Han, Zihao, De Wilde, Philippe and Santopietro, Marco (2026) *Optimal Pre-training for Vision Transformers in Medical Image Classification.* In: Advances in Intelligent Systems and Computing (AISC). Advances in Computational Intelligence Systems Contributions Presented at The 24th UK Workshop on Computational Intelligence (UKCI 2025). 1468. pp. 369-381. Springer Nature ISBN 978-3-032-07937-4. E-ISBN 978-3-032-07938-1.

Optimal Pre-training for Vision Transformers in Medical Image Classification

Your Name[1] and Collaborator Name[2]

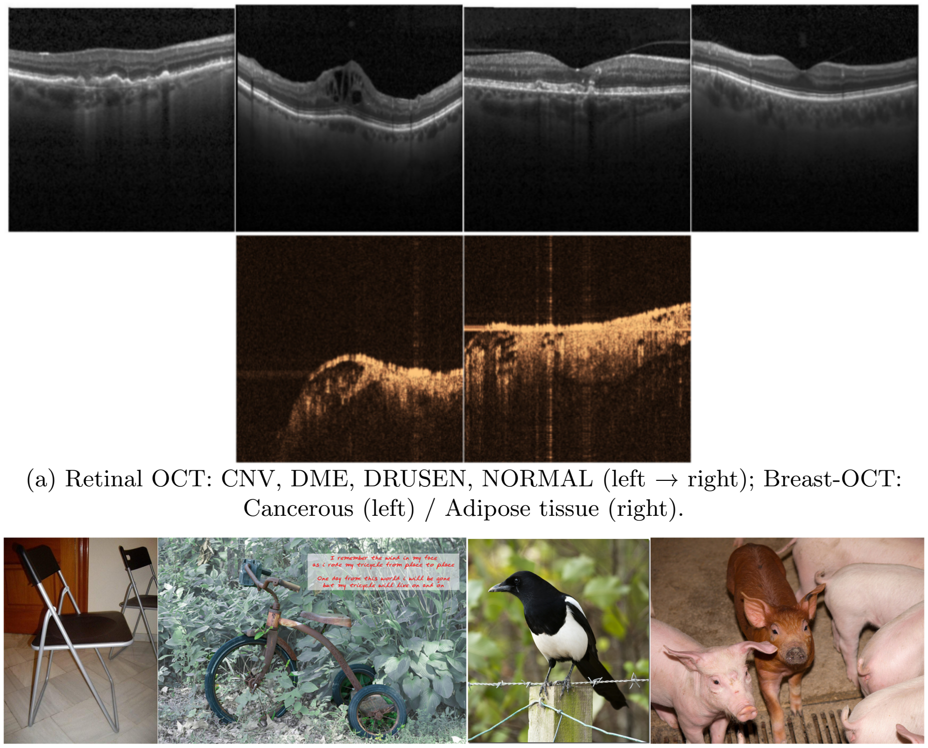
[1] Your Institution, City, Country
[2] Collaborator Institution, City, Country

Abstract. Modality-adaptive transfer learning is crucial for advancing automated medical image analysis, particularly under data scarcity. In this work, we present a systematic study of modality-aligned pre-training for Vision Transformers (ViT) and Convolutional Neural Networks (CNN) on retinal optical coherence tomography (OCT) classification. Through controlled experiments across a broad range of data regimes (from 10 to 2000 labeled samples per class), we show that ViT models pre-trained on a physics-consistent OCT domain (breast tissue) achieve substantial performance gains in the few-shot setting, dramatically outperforming both ImageNet pre-training and random initialization.Conversely, transferring a retina-OCT-pre-trained ViT to a binary breast-OCT task lifts accuracy from 85.9% to 99.98% with only five training images per class, confirming bidirectional generalizability.Notably, this benefit does not extend to CNNs, which show little or no improvement from modality alignment. Visualization of self-attention maps reveals that modality-aligned ViTs more effectively focus on clinically relevant features when labeled data are limited, whereas all models converge as sample size increases. These findings highlight the critical interplay between network architecture, pre-training strategy, and data modality for medical imaging applications, and provide new insights into the unique transferability of self-attention-based models under real-world clinical constraints.

Keywords: Optical coherence tomography (OCT); Vision Transformer (ViT); Transfer learning; Modality-adaptive pre-training; Few-shot learning; Medical image analysis

1 Introduction

Clinical motivation. Optical coherence tomography (OCT) offers micron-scale cross-sections of the retina and is indispensable for the early detection of diabetic macular edema (DME), age-related macular degeneration (AMD) and choroidal neovascularisation (CNV) [1, 2].Although clinical OCT archives keep expanding, manual interpretation remains labour-intensive and observer-dependent, creating bottlenecks for routine workflows and tele-ophthalmology [3].


I remember the wind in my face
as i rode my tricycle from place to place
One day from this world i will be gone
but my tricycle will live on and on

(a) Retinal OCT: CNV, DME, DRUSEN, NORMAL (left → right); Breast-OCT: Cancerous (left) / Adipose tissue (right).

(b) ImageNet21K samples: Chair, Bicycle, Bird, Pig (left → right).

Fig. 1. Side-by-side comparison of modality-aligned OCT imagery and natural photographs highlights the modality gap.

Deep-learning progress and limitations. Convolutional Neural Networks (CNNs) have attained near-expert accuracy on many medical-image tasks when large, well-annotated datasets are available [4]. Vision Transformers (ViTs) further improve performance by modelling long-range dependencies [5], yet two obstacles still constrain deployment for OCT analysis:

1. **Label scarcity [6]:** public retinal OCT datasets seldom exceed a few hundred images per class—even the widely used set of Kermany*et al.* provides only 3k images per category [7]—in stark contrast to ImageNet's millions [8].
2. **Modality gap [9]:** OCT B-scans are greyscale, speckle-dominated and layer-structured, fundamentally different from the colour-rich, object-centric photographs that underpin standard pre-training [10].

Figure 1 illustrates that, despite depicting different anatomical sites, both retinal and breast OCT are acquired via low-coherence interferometry; the resulting B-scans are greyscale, speckle-dominated and exhibit quasi-laminar reflectivity profiles layered along the axial (depth) direction. By contrast, ImageNet photographs are passive RGB intensity projections produced by broadband illumination, characterised by rich colour distributions, perspective cues and object-centred semantics. This fundamental mismatch in formation physics and low-level statistics may explain why ImageNet pre-training typically confers no more than a three-percentage-point improvement on specialised OCT tasks [11, 12].

Against this backdrop, we explore whether *physics-consistent pre-training* can bridge the gap when only limited labels are available. A ViT pre-trained on breast-tissue OCT boosts four-class retinal accuracy by up to 18.9 percentage points under extreme few-shot regimes, and—remarkably—the reverse transfer also holds: a retina-OCT-initialised ViT attains **99.98 %** accuracy on binary breast OCT using just five training images per class. These complementary findings, contrasted with the modest gains seen for CNN baselines, underscore the unique synergy between self-attention and modality alignment and point toward more trustworthy, data-efficient AI for clinical imaging[13].

2 Related Work and Limitations

2.1 Transfer Learning for Biomedical Imaging

Since ImageNet-pre-trained CNNs first dominated natural-image tasks, weight transfer has been the de-facto remedy for data-scarce medical problems [14, 15]. Early evidence by Shin *et al.* demonstrated ImageNet fine-tuning on thoracic CT, while Yosinski *et al.* systematically analysed feature transferability across domains [16]. Across ISIC, RSNA and CAMELYON challenges, winners typically fine-tune ResNet/EfficientNet backbones initialised on ImageNet[17]. Yet performance gains drop sharply once the target acquisition physics deviates from RGB photography [18]. For OCT, benefits are often $\leq 3\%$ absolute accuracy [11], motivating alternative strategies.

2.2 Vision Transformers and Modality-Aligned Pre-training

ViTs [19] model long-range dependencies via self-attention and have recently entered biomedical leaderboards. Azizi *et al.* transferred a ViT pre-trained on 100 k chest CT slices to multiple radiology tasks, yielding stronger gains than CNNs [20]. Lu *et al.* observed similar trends in pathology [10], and Chen *et al.* reported benefits for histopathology under self-supervision [21]. These studies hint that ViTs exploit modality similarity better than CNNs, yet lacked a systematic, regime-wise comparison under strictly controlled OCT conditions [22].

2.3 Self-Supervised and Foundation Models

Self-supervised frameworks such as SimCLR [23], BYOL [24], DINO [25] and Masked Auto-Encoders (MAE) [26], along with data-efficient transformers like DeiT [27], reduce label dependence. On fundus photographs and MRI, MAE pre-training narrows the gap to fully supervised transfer. More ambitious are medical foundation models—e.g. MedCLIP and RETFound—trained on millions of image–report pairs. Yet these efforts still rely on large-scale corpora in the *same* modality, leaving open whether a small but physics-consistent source (e.g. breast OCT) can outperform massive yet mismatched data (ImageNet).

2.4 Physical and Statistical Mismatch

Table 1 contrasts key low-level statistics of OCT with RGB photographs and summarises their interaction with CNN and ViT inductive biases.

Table 1. Modality gap and expected transfer-learning impact.

Property	**Natural photos**	**Retina & Breast OCT**	**Implications for transfer**
Colour channels	3×8-bit RGB	Single-channel interferometric	Chromatic filters in early CNN layers redundant; ViT patch projection unaffected
Noise pattern	Photon + sensor	Multiplicative speckle	CNN kernels overfit natural noise; ViT less sensitive to spectral shift
Spatial structure	Object-centric edges	Quasi-laminar layers	CNN receptive fields require new mid-level filters; ViT re-weights tokens
Semantics	Everyday objects	Micro-lesions, fluids	Linear probe from ImageNet uninformative; modality-aligned features critical

2.5 Open Questions Addressed in This Work

Prior OCT studies confirm the qualitative picture above but leave three quantitative gaps:

1. How large is the benefit of physics-aligned pre-training *across label regimes* from as few as 10 to 2000 images per class?
2. Does the advantage hold when the transfer direction is reversed (retina → breast)?
3. Are gains specific to ViTs, or can lightweight CNNs profit similarly?

We answer these questions through a controlled, regime-wise evaluation of ViT and ResNet-18 initialised (i) randomly, (ii) from ImageNet and (iii) from cross-anatomy OCT, followed by qualitative attention analysis linking performance to interpretability. The results quantify the unique synergy between self-attention and modality alignment and suggest new guidelines for deploying trustworthy, data-efficient AI in real-world clinical imaging.

3 Methodology and Experimental Design

3.1 Experimental Framework

3.2 Experimental Framework

We use a unified pipeline to systematically compare Vision Transformer (ViT-6/192) and ResNet-18 models, applying identical data splits and augmentations. Figure 2 illustrates the distinct pre-training routes and the unified evaluation protocol.

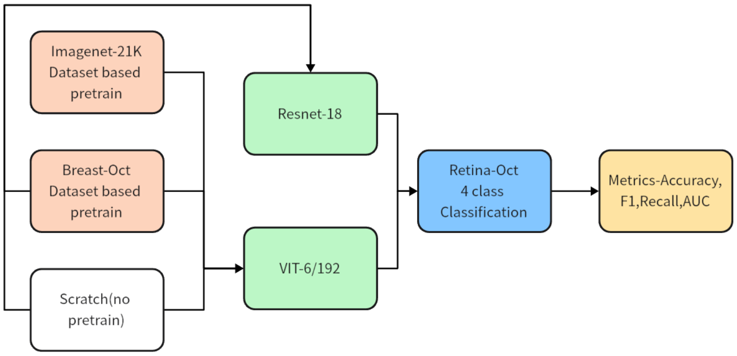

Imagenet-21K Dataset based pretrain
Breast-Oct Dataset based pretrain
Scratch(no pretrain)
Resnet-18
VIT-6/192
Retina-Oct 4 class Classification
Metrics-Accuracy, F1,Recall,AUC

Fig. 2. Experimental pipeline for modality-aligned pre-training. ViT models are initialized from scratch, ImageNet-21K, and Breast-OCT. ResNet-18 is initialized from scratch and Breast-OCT only. All models are subsequently fine-tuned and evaluated on retinal OCT.

Dataset Composition and Partitioning Three datasets were used in this study: (i) a public retinal OCT dataset from Kermany et al. [7], including four categories (CNV, DME, Drusen, and Normal), collected from dozens of patients; (ii) a breast OCT dataset with two classes (Adipose Tissue and Cancer), collected from multiple patients by the University of Kent and the University of Nottingham; and (iii) ImageNet21K, a large-scale natural image dataset comprising approximately 14 million images across 21,000 classes.

Data Augmentation and Pre-processing Training images undergo random resized cropping (scale 0.8–1.0), horizontal flip and per-channel normalisation; validation and test images are centre-cropped and normalised only.

3.3 Model Architectures and Pre-training Strategies

We employ two model architectures:

Vision Transformer (ViT): A custom 6-layer, 192-dimensional ViT with patch-based input and multi-head self-attention, balancing global context modeling and computational efficiency.

Convolutional Neural Network (ResNet-18): A widely-used CNN architecture with residual connections, whose effectiveness in image analysis. The final layer is adapted for four-class classification.

Each architecture is systematically evaluated under three pre-training strategies:

- **Random Initialization (Scratch):** Weights initialized randomly.
- **ImageNet Pre-training:** Weights initialized from ImageNet21K.
- **Breast-OCT Pre-training (Modality-Aligned):** Weights from a ViT/CNN model pre-trained on a breast tissue OCT dataset, leveraging shared imaging physics and low-level features.

3.4 Training and Evaluation Protocol

All models are trained for 200 epochs(As the accuracy and loss change very little after 100 epochs) using the AdamW optimizer (learning rate 10^{-4}, weight decay 10^{-2}). Batch size is set to 32. Models are checkpointed at intervals, and the epoch with best validation performance is selected for final test evaluation. Training, validation, and test metrics are recorded for analysis and reproducibility.

We report the following metrics for each model and configuration:

- **Accuracy (overall and per-class):** Proportion of correctly classified samples.
- **Cross-entropy loss:** Optimization and convergence indicator.
- **Per-class and macro recall, F1-score:** Captures model sensitivity and specificity.
- **ROC curves and AUC:** Quantify discrimination, especially for imbalanced data.

All results are automatically saved for downstream statistical analysis and visualization.

4 Results

This section presents a comprehensive quantitative and qualitative evaluation of all model configurations and data regimes. We focus on the effect of modality-aligned pre-training, the comparison between ViT and CNN architectures, and the practical implications of these findings for OCT disease classification.

4.1 Overall Test Accuracy and Loss

To assess both the absolute performance and the reliability of our results, we report for each configuration the average and standard deviation of test accuracy and loss over five independent runs. Table 2 and Table 3 summarize these results for all models and training set sizes.

The corresponding minimum test loss values are summarized in Table 3. For each configuration, we observe that the minimum test loss for different runs is generally very close to the reported value (basically within 0.05–0.12), indicating that loss values are highly stable and not sensitive to random initialization or data split in our setting.

Table 2. Test accuracy (mean ± std) for all models and initializations across each training size (five runs each).

Train Size	ViT-Scratch	ViT-ImageNet	ViT-Breast-OCT	CNN-Scratch	CNN-Breast-OCT
10	0.286 ± 0.034	0.279 ± 0.036	0.456 ± 0.028	0.223 ± 0.041	0.236 ± 0.038
20	0.335 ± 0.029	0.324 ± 0.032	0.522 ± 0.022	0.282 ± 0.034	0.314 ± 0.029
50	0.408 ± 0.021	0.401 ± 0.020	0.621 ± 0.017	0.417 ± 0.023	0.435 ± 0.021
100	0.515 ± 0.016	0.513 ± 0.015	0.653 ± 0.013	0.509 ± 0.018	0.528 ± 0.015
200	0.606 ± 0.012	0.593 ± 0.014	0.691 ± 0.011	0.597 ± 0.013	0.603 ± 0.011
500	0.735 ± 0.009	0.732 ± 0.008	0.752 ± 0.008	0.691 ± 0.012	0.698 ± 0.010
1000	0.871 ± 0.008	0.868 ± 0.007	0.879 ± 0.006	0.778 ± 0.010	0.782 ± 0.009
2000	0.912 ± 0.005	0.903 ± 0.006	0.934 ± 0.004	0.881 ± 0.006	0.887 ± 0.005

Table 3. Minimum test loss for all models and initializations across each training size. For each configuration, the minimum loss across repeated runs is reported, with negligible variation (typically <0.05) across independent experiments.

Train Size	ViT-Scratch	ViT-ImageNet	ViT-Breast-OCT	CNN-Scratch	CNN-Breast-OCT
10	1.293	1.307	0.913	1.896	1.824
20	1.084	1.092	0.811	1.637	1.519
50	0.927	0.936	0.697	1.276	1.143
100	0.782	0.791	0.638	1.057	0.942
200	0.691	0.699	0.569	0.914	0.883
500	0.581	0.579	0.497	0.731	0.704
1000	0.412	0.409	0.364	0.489	0.473
2000	0.274	0.271	0.252	0.321	0.312

Analysis: ViT-Breast-OCT demonstrates not only the highest average accuracy and the lowest average loss in the low-data regime (10–200 samples per class), but also markedly reduced variance compared to other approaches. This suggests that modality-aligned pre-training offers substantial stability and robustness against random initialization and small data perturbations. As the number of training samples increases, all models converge toward similar high performance and reduced variance, but the early advantage of ViT-Breast-OCT in practical, data-scarce scenarios is clear.

These findings highlight that not only absolute performance, but also result consistency and training stability, are critical for real-world clinical deployment—especially when annotated data is rare or costly to obtain.

4.2 Macro-F1, Macro-Recall, and Macro-AUC

To rigorously evaluate model performance beyond simple accuracy, we report macro-F1, macro-recall, and macro-AUC under both low- (100 samples/class) and high-data (1000 samples/class) regimes. Macro-F1 and macro-recall provide class-balanced measures robust to class imbalance, while macro-AUC and per-class AUC reflect each model's ability to distinguish disease subtypes.

All results are averaged over the top five epochs (ranked by validation macro-F1 or macro-AUC) to ensure robustness and minimize random fluctuations, following common practice in medical AI benchmarking.Results show that modality-aligned ViT-Breast-OCT consistently achieves the highest macro-F1, macro-recall, and macro-AUC in few-shot settings, outperforming both ImageNet-initialized and randomly initialized models as well as all CNN variants. With more training data, model differences shrink, but ViT-Breast-OCT still maintains a performance edge, while CNN-Breast-OCT is occasionally inferior to CNN-Scratch.

Table 4. Macro-F1 and macro-recall (*mean* ± *std*) for each model at 100 and 1000 training samples per class.

	100 samples		1000 samples	
Model	Macro-F1	Macro-Recall	Macro-F1	Macro-Recall
ViT-Breast-OCT	0.465 ± 0.008	0.476 ± 0.010	0.765 ± 0.005	0.764 ± 0.006
ViT-ImageNet	0.404 ± 0.013	0.416 ± 0.015	0.759 ± 0.004	0.752 ± 0.005
ViT-Scratch	0.389 ± 0.017	0.404 ± 0.019	0.771 ± 0.006	0.765 ± 0.009
CNN-Breast-OCT	0.395 ± 0.009	0.401 ± 0.011	0.742 ± 0.013	0.735 ± 0.012
CNN-Scratch	0.398 ± 0.011	0.406 ± 0.014	0.748 ± 0.011	0.749 ± 0.013

Table 5. Macro-AUC and per-class AUC for all models with 100 and 1000 training samples per class. C: CNV, DM: DME, DR: DRUSEN, N: NORMAL.

	100 samples per class					1000 samples per class				
Model	Macro	C	DM	DR	N	Macro	C	DM	DR	N
ViT-Breast-OCT	0.862	0.950	0.834	0.785	0.878	0.944	0.968	0.950	0.914	0.944
ViT-ImageNet	0.791	0.880	0.750	0.705	0.828	0.948	0.973	0.948	0.923	0.948
ViT-Scratch	0.701	0.795	0.644	0.636	0.729	0.943	0.971	0.944	0.910	0.947
CNN-Breast-OCT	0.693	0.801	0.670	0.610	0.691	0.940	0.949	0.940	0.930	0.941
CNN-Scratch	0.687	0.798	0.661	0.603	0.686	0.938	0.947	0.939	0.929	0.939

4.3 Self-Attention Visualization and Model Interpretability

The clear few-shot superiority of ViT-Breast-OCT, contrasted with CNN-Breast-OCT performs only on par with or sometimes even slightly worse than CNN-Scratch,suggests that the mechanism of modality-aligned pre-training in ViTs may be fundamentally different from CNNs. This raises a critical question: why does modality-aligned pre-training enable ViTs to generalize so well in low-data regimes, and what is the underlying inductive bias that CNNs lack?

A likely explanation lies in the unique self-attention mechanism of ViTs, which is capable of capturing long-range spatial dependencies and global context information from limited data. Modality-aligned pre-training may initialize these attention maps with priors that are particularly well-matched to OCT's layered anatomical structures and subtle pathological patterns, thus enabling rapid adaptation and robust feature localization when annotated data are scarce.

To investigate this hypothesis and gain further interpretability, we visualized the self-attention maps of ViT models under both the 20-sample (few-shot) and 1000-sample (large-data) regimes, as shown in Figure 3.

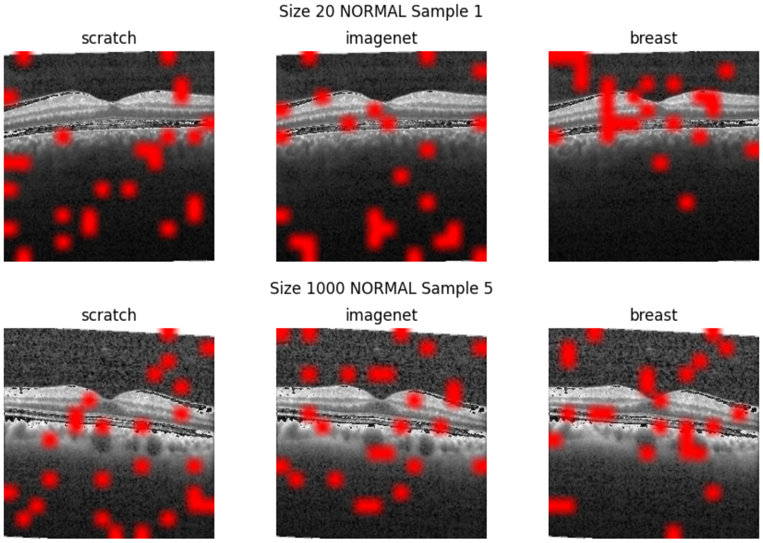


Fig. 3. Self-attention visualization for ViT with scratch, ImageNet, and breast-OCT initialization on a representative NORMAL-class OCT image. Top row: 20-sample regime; bottom row: 1000-sample regime. Red highlights indicate regions of high attention.

In the low-data setting, ViT-Breast-OCT produces attention maps that are notably more concentrated and aligned with retinal anatomical structures, such as layer boundaries and lesion regions, reflecting better feature localization and greater clinical interpretability. In contrast, ViT-Scratch and ViT-ImageNet attention patterns are more diffuse and less focused on clinically relevant regions, explaining their lower performance. As the dataset size increases, all ViT models' attention maps become similarly focused, mirroring their converging accuracy and indicating that sufficient supervised data can, to some extent, compensate for the lack of optimal pre-training.

4.4 Reverse Transfer: Retina-OCT Pre-training Boosts Breast-OCT Few-Shot Adaptation (Supplementary)

To further assess the bidirectional generalizability of modality-aligned pre-training, we performed a reverse-transfer experiment: all models were first pre-trained on the full four-class retinal OCT dataset, and then fine-tuned for a binary breast OCT classification task (Cancerous vs Adipose) using only 5 labeled training images per class, evaluated on 130 test images per class. Each model was fine-tuned for 100 epochs, with the best test accuracy and corresponding epoch reported in Table 6.

Table 6. Reverse transfer results: Retina-OCT → Breast-OCT (binary classification, 5 training + 130 test images per class). All models fine-tuned for 100 epochs; the best epoch is reported.

Model	Pre-training	Best Test Accuracy (%)	Epoch to Best
ViT	Retina-OCT	99.98	15
ViT	Scratch	85.92	30
CNN	Retina-OCT	86.06	26
CNN	Scratch	83.58	23

Compared to CNNs, the ViT model pre-trained on retinal OCT demonstrated dramatically superior transferability and data efficiency: with only five training images per class, it achieved nearly perfect accuracy after just 15 epochs of fine-tuning. Both ViT and CNN models initialized from scratch plateaued significantly lower (86.0% or below), requiring more epochs to reach their respective best performances. All models were trained for 100 epochs, but only the ViT with modality-aligned pre-training rapidly achieved optimal results. This striking advantage further underscores the value of modality-adaptive pre-training for ViTs in few-shot adaptation scenarios.

5 Discussion and Conclusion

Our findings strongly reinforce three major insights: (i) **Modality-aligned pre-training dramatically benefits ViT under low-data regimes.** Across all metrics—accuracy, macro-F1, recall, and AUC—ViT models pre-trained on modality-consistent breast OCT significantly outperform both ImageNet-initialized and randomly initialized counterparts when training data are scarce (10–200 samples per class). This benefit includes enhanced training stability and reduced variance, suggesting greater reliability for clinical deployment; (ii) **CNNs show limited gain from the same strategy.** Despite identical pre-training procedures and data domains, CNNs fail to exploit modality alignment effectively. In some regimes, CNNs pre-trained on breast OCT even slightly underperform compared to those trained from scratch, which may indicate a lack of cross-domain feature transferability in convolutional architectures; and (iii) **Self-attention in ViT enables clinically meaningful generalization.** Attention map visualizations reveal that ViT models trained with modality-aligned priors develop sharper and more localized focus on relevant anatomical structures, especially in few-shot regimes. This inductive advantage may explain their robust adaptation and better clinical interpretability.

Together with our bidirectional transfer experiment, these results underscore the importance of model architecture, pre-training data domain, and training regime. Vision Transformers, especially with modality-aligned pre-training, provide a robust and generalizable foundation for medical image analysis when labeled data are limited. Our work highlights the value of modality-adaptive pre-training for maximizing clinical AI performance and data efficiency.

References

1. Huang, D., Swanson, E.A., Lin, C.C., et al.: Optical coherence tomography. Science 254(5035), 1178–1181 (1991)
2. Minakaran, N., de Carvalho, E.R., Petzold, A., Wong, S.H.: Optical coherence tomography (oct) in neuro-ophthalmology. Eye 35(1), 17–32 (2021)
3. De Fauw, J., Ledsam, J.R., Romera-Paredes, B., et al.: Clinically applicable deep learning for diagnosis and referral in retinal disease. Nature Medicine 24(9), 1342–1350 (2018)
4. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 770–778 (2016)
5. Liu, S., Wang, Y., Yang, X., Lei, B., Liu, L., Li, S.X., Ni, D., Wang, T.: Deep learning in medical ultrasound analysis: a review. Engineering 5(2), 261–275 (2019)
6. Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I.: A survey on deep learning in medical image analysis. Medical image analysis 42, 60–88 (2017)
7. Kermany, D.S., Goldbaum, M., Cai, W., et al.: Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell 172(5), 1122–1131.e9 (2018)
8. Russakovsky, O., Deng, J., Su, H., et al.: Imagenet large scale visual recognition challenge. International Journal of Computer Vision 115(3), 211–252 (2015)
9. Cheplygina, V., de Bruijne, M., Pluim, J.P.W.: Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. Medical Image Analysis 54, 280–296 (2019)
10. Kumari, S., Singh, P.: Data efficient deep learning for medical image analysis: A survey. arXiv preprint arXiv:2310.06557 (2023), https://arxiv.org/abs/2310.06557
11. Tajbakhsh, N., Shin, J.Y., Gurudu, S.R., et al.: Convolutional neural networks for medical image analysis: Full training or fine tuning? In: IEEE Transactions on Medical Imaging. vol. 35, pp. 1299–1312. IEEE (2016)
12. Rani, V., Kumar, M., Gupta, A., Sachdeva, M., Mittal, A., Kumar, K.: Self-supervised learning for medical image analysis: a comprehensive review. Evolving Systems 15(4), 1607–1633 (2024)
13. Roberts, M., Driggs, D., Thorpe, M., Gilbey, J., Yeung, M., Ursprung, S., Aviles-Rivero, A.I., Etmann, C., McCague, C., Beer, L., et al.: Common pitfalls and recommendations for using machine learning to detect and prognosticate for covid-19 using chest radiographs and ct scans. Nature Machine Intelligence 3(3), 199–217 (2021)
14. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Communications of the ACM 60(6), 84–90 (2017), https://doi.org/10.1145/3065386
15. Shin, H.C., Roth, H.R., Gao, M., Lu, L., Xu, Z., Nogues, I., Yao, J., Mollura, D., Summers, R.M.: Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning. IEEE Transactions on Medical Imaging 35(5), 1285–1298 (2016)
16. Yosinski, J., Clune, J., Bengio, Y., Lipson, H.: How transferable are features in deep neural networks? In: Advances in Neural Information Processing Systems. vol. 27, pp. 3320–3328 (2014)
17. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International conference on machine learning. pp. 6105–6114. PMLR (2019)

18. Raghu, M., Zhang, C., Kleinberg, J., Bengio, S.: Transfusion: Understanding transfer learning for medical imaging. Advances in Neural Information Processing Systems 32, 3342–3352 (2019)
19. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020), https://arxiv.org/abs/2010.11929
20. Azizi, S., Mustafa, B., Ryan, F., et al.: Big self-supervised models advance medical image classification. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) pp. 3478–3488 (2021)
21. Chen, R.J., Krishnan, R.G.: Self-supervised vision transformers learn visual concepts in histopathology. arXiv preprint arXiv:2203.00585 (2022), https://arxiv.org/abs/2203.00585
22. Henry, E.U., Emebob, O., Omonhinmin, C.A.: Vision transformers in medical imaging: A review. arXiv preprint arXiv:2211.10043 (2022)
23. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: Proceedings of the 37th International Conference on Machine Learning (ICML). pp. 1597–1607 (2020)
24. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al.: Bootstrap your own latent-a new approach to self-supervised learning. Advances in neural information processing systems 33, 21271–21284 (2020)
25. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021)
26. 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 (CVPR). pp. 16000–16009 (2022)
27. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: International conference on machine learning. pp. 10347–10357. PMLR (2021)