# Domain-Aligned OCT Pre-training: Enhancing Retinal Disease Diagnosis Through Cross-Anatomy Vision Transformers

Anonymous Authors

Anonymous Institution

**Abstract.** Medical imaging often suffers from limited labeled data and substantial domain gaps when transferring models pre-trained on general-purpose benchmarks such as ImageNet. This study systematically compares three training strategies for Vision Transformers (ViTs) on a four-class retinal Optical Coherence Tomography (OCT) dataset(CNV, DME, Drusen, Normal): (1) training from scratch, (2) conventional ImageNet-based pre-training, and (3) a novel domain-specific pre-training method using OCT breast cancer images (adipose tissue vs. cancer). Experimental results clearly show that the domain-specific OCT breast pre-training significantly improves classification accuracy compared to both ImageNet pre-training and training from scratch, particularly under limited-data scenarios. These findings challenge the prevailing view that general-domain pre-training has limited utility in medical imaging, instead emphasizing the essential role of domain alignment in pre-training datasets. Our results highlight the critical advantage of domain-specific pre-training in medical imaging AI, demonstrating improved accuracy and potential for earlier retinal disease detection even with scarce labeled data. Future research should focus on constructing larger OCT-specific pre-training datasets and exploring advanced self-supervised methods tailored explicitly for medical imaging tasks.

**Keywords:** Optical Coherence Tomography · Vision Transformer · OCT Image Classification · Transfer Learning · Medical Imaging

## 1 Introduction

The effectiveness of deep learning in medical imaging is frequently hindered by the scarcity of annotated datasets, which limits the training and generalization capabilities of complex models such as Transformers [1]. Originally introduced for natural language processing (NLP), Transformers utilize self-attention mechanisms to effectively model long-range dependencies, significantly outperforming previous models in capturing contextual relationships [1]. The adaptation of this architecture into computer vision resulted in the Vision Transformer (ViT), which partitions images into patches and processes them as sequential data [2]. ViTs have demonstrated superior performance relative to traditional convolutional neural networks (CNNs) across diverse image classification tasks,

attributed to their ability to capture global visual context. However, ViTs are inherently data-hungry and typically require large-scale datasets for optimal performance, which poses a critical challenge in medical imaging domains characterized by limited labeled data [3].

Optical Coherence Tomography (OCT) is a vital medical imaging modality providing high-resolution, cross-sectional views of the retina, facilitating early detection and monitoring of retinal diseases such as *diabetic macular edema (DME)*, *age-related macular degeneration (DRUSEN)*, and *choroidal neovascularization (CNV)* [4].Compared to natural images, OCT scans exhibit three distinctive characteristics: (1)layer-specific contrast patterns (e.g., hyperreflective retinal layers) [5], (2)coherent speckle noise distribution [6], and (3)anisotropic spatial dependencies along depth and lateral axes [7]. These domain-specific attributes complicate the direct application of models originally developed for natural-image benchmarks.

Previous studies generally concluded that pre-training ViTs on natural-image datasets such as ImageNet offers minimal accuracy improvements for medical imaging applications, typically less than 3% [8]. However, these conclusions overlook the potential gains achievable when using closely related domain-specific pre-training datasets. We hypothesize that pre-training on domain-aligned OCT datasets—even those from different anatomical contexts—can significantly enhance diagnostic performance by addressing fundamental domain discrepancies including imaging physics (coherent OCT interferometry vs. natural image photon detection), structural representations (layered biological tissues vs. object-centric scenes), and pathological scales (micron-level tissue distortions vs. macroscopic object features).

Motivated by this hypothesis, we propose a novel two-stage cross-domain pre-training strategy:

1. Pre-training ViTs on a large OCT dataset originally collected for breast cancer detection (adipose tissue vs. cancer), leveraging the imaging similarity between different OCT modalities.
2. Fine-tuning this pre-trained model on a four-class retinal OCT dataset (CNV, DME, Drusen, Normal).

We systematically evaluate and compare three distinct training paradigms:

- **ViT-ImageNet21K:** General-purpose pre-training using ImageNet21K.
- **ViT-OCT-Breast:** Proposed OCT-specific pre-training using breast OCT images.
- **ViT-Scratch:** Direct training from random initialization on retinal OCT images.

Our experimental findings demonstrate a substantial performance gain when employing domain-specific OCT pre-training, particularly under limited-data conditions, improving classification accuracy from around 60% (general pre-training) to approximately 80%. Additionally, our results highlight the limited benefits derived from traditional ImageNet pre-training in OCT imaging tasks

due to substantial domain discrepancies. These insights emphasize the necessity and potential of domain-aligned pre-training datasets, calling for more extensive domain-specific OCT datasets and advanced self-supervised approaches specifically designed for medical imaging scenarios.

## 2 Related Work

Recent years have witnessed substantial progress in medical imaging analysis, driven by the convergence of deep learning algorithms, large-scale computing resources, and continuously expanding datasets. Table 1 provides a concise overview of representative methods in this domain, illustrating the historical dominance of CNNs, emerging Transformer-based approaches, and ongoing debates surrounding transfer learning strategies.

Table 1: Representative Related Works in Medical Imaging and Transfer Learning

| **Approach** | **Dataset** | **Key Findings** | **Refs** |
|---|---|---|---|
| CNN for OCT lesions | OCT datasets | High accuracy but data-hungry; limited long-range context modeling | [9, 10] |
| Data augmentation & GAN | Liver lesion | Synthetic data addresses class imbalance and small samples | [11] |
| Vision Transformer | ImageNet, COCO | Global attention outperforms CNNs, requires large datasets | [2] |
| Medical ViT variants | Brain MRI, CT | Hybrid architectures improve medical data efficiency | [12, 13] |
| OCT domain challenges | Retinal OCT/OCTA | Domain-specific structures limit standard models | [14] |
| ImageNet pre-training | Multi-modal | Speeds convergence but domain gap reduces benefits | [15, 8] |
| Scratch training | Medical tasks | Can match pre-trained models under specific conditions | [8, 16] |

### 2.1 CNNs in Medical Imaging: Achievements and Limitations

Convolutional neural networks (CNNs) have served as the cornerstone of medical image analysis for over a decade, demonstrating remarkable success in both anatomical imaging modalities (e.g., CT/MRI [17]) and functional modalities like OCT [9]. The hierarchical feature learning mechanism of CNNs enables effective pattern recognition when trained on large-scale datasets, with Kermany et al. [10] achieving 96.6% classification accuracy on OCT scans using over 100,000 annotated samples.

However, three fundamental challenges persist in clinical deployment scenarios:
1) **Data scarcity constraints**: Most medical institutions possess OCT datasets below 5,000 samples, leading to over 15% accuracy degradation compared to ideal training conditions as shown by Ran et al. [18];
2) **Domain-specific discrepancies**: OCT's inherent characteristics—including coherent speckle noise patterns and micron-scale retinal layer structures—result in approximately 40% feature mismatch with natural images [19];
3) **Local receptive field limitations**: Conventional 3×3 convolution kernels struggle to model the long-range spatial dependencies inherent in OCT's multi-layered anatomical structures [20].

Current solutions employ data augmentation via GANs [11] and deeper network architectures. Yet even 152-layer ResNet variants yield diminishing returns, with Ran et al. [21] reporting only 2.1% accuracy improvement despite quadrupling parameters. This suggests intrinsic limitations in CNN's local inductive bias for OCT analysis, motivating exploration of architectures with global context modeling capabilities.

### 2.2 Vision Transformers and Their Role in Medical Imaging

While convolutional neural networks (CNNs) have long dominated medical image analysis,ViTs [2] have recently emerged as a powerful alternative, leveraging a self-attention mechanism to capture rich global context across an image without relying strictly on local convolutional filters. In OCT, where high-resolution cross-sectional images often display subtle structural details and noise patterns, ViTs can be particularly advantageous. By treating OCT scans as sequences of non-overlapping patches, ViTs learn long-range dependencies and inter-patch relationships, which are essential for identifying small or diffuse lesions and complex tissue boundaries.

Recent studies highlight the effectiveness of ViTs in diverse OCT imaging tasks. For instance, the **Swin-Poly Transformer**, which integrates multi-scale feature modeling with polynomial loss optimization, has demonstrated superior performance in classifying retinal diseases from OCT images, achieving an accuracy of 99.80% and an AUC of 99.99% [22]. Another work introduces the **TESR** network (Transformer-based OCT retinal image super-resolution), which employs an *edge enhancement* module to emphasize layer boundaries, boosting the clarity of high-frequency features [23]. These specialized architectures demonstrate that ViT-style models can not only perform classification but also excel at super-resolution and lesion detection, tasks that demand precise global understanding of retinal or tissue structures.

Overall, ViTs offer a flexible, attention-driven way to encode both the local and global cues crucial in OCT images, whose morphological features may be distributed throughout the scan. Unlike purely convolutional models, ViTs are less constrained by receptive field sizes and can therefore capture subtle, global variations — an important property for applications like early disease detection and high-fidelity image reconstruction. As OCT datasets continue to grow and

diversify, domain-tailored ViT architectures stand poised to play an increasingly prominent role in advancing AI-aided medical imaging.

Nevertheless, ViT's data-hungry nature poses challenges for medical imaging, where annotated datasets are often limited. To address these issues, hybrid or specialized Transformer architectures have been proposed, combining convolutional layers with attention blocks to strike a balance between local and global feature learning. Recent studies also explore self-supervised or large-scale pre-training tailored for medical images, showing promising results in reducing domain gaps and improving downstream tasks [24].

### 2.3 Pre-training Strategies in Medical Imaging

Transfer learning from natural image benchmarks (e.g., ImageNet) is widely adopted to address medical data scarcity, yet its efficacy diminishes in specialized modalities like OCT due to profound domain gaps [25].The following three points further summarize the differences caused by OCT image properties during cross-domain transfer:(1) imaging physics (interferometric signals vs. photon reflection) [26], (2) structural patterns (layered tissues vs. object-centric scenes) [27, 28], and (3) pathological scale (micron-level lesions vs. macroscopic features) [29]. While generic pre-training accelerates convergence, studies report marginal gains in OCT tasks, as natural image priors fail to capture domain-specific attributes like speckle noise and anisotropic dependencies [30, 31].

To bridge this gap, we propose cross-anatomy OCT pre-training, leveraging shared imaging mechanisms across anatomical contexts. By pre-training on breast OCT data (adipose vs. cancer), models acquire transferable features—tissue texture sensitivity, noise robustness, and micron-scale pattern recognition—directly applicable to retinal OCT [21]. This strategy aligns with evidence from CT/MRI hybrid architectures, where domain-aligned pre-training outperforms ImageNet initialization despite anatomical differences. Our study directly addresses this gap by evaluating the efficacy of cross-anatomy OCT pre-training strategies, offering a scalable solution to overcome data scarcity while preserving OCT-specific diagnostic priors.

## 3 Materials and Methods

### 3.1 Dataset and Preprocessing

This study utilizes three datasets for evaluation. The first is a publicly available retinal OCT dataset introduced by Kermany et al. [10], which includes four categories of retinal pathologies: Choroidal Neovascularization (CNV), Diabetic Macular Edema (DME), Drusen, and Normal. The second dataset comprises breast OCT images,categorized into two classes: Adipose Tissue and Cancer. The third dataset is ImageNet21K, a large-scale natural-image dataset consisting of approximately 14 million images distributed across roughly 21,000 classes.

To systematically evaluate model performance, we define clear experimental settings based on the datasets:

– **Retinal OCT dataset (4 classes)**: The training set contains 360 images per class (1440 images total), trained for 200 epochs, with a test set of 90 images per class (360 images total).

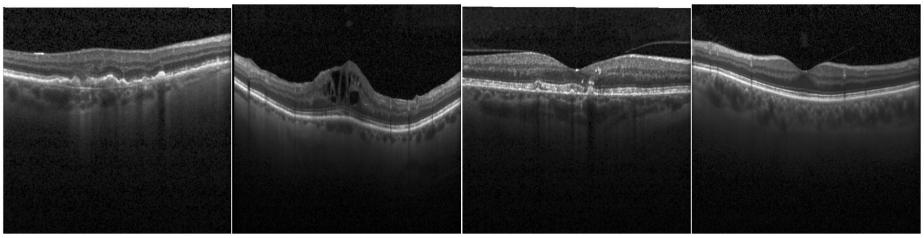


Fig. 1: Examples of the four categories from the retinal OCT dataset: left to right—CNV, DME, Drusen, and Normal

– **OCT breast dataset (2 classes)**: The training set contains 5000 images per class (10,000 images total), and the test set contains 1000 images per class (2000 images total).

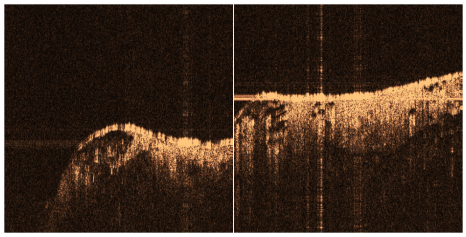


Fig. 2: Sample images from the OCT breast cancer dataset: left to right—Cancer and Adipose Tissue

– **ImageNet21K pre-trained model (21,000 classes)**: We utilize publicly available pre-trained weights originally trained on approximately 14 million images across diverse categories such as animals, vehicles, plants, and common objects.

Fig. 3: Samples from ImageNet21K categories: left to right—Chair, Bike, Bird, and Pig

All OCT images (both retinal and breast) are resized or cropped to a standardized resolution of 224×224 pixels, maintaining their single-channel (grayscale) characteristic. Basic data augmentation, such as random flips, small rotations, and ViT-based denoising techniques, is applied to enhance model generalization unless otherwise specified.

### 3.2 Implementation Details

**Vision Transformer Setup** We adopt the standard ViT architecture described in [2], where each image is divided into non-overlapping $16 \times 16$ patches. Each patch is linearly projected into embeddings, combined with positional encodings, and processed through multiple Transformer encoder layers. We specifically evaluate three initialization strategies:

- **ViT-ImageNet21K**: ViT pre-trained on the large-scale ImageNet21K dataset, then fine-tuned on the Retina OCT dataset.
- **ViT-OCT-Breast**: ViT pre-trained on OCT Breast images, then fine-tuned on the Retina OCT dataset. Pre-trained weights were loaded excluding the classifier head to ensure effective transfer of backbone features.
- **ViT-Scratch**: ViT trained directly from random initialization on the Retina OCT dataset.

**Training Protocol** We utilize the AdamW optimizer with a base learning rate of $1 \times 10^{-3}$ and a batch size of 32. The models are trained for 200 epochs without applying early stopping, as empirical observations suggested that training loss plateaued in later epochs, indicating sufficient convergence. A learning rate decay strategy (e.g., reducing the learning rate upon significant plateauing of validation accuracy) was applied to ensure more stable and robust convergence. All OCT images are first pre-processed into grayscale (single-channel) format prior to patch embedding.

**Evaluation Metrics** We evaluate the model performance using multiple complementary metrics:

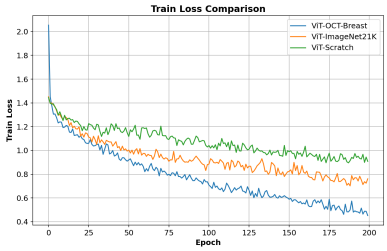
- **Accuracy (%)**: $\frac{\text{correct predictions}}{\text{total samples}} \times 100$ .Overall classification accuracy
- **Loss (Cross-Entropy)**: Cross-entropy loss measures the discrepancy between predicted probabilities and the true class distributions, guiding model optimization by penalizing uncertain or incorrect predictions. We use this loss not only to monitor convergence but also as the primary criterion for deciding on learning rate adjustments during training.Lower values indicate closer alignment between the predicted probability $\hat{p}_c$ and the true label $y_c$, guiding gradient-based optimisation during training.
- **Confusion Matrices**: Visualise the counts of true vs. predicted labels, pinpointing per-class strengths and systematic misclassifications.

– **ROC Curves and AUC**: ROC curves plot the true-positive rate (TPR) against the false-positive rate (FPR) across classification thresholds. The AUC (Area Under the Curve) summarizes overall discriminative ability, where 1.0 indicates perfect separability and 0.5 implies random guessing. For this multi-class tasks, one-vs-rest AUC is reported.

## 4 Experiments and Results

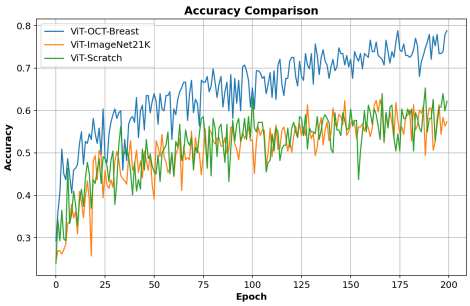
### 4.1 Loss and Accuracy

As illustrated in figure 4 , the ViT-OCT-Breast model consistently achieves lower training and test loss values and significantly higher accuracy than both ViT-ImageNet21K and ViT-Scratch. Specifically, the ViT-OCT-Breast achieves a final accuracy around 80%, outperforming ViT-ImageNet21K and ViT-Scratch, which stabilize near 55% and 50%, respectively.




(a) Training loss comparison


(b) Test loss comparison




(c) Accuracy comparison

Fig. 4: Comparison among ViT-OCT-Breast, ViT-ImageNet21K, and ViT-Scratch over 200 epochs.

Additionally, we note the relative performance of ViT-ImageNet21K and ViT-Scratch. ViT-ImageNet21K demonstrates marginally better performance

compared to ViT-Scratch, reflected in slightly lower training and test loss values and modestly higher accuracy. However, this advantage remains minimal, underscoring the limited benefit of ImageNet21K pre-training for highly specialized OCT image classification tasks.

It's important to note that the poor performance of the ImageNet-pretrained model in this task doesn't mean the model itself is bad. In fact, a ViT pretrained on ImageNet21K performs very well on general datasets—about 98.7% on CIFAR-10, 92.3% on CIFAR-100. This shows that domain relevance matters—a strong model may still perform poorly if it's not adapted to the specific type of data.

### 4.2 Confusion Matrices

Confusion matrices provide insights into the specific types of errors made by the classification models, showing how often each class is correctly identified and how often it is misclassified into other categories. High values along the diagonal indicate accurate predictions, while off-diagonal values represent misclassifications.

Figure 5 illustrates the confusion matrices for ViT-OCT-Breast, ViT-ImageNet21K, and ViT-Scratch after 200 epochs.

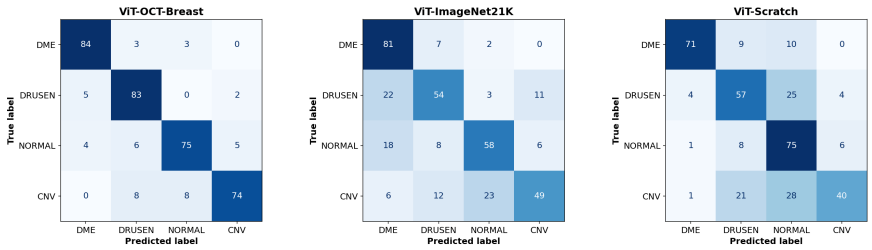



Fig. 5: Confusion matrices comparison for ViT-OCT-Breast, ViT-ImageNet21K, and ViT-Scratch on the small-scale 4-class OCT dataset.

From the confusion matrices, ViT-OCT-Breast shows superior class-wise accuracy, particularly in accurately classifying challenging classes such as DRUSEN and CNV, with very few misclassifications. In contrast, ViT-ImageNet21K and ViT-Scratch show notably higher rates of misclassification across multiple classes, further emphasizing the effectiveness of OCT-specific pre-training in enhancing class-wise discriminative performance.

### 4.3 ROC Curves and AUC

The Receiver Operating Characteristic (ROC) curve visualizes the trade-off between true positive rate (sensitivity) and false positive rate (1-specificity) across

different classification thresholds. The Area Under the Curve (AUC) metric quantifies the overall performance, where an AUC of 1 indicates perfect classification, and an AUC of 0.5 suggests no discriminative capability.

Figure 6 provides a detailed view of the ROC curves and corresponding AUC values for each of the four classes (DME, DRUSEN, NORMAL, CNV).

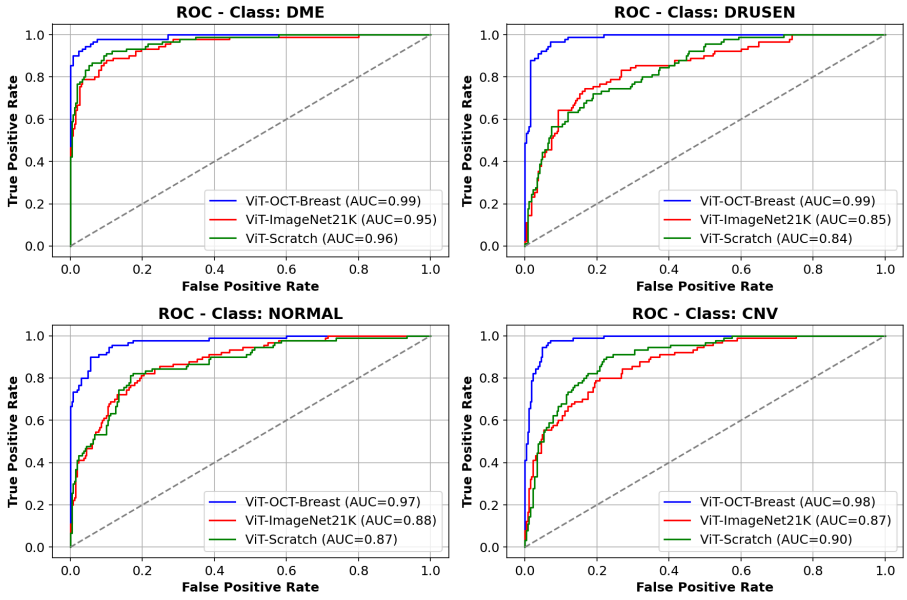



Fig. 6: ROC curves and AUC values comparison for ViT-OCT-Breast, ViT-ImageNet21K, and ViT-Scratch across the four OCT classes.

As shown in the ROC curves, ViT-OCT-Breast consistently achieves superior AUC values, close to or above 0.98 across all classes, significantly outperforming ViT-ImageNet21K and ViT-Scratch, especially in classes DRUSEN and CNV. ViT-ImageNet21K and ViT-Scratch show lower AUC values, often below 0.90, particularly evident in classes DRUSEN and NORMAL. This demonstrates the clear benefit of domain-specific pre-training in achieving robust class-wise predictive performance in specialized medical imaging classification tasks.

To further evaluate model robustness across data scales, systematic experiments were conducted on datasets containing 20, 50, 100, 360, 500, and 1,000 images. Key findings include: (1) In few-shot scenarios (e.g., 20 images), accuracy improved from the baseline 25% (ViT-Scratch and ViT-ImageNet21K) to 50% (ViT-OCT-Breast); (2) On the 1,000-image dataset, ViT-OCT-Breast achieved 83% accuracy within only 20 epochs, significantly outperforming 55% (ViT-Scratch and ViT-ImageNet21K). The 360-image configuration was selected as the primary experimental benchmark, as it effectively demonstrates small-data

improvements while avoiding performance interference from pretraining advantages observed in larger datasets.

## 5 Discussion

### 5.1 The Value of Domain-Specific Pre-training

Our results clearly demonstrate the superiority of domain-aligned pre-training over both generic ImageNet initialization and training from scratch. The ViT-OCT-Breast model consistently outperformed other variants in accuracy, AUC, and class-wise discrimination. This indicates that even OCT images from anatomically unrelated domains—such as breast tissue—can significantly improve model performance when their imaging modalities share structural and physical characteristics (e.g., speckle noise, layer boundaries, and micron-level resolution).

While the ViT-ImageNet21K model exhibited faster initial convergence, its final performance was only marginally better than the randomly initialized ViT-Scratch model. This underscores the limited utility of natural-image pre-training in highly specialized domains like OCT, where imaging physics and texture characteristics deviate substantially from conventional visual datasets.

In contrast, training from scratch (ViT-Scratch) struggled to learn effective representations in low-data regimes, leading to higher misclassification rates and lower AUCs across all four classes. This highlights the importance of meaningful inductive priors—especially when training data is scarce.

### 5.2 Data Scale and Domain Gap: Key Determinants

The performance gap between ViT-OCT-Breast and other models becomes especially prominent under limited-data conditions. In few-shot scenarios (e.g., 20 images per class), ViT-OCT-Breast achieved nearly double the accuracy of ViT-Scratch, suggesting strong feature transferability from domain-aligned OCT sources. These findings challenge the common notion that large-scale generic pre-training suffices for medical imaging and instead emphasize that *domain relevance outweighs data scale* when deploying Vision Transformers in clinical applications.

Our study also demonstrates that anatomical similarity is not a prerequisite for effective transfer. Despite the differences between breast and retinal tissue, their shared OCT acquisition principles enable useful feature extraction and transfer, supporting the hypothesis that *modality-specific alignment* is more critical than anatomical correspondence.

### 5.3 Limitations and Future Directions

Although this study provides evidence for the effectiveness of cross-anatomy domain-aligned pre-training, broader validation across diverse OCT tasks remains an important future step. One practical limitation is the difficulty of

acquiring large, annotated OCT datasets from other anatomical sites such as skin, cornea, or gastrointestinal tissues, due to clinical access restrictions, patient variability, and the high cost of expert labeling. These factors currently hinder broader experimentation but also highlight the potential impact of methods that can transfer knowledge across OCT modalities.

## 6 Conclusion

This study investigated whether domain-specific pre-training could enhance the performance of Vision Transformers in OCT image classification. Through extensive experiments across multiple training regimes, our findings lead to the following key conclusions:

- Pre-training ViTs on domain-aligned OCT data—even from unrelated anatomical regions—substantially improves classification accuracy (up to 80%) and AUC ($\geq 0.98$) in small-data retinal OCT settings.
- In contrast, ImageNet-based pre-training offers minimal performance gains, reaffirming the challenge of domain shift in medical imaging tasks.
- Training from scratch remains a viable baseline but is significantly less effective in low-resource conditions, highlighting the value of meaningful inductive priors from domain-specific sources.

Our findings also suggest a promising direction: by pre-training on a large-scale OCT dataset from one anatomical site (e.g., breast), it is possible to significantly enhance performance on smaller datasets from another site (e.g., retina). This highlights the potential of building general-purpose OCT transformer backbones that can be efficiently fine-tuned for diverse clinical tasks. In settings where labeled data is scarce, such cross-anatomy transfer learning offers a scalable and practical solution for advancing medical image analysis.

**Ethical Statement:** Retina Images for DME and Drusen: These images are publicly available and were sourced from Kaggle, based on [10]. This dataset includes high-resolution retinal images specifically related to Diabetic Macular Edema (DME) and Drusen, utilized for various diagnostic model training purposes. Breast Cancer Images: The breast cancer images were obtained from the Kent Applied Optics Group (AOG) and the University of Nottingham.The ethical approval for the use of these images in research was rigorously obtained from the City Hospital Nottingham.

## References


1. Vaswani, A., Shazeer, N., Parmar, N., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems (NIPS), pp. 5998–6008 (2017)
2. Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (ICLR), (2021)
3. Usman, M., Zia, T., Tariq, A.: Analyzing transfer learning of vision transformers for interpreting chest radiography. Journal of Digital Imaging **35**(6), 1445–1462 (2022)
4. Nassif, N., Cense, B., Park, B., et al.: In vivo human retinal imaging by ultrahigh-speed spectral domain optical coherence tomography. Optics Letters, 29(5):480–482 (2004)
5. von der Emde, L., Saßmannshausen, M., Morelle, O., et al.: Reliability of retinal layer annotation with a novel, high-resolution optical coherence tomography device: a comparative study. Bioengineering, 10(4):438 (2023)
6. Liba, O., Lew, M.D., SoRelle, E.D., et al.: Speckle-modulating optical coherence tomography in living mice and humans. Nature Communications, 8(1):15845 (2017)
7. Abt, D.L., Fischer, K.M.: Resolving three-dimensional anisotropic structure with shear wave splitting tomography. Geophysical Journal International, 173(3):859–886 (2008)
8. He, K., Girshick, R., Dollár, P.: Rethinking ImageNet pre-training. In: IEEE International Conference on Computer Vision (ICCV), pp. 4918–4927 (2019)
9. Esteva, A., Kuprel, B., Novoa, R., et al.: Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639):115–118 (2017)
10. 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 (2018)
11. Frid-Adar, M., Diamant, I., Klang, E., et al.: GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. Neurocomputing, 321:321–331 (2018)
12. Touvron, H., Cord, M., Sablayrolles, A., Jégou, H., et al.: Going deeper with Image Transformers. In: IEEE International Conference on Computer Vision (ICCV), pp. 32–42 (2021)
13. Zhou, Z., Sodha, V., Pang, J., et al.: Models genesis. Medical Image Analysis, 75:102254 (2022)
14. Abràmoff, M.D., Garvin, M.K., Sonka, M.: Retinal imaging and image analysis. IEEE Reviews in Biomedical Engineering, 3:169–208 (2010)
15. Deng, J., Dong, W., Socher, R., et al.: ImageNet: A large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 248–255 (2009)
16. Raghu, M., Zhang, C., Kleinberg, J., Bengio, S.: Transfusion: Understanding transfer learning with applications to medical imaging. In: Advances in Neural Information Processing Systems (NIPS), pp. 3342–3352 (2019)
17. Litjens, G., Kooi, T., Bejnordi, B., et al.: A survey on deep learning in medical image analysis. Medical Image Analysis, 42:60–88 (2017)
18. Ran, A.R., Cheung, C.Y.: Deep learning-based detection of retinal diseases from OCT: A review. Asia-Pacific Journal of Ophthalmology, 10(3):268–276 (2021)
19. Li, M., Idoughi, R., Choudhury, B., et al.: Statistical model for OCT image denoising. Biomedical Optics Express, 8(9):3903–3917 (2017)

20. Luo, W., Li, Y., Urtasun, R., et al.: Understanding the effective receptive field in deep convolutional neural networks. Advances in Neural Information Processing Systems, 29 (2016)
21. Ran, A.R., Chan, P.P., Li, F., et al.: In-depth OCT imaging for automated classification using hybrid CNN-transformers. Ophthalmology Science, 1(4):100101 (2021)
22. He, J., Wang, J., Han, Z., et al.: An interpretable transformer network for the retinal disease classification using optical coherence tomography. Scientific Reports, 13(1):3637 (2023)
23. Lu, Y., Chen, M., Qin, K., Wu, Y., Yin, Z., Yang, Z.: Super-Resolution Reconstruction of OCT Image Based on Pyramid Long-Range Transformer. Chinese Journal of Lasers, 50(15):1507107 (2023)
24. Azizi, S., Mustafa, B., Ryan, F., et al.: Big self-supervised models advance medical image classification. In: IEEE International Conference on Computer Vision (ICCV), pp. 3458–3468 (2021)
25. Cheplygina, V.: Cats or CAT scans: Transfer learning from natural or medical image source data sets? Current Opinion in Biomedical Engineering, 9:21–27 (2019)
26. Huang, D., Swanson, E.A., Lin, C.P., et al.: Optical coherence tomography. Science, 254(5035):1178–1181 (1991)
27. He, Y., Carass, A., Liu, Y., et al.: Structured layer surface segmentation for retina OCT using fully convolutional regression networks. Medical Image Analysis, 68:101856 (2021)
28. Russakovsky, O., Deng, J., Su, H., et al.: ImageNet large scale visual recognition challenge. International Journal of Computer Vision, 115:211–252 (2015)
29. Drexler, W., Morgner, U., Ghanta, R.K., et al.: Ultrahigh-resolution ophthalmic optical coherence tomography. Nature Medicine, 7(4):502–507 (2001)
30. Guan, H., Liu, M.: Domain adaptation for medical image analysis: a survey. IEEE Transactions on Biomedical Engineering, 69(3):1173–1185 (2021)
31. Karthik, K., Mahadevappa, M.: Convolution neural networks for optical coherence tomography (OCT) image classification. Biomedical Signal Processing and Control, 79:104176 (2023)