR 2020***

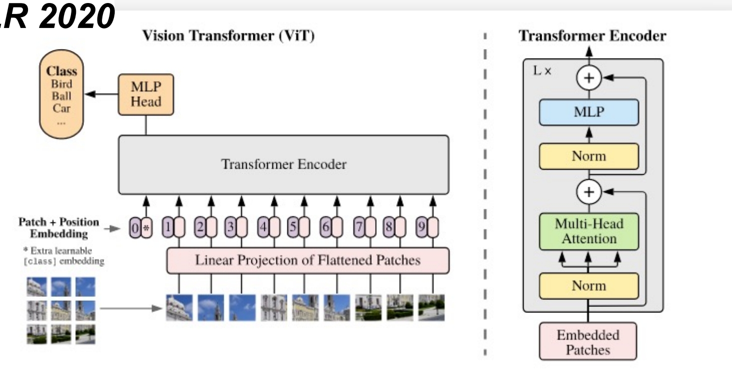
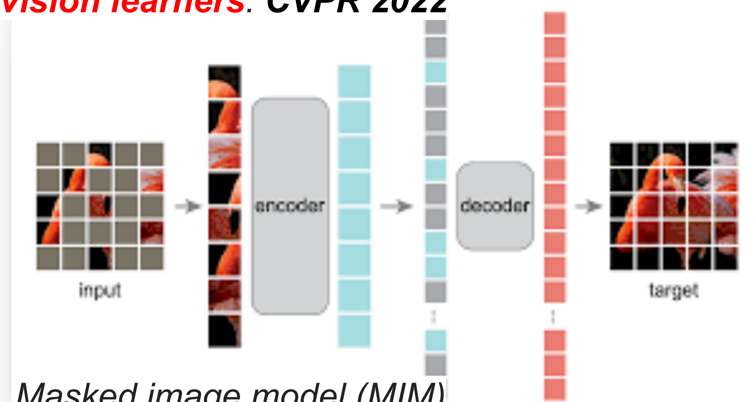


Figure 1: Model overview. We split an image into fixed-size patches, linearly embed each of them, add position embeddings, and feed the resulting sequence of vectors to a standard Transformer encoder. In order to perform classification, we use the standard approach of adding an extra learnable "classification token" to the sequence. The illustration of the Transformer encoder was inspired by Vaswani et al. (2017).

Pretraining with self-supervised learning(SSL) *utilizing the web-scale images without labels*

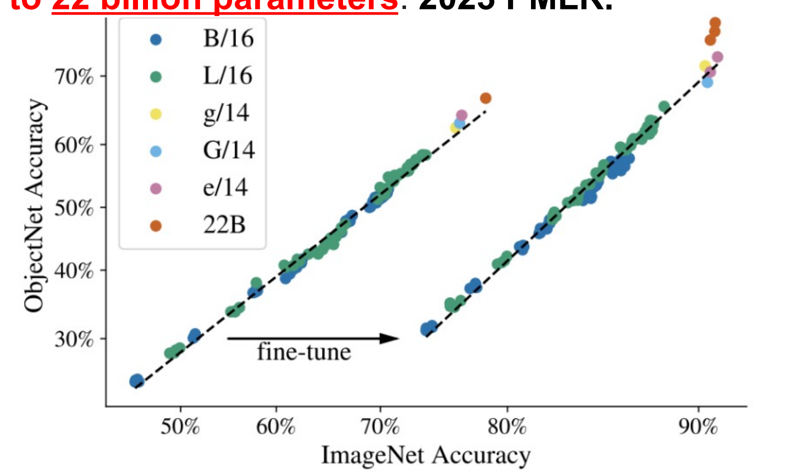
He, K., et. al., ***Masked autoencoders are scalable vision learners****.* ***CVPR 2022***



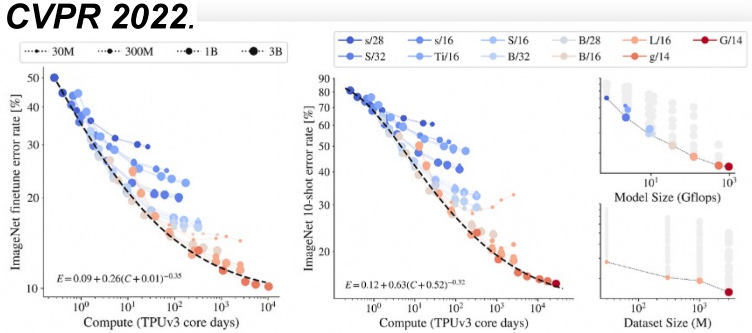
Masked image model (MIM)

Finetuning for downstream task

Dehghani, M., et. al. **Scaling vision transformers to 22 billion parameters. 2023 PMLR.**

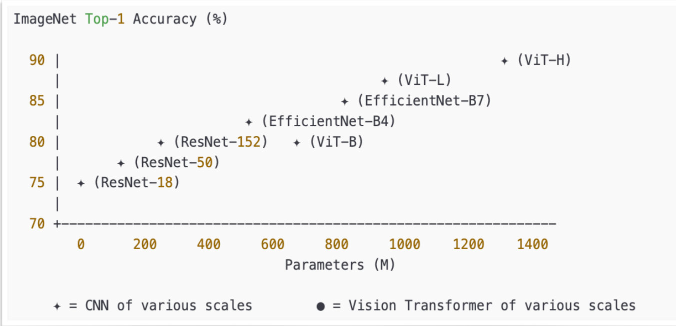


Zhai X, et. al. ***Scaling vision transformers.*** ***CVPR 2022.***



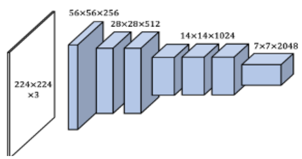
1. Background: Model scaling and pretrain-finetune paradigm

Scaling laws in vision model: Bigger models + more data + massive pretraining → better transfer, and that error decays as a power law.



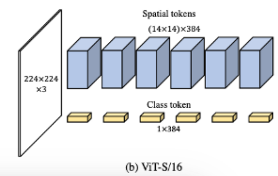
ImageNet1K dataset: This dataset spans 1000 object classes and contains 1,281,167 training images, 50,000 validation images and 100,000 test images

Typical CNN models



Method	Depth	Params
ResNet [11]	110	1.7M
ResNet (reported by [13])	110	1.7M
ResNet with Stochastic Depth [13]	110	1.7M
	1202	10.2M
Wide ResNet [41]	16	11.0M
	28	36.5M
with Dropout	16	2.7M
ResNet (pre-activation) [12]	164	1.7M
	1001	10.2M
DenseNet ($k = 12$)	40	1.0M
DenseNet ($k = 12$)	100	7.0M
DenseNet ($k = 24$)	100	27.2M
DenseNet-BC ($k = 12$)	100	0.8M
DenseNet-BC ($k = 24$)	250	15.3M
DenseNet-BC ($k = 40$)	190	25.6M

Typical vision transformers



Model	Layers	Hidden size D	MLP size	Heads	Params
ViT-Base	12	768	3072	12	86M
ViT-Large	24	1024	4096	16	307M
ViT-Huge	32	1280	5120	16	632M

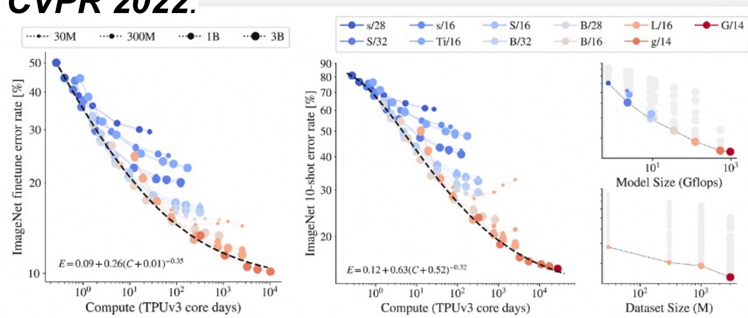
Table 1: Details of Vision Transformer model variants.

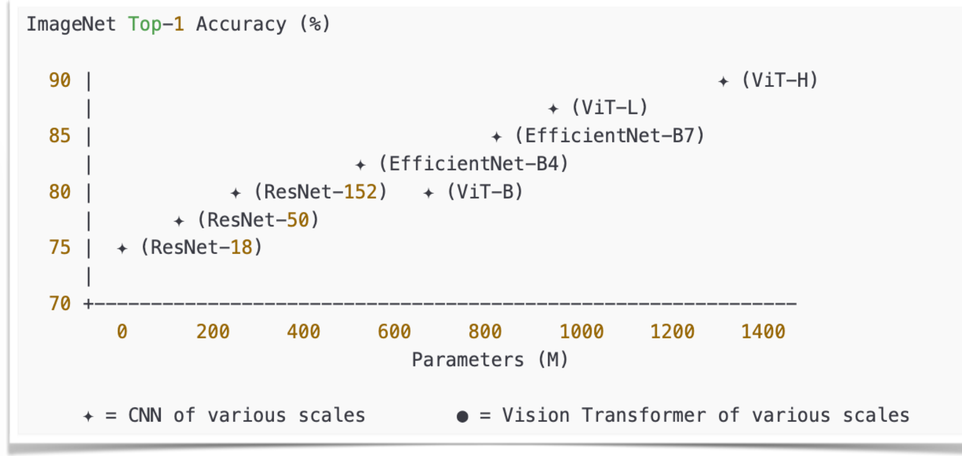
Table 1: ViT-22B model architecture details.

Name	Width	Depth	MLP	Heads	Params [M]
ViT-G	1664	48	8192	16	1843
ViT-e	1792	56	15360	16	3926
ViT-22B	6144	48	24576	48	21743

1. Background: Model scaling and pretrain-finetune paradigm

Zhai X, et. al. ***Scaling vision transformers. CVPR 2022.***



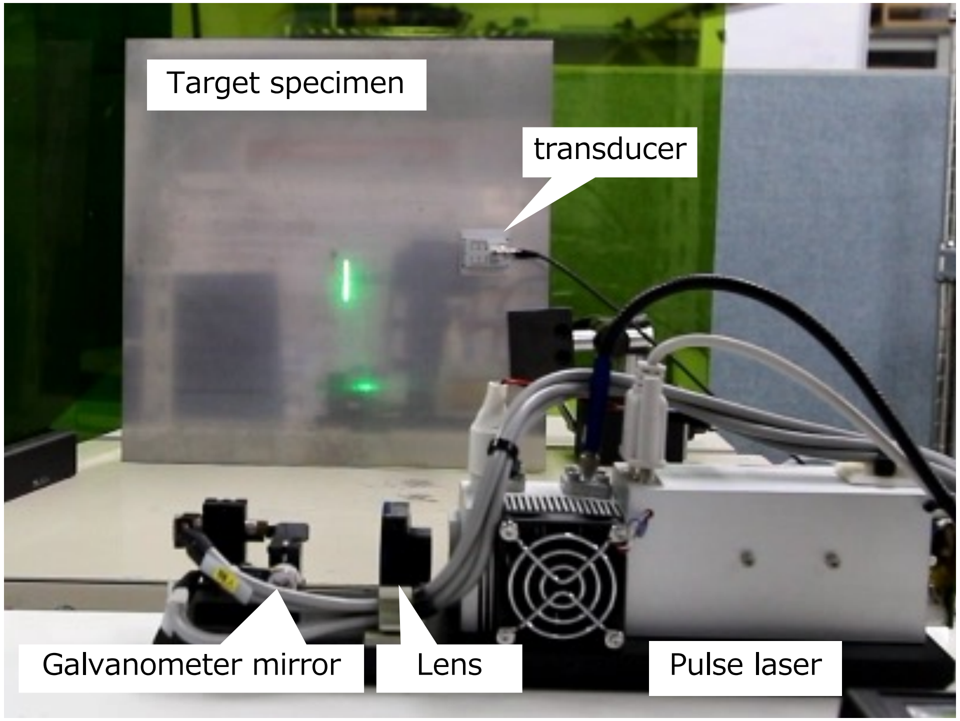


The major challenges brought by the pretrain-finetune paradigm:

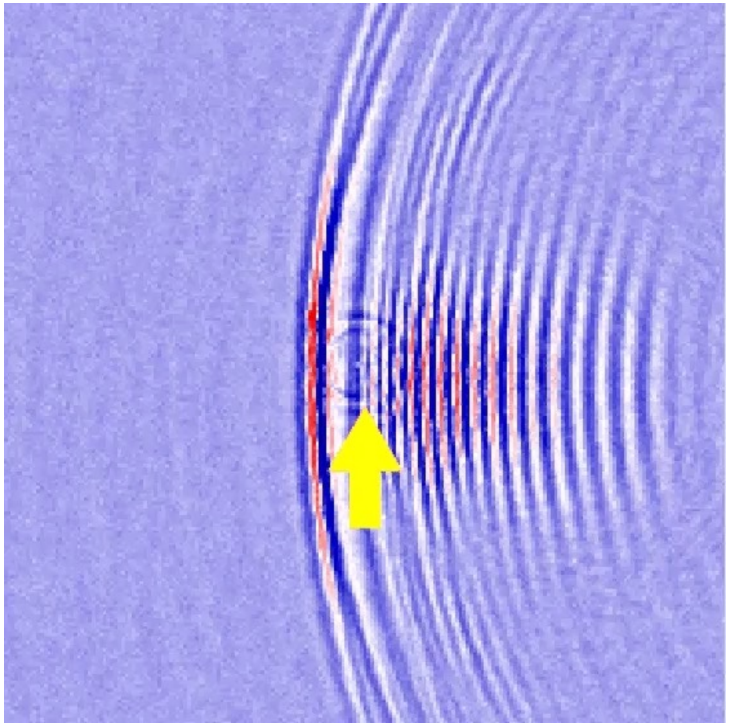
1. **Domain gaps**: the target downstream dataset can be different from the pretrain cases.
2. **Data scarcity**: In-domain data is rather limited in scale
3. **High computation cost:** Larger models demand for more computations for both finetuning and deployment stages.

2. Application: Ultrasonic wavefields pattern recognition

In our research institute, the *Non-destructive evaluation research group, AIST, Japan*, developed Ultrasonic Wave Propagation Imaging (UWPI) System for defect detection with extensive industrial applications in vehicle engineering, airplane safety testing, and welding assessment, etc.

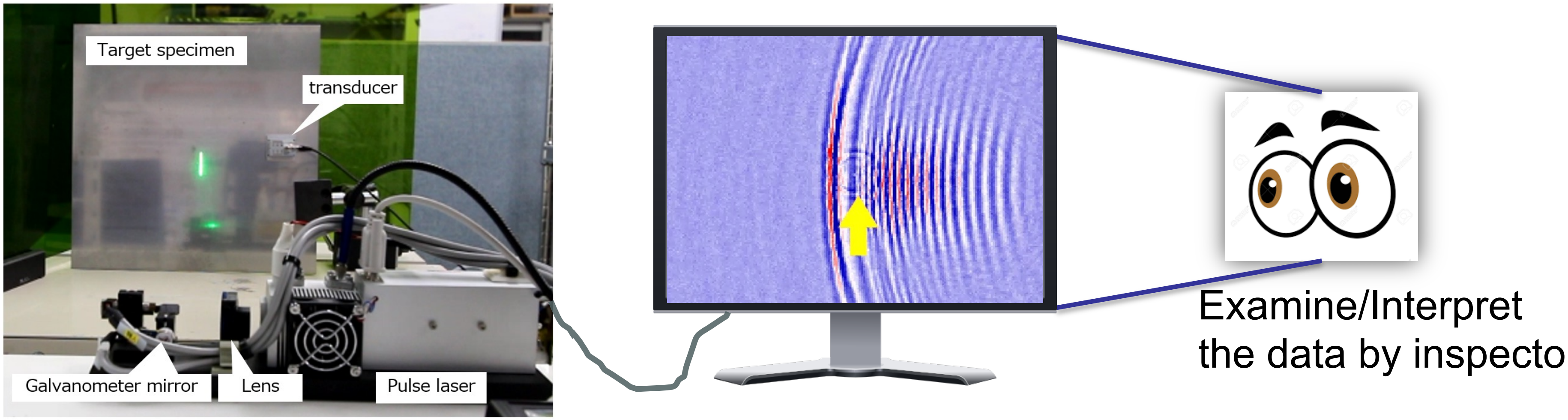


Ultrasonic propagation visualization system for non-destructive evaluation



Detect-induced echo visualization [Superposition of waves]

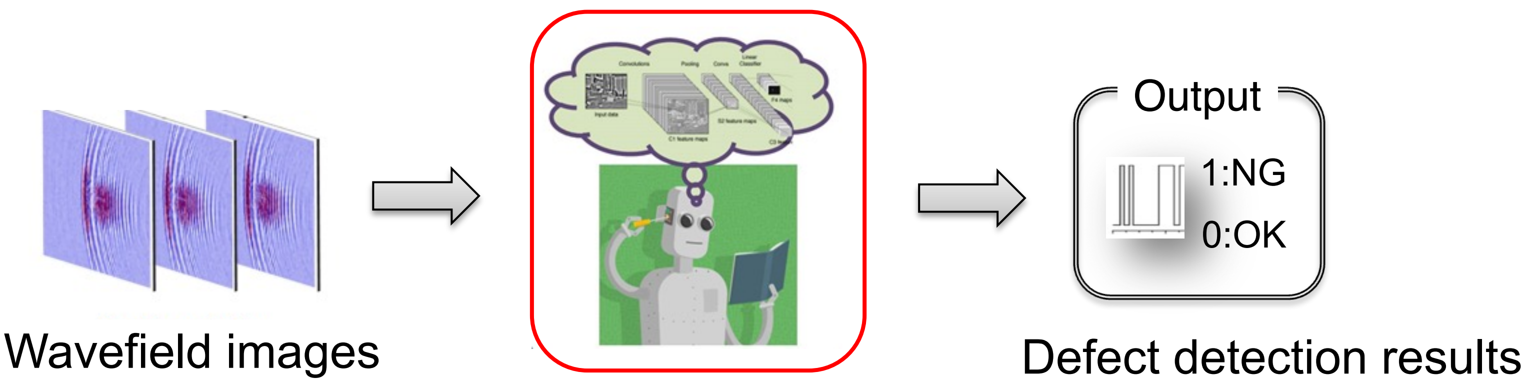
2. Application: Ultrasonic wavefields pattern recognition



Why employing AI/CV for Ultrasonic Wave Propagation Image pattern investigation:

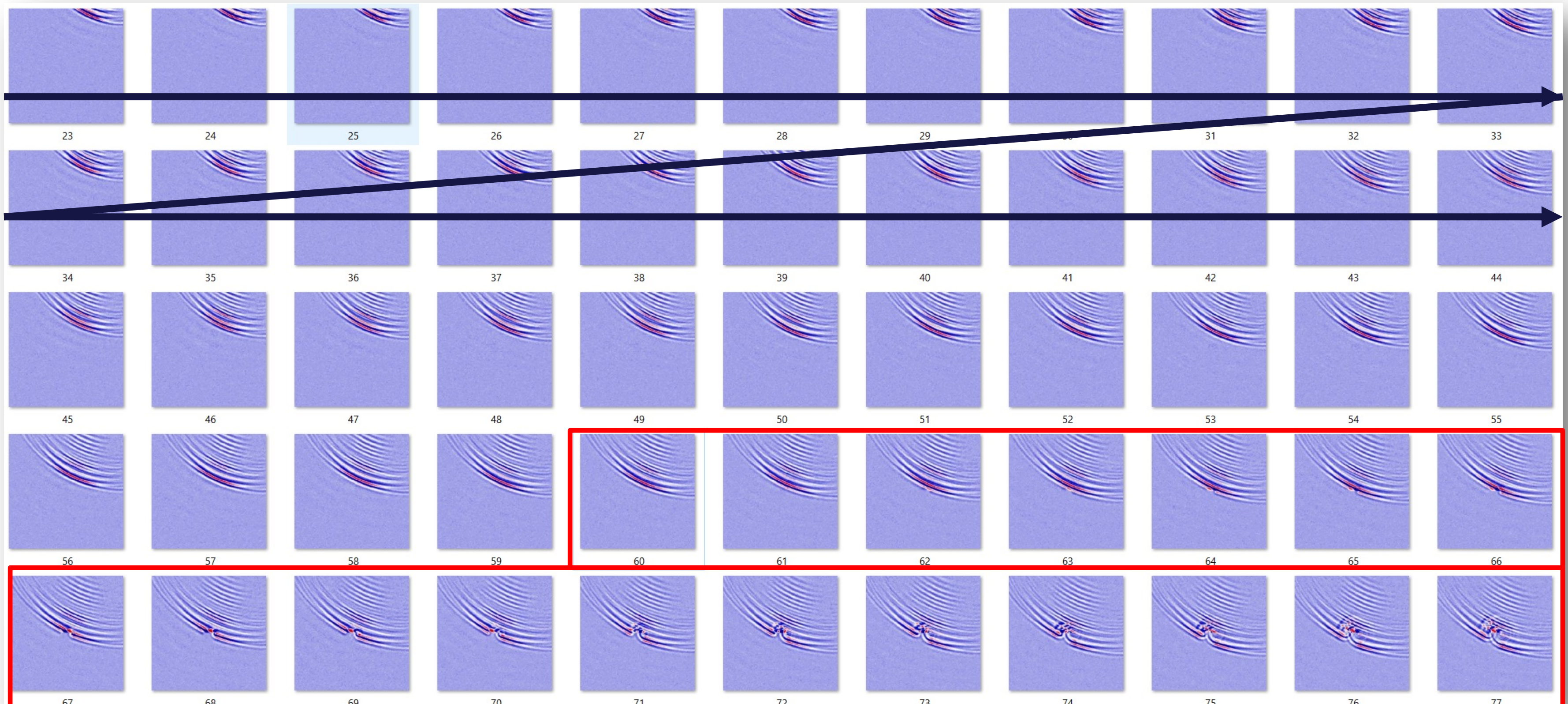
1. Reducing human efforts. In a production line scenario, almost all specimens are normal and the tedious work is quite tedious and exhausting.
2. Ambiguous standard/criterion that rely on inspector individual judgement/intuition.
3. Opening prospective to derive further defect information such as defect size/type.

To alleviate the above issues, we introduce computer vision approach for the task



2. Application: Ultrasonic wavefields pattern recognition

In our research institute, the *Non-destructive evaluation research group, AIST, Japan*, developed Ultrasonic Wave Propagation Imaging (UWPI) System for defect detection with extensive industrial applications in vehicle engineering, airplane safety testing, and welding assessment, etc.



Defect-induced scattered echoes can be observed

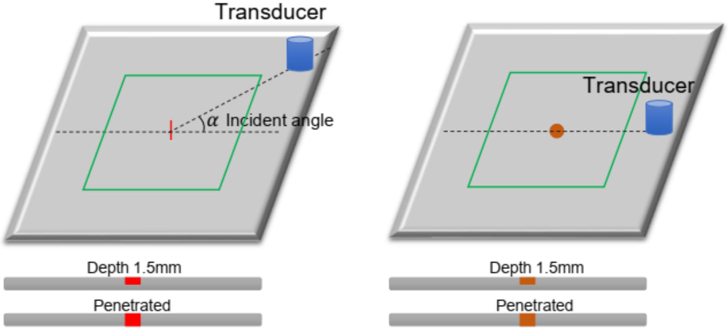
2. Application: Ultrasonic wavefields pattern recognition

Adopting computer vision technique for ultrasonic wavefield pattern analysis

Preparations: Dataset → Prune-then-finetune algorithms → Extensive evaluation

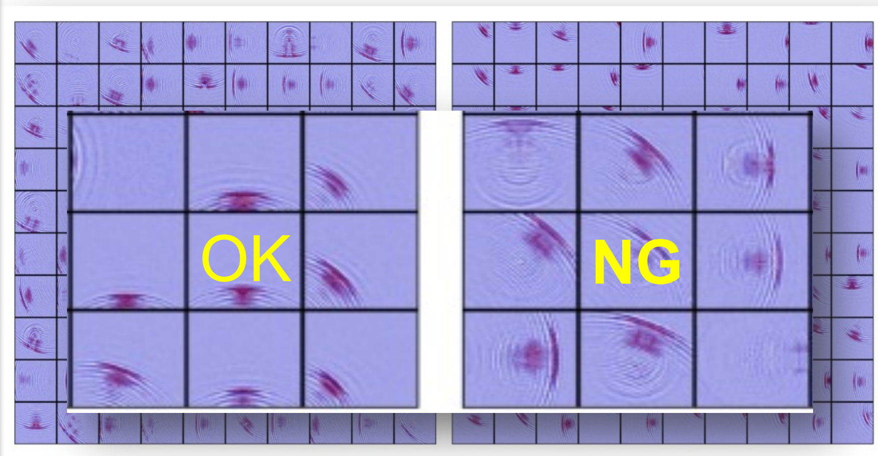
Aluminum plate: 17 (defective) + 1 (normal)

Wavefield dataset: 3235 (Normal) + 3069 (NG)



Slit defect

Hole defect



+

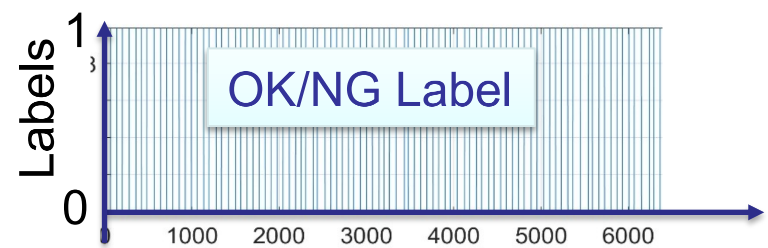
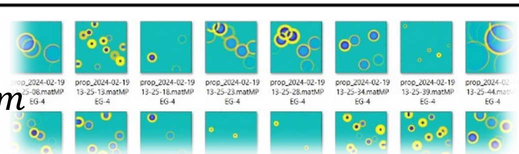


Table 1. Physics signal datasets of ultrasonic wavefield images.

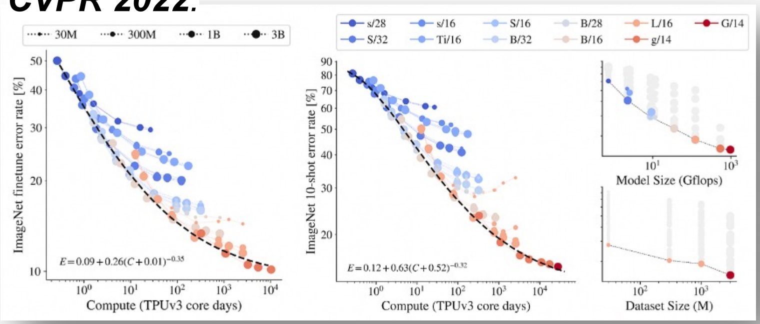
Attribute	$\mathcal{D}_{sim}$	$\mathcal{D}_{real}$
Data source	k-Wave [35]	USimgAIST [13]
Data type	*Simulated*	*Real*
Image size	224 × 224 × 3	224 × 224 × 3
# of samples	16,000	7,004 (17 specimens)
Class label	✗ (unlabeled)	✔ (defect/non-defect)

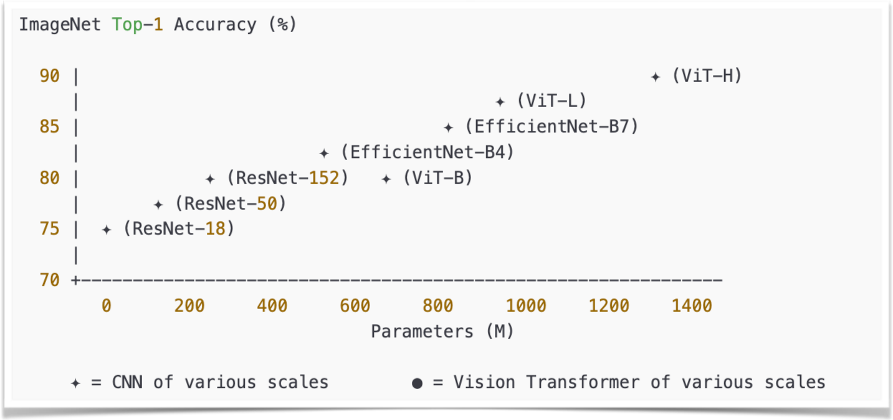
D_{sim}



3. Method: prune-then-finetune of ViTs for physics data

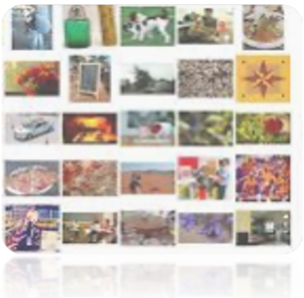
Zhai X, et. al. *Scaling vision transformers*. **CVPR 2022**.



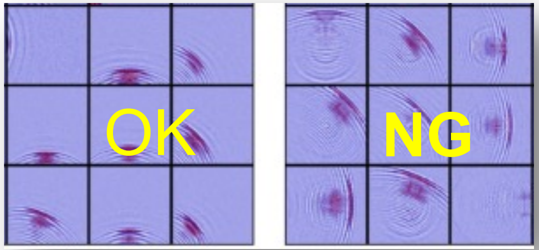


The major challenges brought by the pretrain-finetune paradigm:

1. Domain gap: the target downstream dataset can be different from the pretrain cases.



Significant Domain gap

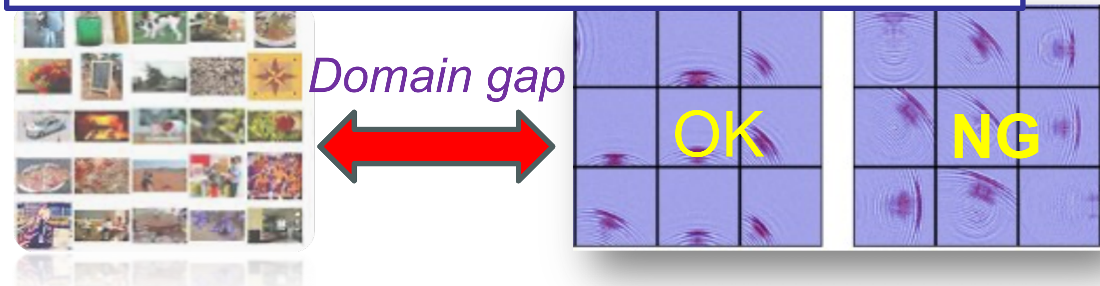


2. High computation cost for both finetuning and deployment stages.

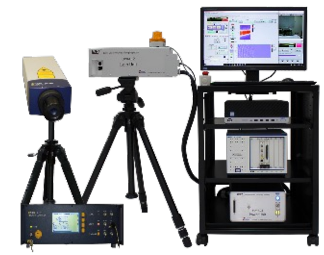
Onsite inspection demands for efficient models at inference

3. Method: prune-then-finetune of ViTs for physics data

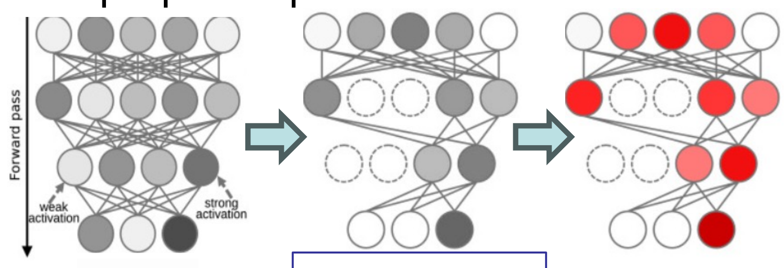
Domain gap: natural images vs. physics data



Onsite analysis requirement



Our concept: The proposed prune-then-finetune scheme



Physics-ticket for wavefields?

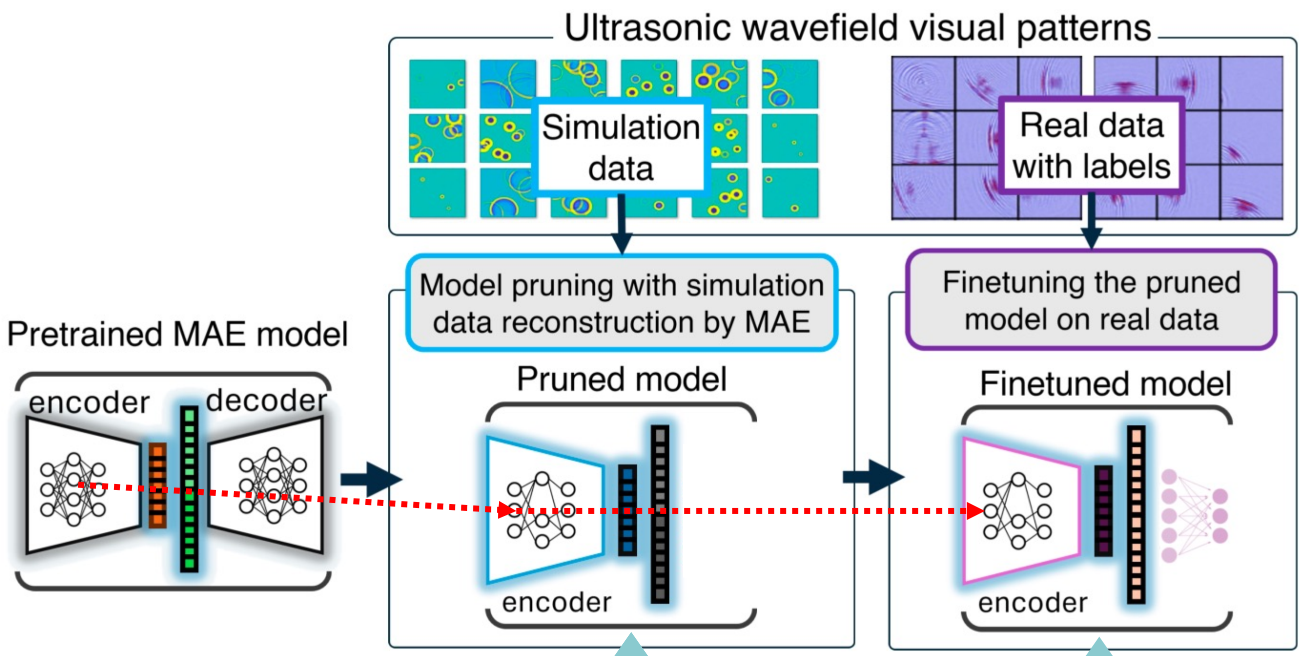
Our objective:

⇒ Tackling real-world problem of pattern classification of physics data (ultrasonic wavefields for inspection) with significant pretrain-finetune domain gap.

⇒ Propose novel prune-then-finetune scheme to facilitate efficient finetuning/deployment.

⇒ To mitigate data scarcity of physics measurement data, we incorporate simulation data to enhance generalization.

3-1. Flowchart of the proposed method



Phase 1 (Pruning): Performing auxiliary task of MAE reconstruction with simulation data of wavefields to highlight functional subnetwork for wavefield representation for model pruning.

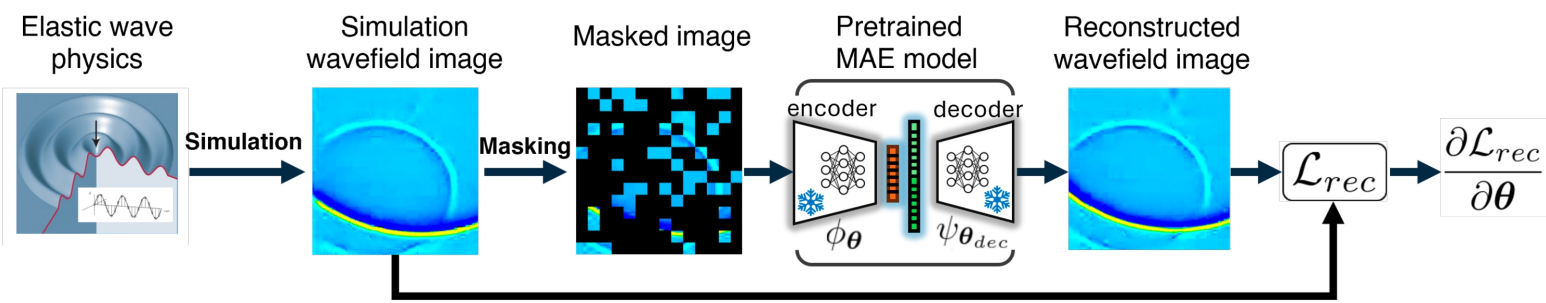
Phase 2 (finetuning): Conducting finetuning with the pruned model using real wavefield imaging data, saving computation cost for both finetuning and deployment stages.

3. Method: prune-then-finetune of ViTs for physics data

3-2. Detail process of model pruning

* The adaptation of Physics-Ticket Hypothesis: asserting that within the overparameterized model, a sparse subnetwork—Physics-Ticket—encodes intrinsic representations of physics signal (ultrasonic wavefield) patterns.

* To identify the functional substructures, we apply gradient-based structured pruning based on masked-autoencoder (MAE) reconstruction using simulation data.



Masked auto encoder (MAE) on ***simulation data*** to discover physics-ticket. Note that the pretrained model is fixed (frozen).

The loss of MAE reconstruction: $\mathcal{L}_{rec}(\theta^*, \theta^*_{dec}; \mathcal{D}_{sim}) =$

$$\mathbb{E}X \sim \mathcal{D}_{sim}\mathbb{E}M \sim \texttt{uniform}(p) \sum^{n} M_j \| x_j - [\psi_{\theta^*_{dec}} \circ \phi_{\theta^*}(\{x_k\}_{k|M_k=0})]_j \|_2^2,$$

Specifically, the loss gradient $\partial \mathcal{L}_{rec} / \partial_{\theta}$ more directly reflects the degree of contribution (significance) to describing the physics law of wavefields, thus, can be utilized as a key clue for pruning.

Pruning criterions based on the loss gradients of MAE task

$$C_1(\theta_j^*) = \left\| \frac{\partial \mathcal{L}_{rec}}{\partial \theta_j^*} \theta_j^* \right\|_2^2 \qquad C_2(\theta_j^*) = \frac{\partial^2 \mathcal{L}_{rec}}{\partial {\theta_j^*}^2} {\theta_j^*}^2$$

Pruning criterions based on magnitude of weights

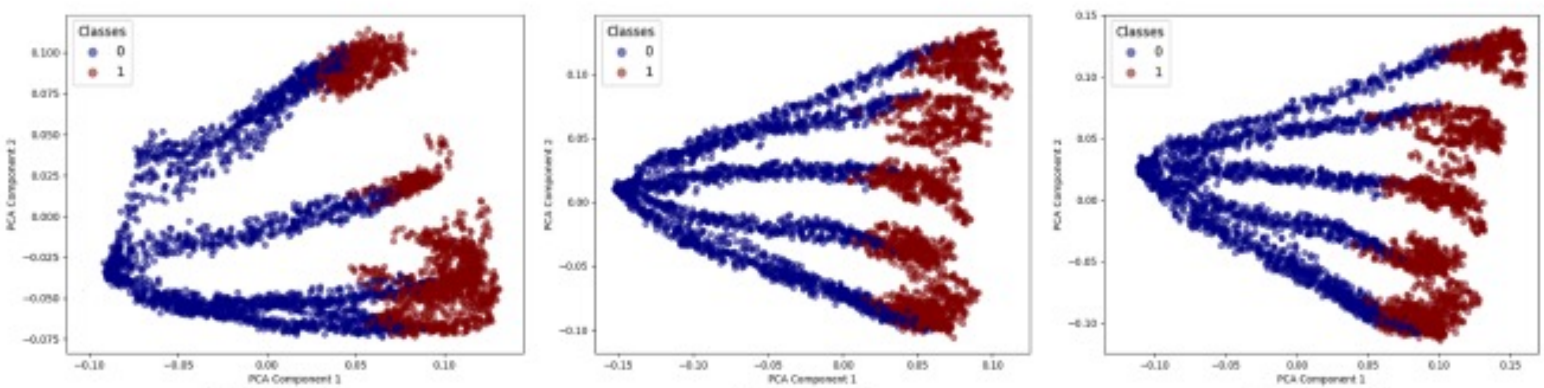
$$C_0^{l_1}(\theta_j^*) = \|\theta_j^*\|_2^2, \quad C_0^{l_2}(\theta_j^*) = \|\theta_j^*\|_1.$$

3. Method: prune-then-finetune of ViTs for physics data

3-3. Experiment 1: Does MAE reconstruction on simulation data yields efficient representations for real wavefield data?

Since we propose to *use simulation data (with MAE task) as stimuli for model pruning*. The first concept ought to be proved is *that the representation learnt from simulation data is indeed useful for real physics measurements*.

To this end, we visualized the latent feature distributions of real wavefield images during MAE training via principal component analysis. From left to right, the figures display the initial latent distributions and the trained ones after 20 and 40 epochs via the MAE loss (1) ***solely on simulated data*** (D_{sim}). Defective (label 1) and normal (label 0) classes are indicated by red and blue points, respectively.



(a) 0 epoch (b) 20 epochs (c) 40 epochs

We therefore can evidently confirm that the feature learnt with simulation data (via MAE model updates) efficiently enhanced the discrimination for real wavefield pattern classification.
It validates that ***MAE task using D_{sim} produces efficient clues for model pruning***.

4. Evaluation results and conclusions

Pruning criterions based on the loss gradients of MAE task

$$C_1(\theta_j^*) = \left\| \frac{\partial \mathcal{L}_{rec}}{\partial \theta_j^*} \theta_j^* \right\|_2^2 \qquad C_2(\theta_j^*) = \frac{\partial^2 \mathcal{L}_{rec}}{\partial {\theta_j^*}^2} {\theta_j^*}^2$$

Pruning criterions based on magnitude of weights

$$C_0^{l_1}(\theta_j^*) = \|\theta_j^*\|_2^2, \quad C_0^{l_2}(\theta_j^*) = \|\theta_j^*\|_1. \qquad (4)$$

Experiment 2: Comparisons on the importance measures for model pruning

Table 3. Comparison of pruning methods under Leave-One-Specimen-Out setting.

#	Method	Accuracy	Precision	Recall	F1-score
Magnitude-based pruning					
1	l_1 (4)	$97.00 \pm 3.30\%$	$96.90 \pm 3.20\%$	$96.20 \pm 3.10\%$	$96.40 \pm 2.30\%$
2	l_2 (4)	$97.40 \pm 1.70\%$	$97.50 \pm 2.40\%$	$96.20 \pm 3.50\%$	$96.70 \pm 1.60\%$
Proposed Physics-guided pruning					
3	Taylor (2)	$97.90 \pm 1.80\%$	$97.80 \pm 2.60\%$	$97.40 \pm 2.80\%$	$97.20 \pm 1.60\%$
4	Hessian (3)	$98.00 \pm 1.70\%$	$97.60 \pm 2.20\%$	$97.50 \pm 2.40\%$	$97.40 \pm 1.40\%$

Experiment 3: investigation on pruning ratio / Pruning ratio vs. performance trade-off.

Table 4. Model efficacy comparison across pruning ratios. The performances are measured in the leave-one-specimen-out protocol.

Pruning ratio	**0.5**	**0.7**	**0.9**	**0.95**
MACs (GFLOPs)	9.86	6.64	3.58	2.78
Param. No. (M)	42.93	25.07	8.10	3.64
Model size (MB)	167.81	98.04	31.75	14.31
Accuracy (%)	98.10 ± 1.70	97.90 ± 1.60	97.90 ± 1.80	96.90 ± 2.00
Precision (%)	97.80 ± 2.40	97.40 ± 2.50	97.80 ± 2.60	96.60 ± 2.60
Recall (%)	97.70 ± 2.20	97.40 ± 2.40	97.40 ± 2.80	95.40 ± 4.80
F1-score (%)	97.70 ± 1.30	97.30 ± 1.30	97.20 ± 1.60	96.00 ± 2.90

4. Evaluation results and conclusions

Experiment 4: Comparison with the practical models

Table 5. Computational metrics comparison with standard architectures.

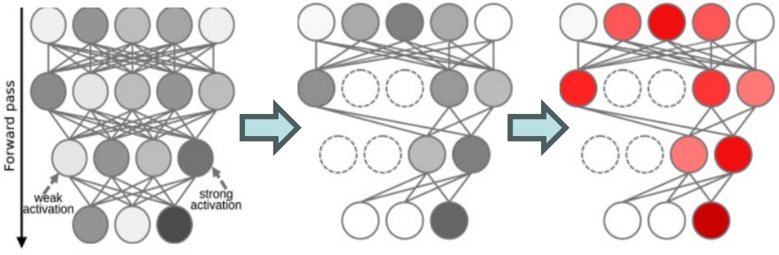
Architecture	ResNet-18	ResNet-50	ViT-B	Proposed
Complexity				
MACs (GFLOPs)	1.82	4.12	17.58	3.58
FLOP efficiency*	1.0×	2.3×	9.7×	**2.0**×
Model Size				
Parameters # (M)	11.69	25.56	86.57	**8.10**
File Size (MB, FP32)	45	98	330	**31.75**
Param. efficiency*	1.0×	2.2×	7.4×	**0.7**×
Performance (F1)				
USimgAIST*	92.20% [13]	95.08% [13]	97.84% [34]	97.24%

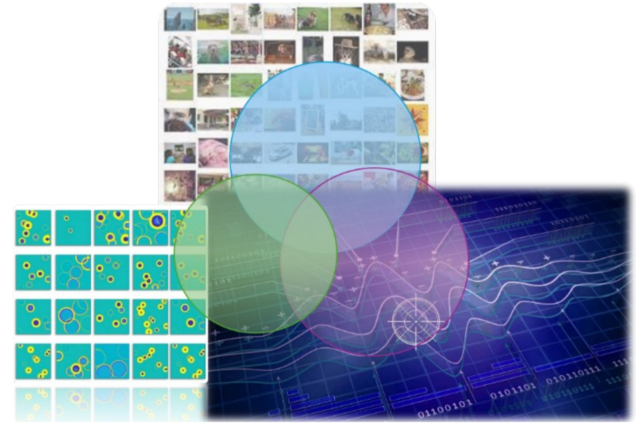
* Baseline results on USimgAIST dataset for defect classification [13], [34]
* Efficiency is measured using a normalized ratio to ResNet-18 (ResNet-18 → 1.0)

Results summary:

1. Achieving 9.7× FLOPs reduction compared to ViT-B, rendering ResNet-18-level efficiency.
2. Delivering significantly higher classification accuracy (**97.24% vs 92.20%**).
3. 10× Smaller Model Footprint, reducing model size from **330 MB to 31 MB**
4. Practical Deployment on Resource-Constrained Platforms

The proposed prune-then-finetune scheme





Key Contributions

Physics-Ticket Discovery: Identified a compact Physics-Ticket subnetwork within a vision foundation model that effectively captures ultrasonic wavefield physics for data-scarce scenarios.
Physics-Guided Pruning: Leveraged self-supervised MAE learning on simulation data to prune physics-irrelevant pathways, yielding a compact and generalizable model.
Real-World Validation: Achieved **90% model pruning** while retaining **97.2% Macro-F1** on real ultrasonic inspection data, demonstrating strong potential for **edge-deployable ultrasonic inspection**.

Thank you for your attention!
Any questions?