



AI-CNet3D: An Anatomically-Informed Cross-Attention Network with Multi-Task Consistency Fine-tuning for 3D Glaucoma Classification

Roshan Kenia ¹, Anfei Li ², Rishabh Srivastava ¹, Kaveri A. Thakoor ^{1,2,3,4}

¹ Department of Computer Science, Columbia University, New York, NY, USA

² Department of Ophthalmology, Columbia University Irving Medical Center, New York, NY, USA

³ Department of Biomedical Engineering, Columbia University, New York, NY, USA

⁴ The Data Science Institute, Columbia University, New York, NY, USA

Abstract

Glaucoma is a progressive eye disease that leads to optic nerve damage, causing irreversible vision loss if left untreated. Optical coherence tomography (OCT) has become a crucial tool for glaucoma diagnosis, offering high-resolution 3D scans of the retina and optic nerve. However, the conventional practice of condensing information from 3D OCT volumes into 2D reports often results in the loss of key structural details. To address this, we propose a novel hybrid deep learning model that integrates cross-attention mechanisms into a 3D convolutional neural network (CNN), enabling the extraction of critical features from the superior and inferior hemiretinas, as well as from the optic nerve head (ONH) and macula, within OCT volumes. We introduce Channel Attention REpresentations (CAREs) to visualize cross-attention outputs and leverage them for consistency-based multi-task fine-tuning, aligning them with Gradient-Weighted Class Activation Maps (Grad-CAMs) from the CNN's final convolutional layer to enhance performance, interpretability, and anatomical coherence. We have named this model AI-CNet3D (AI-'See'-Net3D) to reflect its design as an Anatomically-Informed Cross-attention Network operating on 3D data. By dividing the volume along two axes and applying cross-attention, our model enhances glaucoma classification by capturing asymmetries between the hemiretinal regions while integrating information from the optic nerve head and macula. We validate our approach on two large datasets, showing that it outperforms state-of-the-art attention and convolutional models across all key metrics. Finally, our model is computationally efficient, reducing the parameter count by one-hundred-fold compared to other attention mechanisms while maintaining high diagnostic performance and comparable GFLOPS. Our code is available at [10.5281/zenodo.17082118](https://zenodo.org/record/17082118).

Keywords

glaucoma, optical coherence tomography (OCT), 3D deep learning, cross-attention, parameter efficiency, spatial consistency, volumetric visualization

Article informations

<https://doi.org/10.59275/j.melba.2025-8d4c>

Volume 3, Received: 2025-04-15, Published 2025-09-08

Corresponding author: rk3291@columbia.edu

©2025 Roshan Kenia. License: CC-BY 4.0



1. Introduction

Glaucoma is one of the leading causes of irreversible blindness worldwide (Steinmetz et al., 2021; Quigley and Broman, 2006). However, as a chronic condition, glaucoma has a slow and gradual onset and often shows no noticeable symptoms in its early stages. Regular eye exams are essential for its classification and treatment. As nerve fiber layer damage is thought to be one of the hallmarks of glaucoma, optical coherence tomography (OCT) has become a widely

used tool for glaucoma detection and diagnosis due to its ability to capture high-resolution 3D volumes of the optic nerve and retina (Geevarghese et al., 2021; Bussell et al., 2014). The raw 3D volumes are then preprocessed and formatted into 2D OCT reports that can facilitate clinical decision-making by ophthalmologists.

Current OCT reports extract the retinal nerve fiber layer (RNFL) and ganglion cell complex (GCC) thicknesses to help detect the presence and degree of glaucomatous damage. Traditionally, a RNFL defect with a spatially-correlated

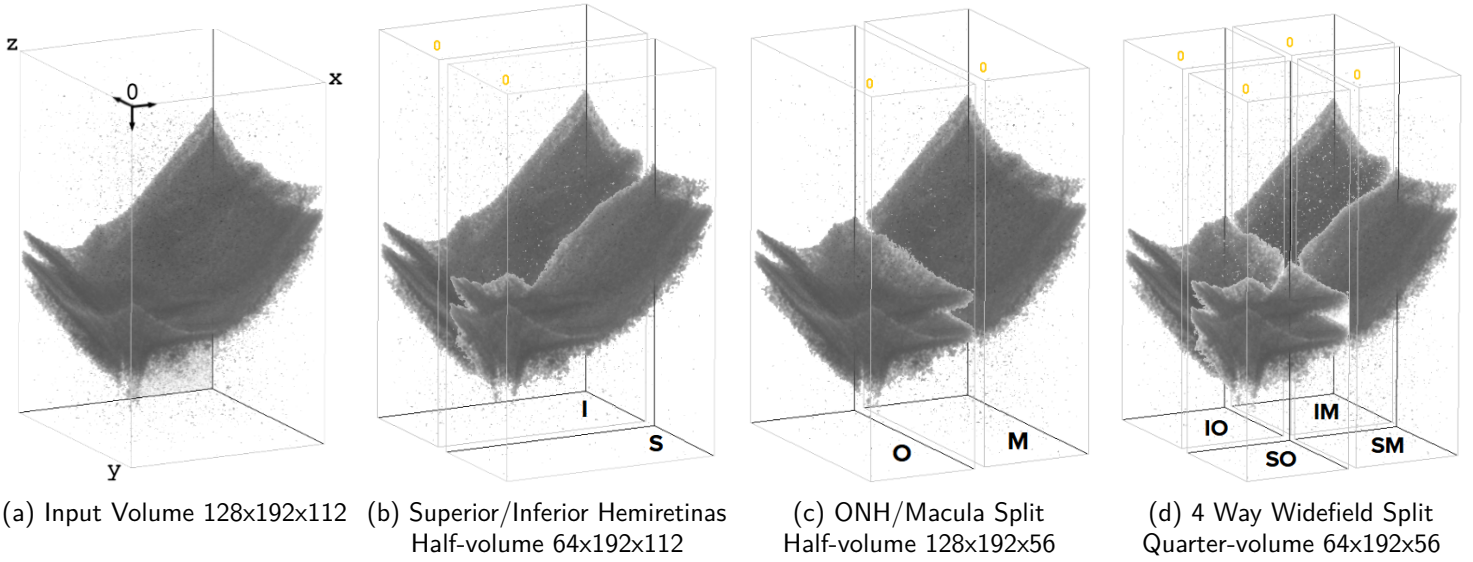


Figure 1: An example of how an OCT volume from Dataset 1 can be split along different axes to separate the anatomy in the volume. CA_H is computed using the inferior (I) and superior (S) hemiretinas. CA_{NA} is computed using the ONH (O) and macula (M). CA_{H-NA} is computed using the inferior ONH (IO), inferior macula (IM), superior ONH (SO), and superior macula (SM).

GCC defect is considered to represent an optic neuropathy, and a specific arced projection of this damage (called an arcuate) involving the superior and/or inferior optic nerve is commonly seen in glaucomatous optic neuropathy. Therefore, the ability to detect correlations between RNFL-GCC defects and correctly identify arcuate patterns are crucial for diagnosing glaucoma using OCT. With deep learning, there is potential to both automate this process and deliver near expert-level care to regions with poor access to glaucoma specialists. Furthermore, the traditional 2D OCT report has heavily relied upon superficial features such as the RNFL and GCC (Steinmetz et al., 2021; Chen et al., 2018). 3D models now allow for analysis of the entire 3D OCT volume that contains both superficial structures as well as previously unused deeper structures, which may lead to improvement in glaucoma classification over traditional approaches.

Deep learning has rapidly evolved into a powerful technology for automatically extracting features from images, enabling tasks such as detection, classification, and segmentation. While initially applied to 2D natural scene images, its capabilities have expanded to handle 3D volumes and video data, broadening its impact across various domains, from activity recognition to disease diagnosis (Tran et al., 2015; Bertasius et al., 2021; George et al., 2020). For glaucoma diagnosis, Maetschke et al. (2019) introduced one of the first 3D CNNs capable of classifying raw, unsegmented OCT volumes of the optic nerve head (ONH). They demonstrated the superiority of using a 3D deep learning model over classical feature-based machine learning algorithms. In addition, by computing Class Activation Maps

(CAM) (Zhou et al., 2016), they found the 3D CNN identified regions typically associated with glaucoma such as the neuroretinal rim, optic disc cupping, and the lamina cribrosa.

Convolutions within CNNs are highly effective at extracting local features (Yu and Koltun, 2016), but when applied to 3D data, their limited receptive fields can pose challenges. Information in 3D volumes is often sparsely distributed or spread over large regions (Ye and Liu, 2012), making it difficult for convolutions alone to capture global context. In contrast, transformer-based methods inherently provide global attention mechanisms, allowing for a more comprehensive understanding of the data. When combined with 3D convolutions, these approaches synergistically capture both local and global information, significantly enhancing feature representation (Shaker et al., 2024). However, this advantage afforded by incorporating attention comes with the drawback of significantly increased computational complexity, which is further amplified when dealing with volumetric data.

In medical imaging, obtaining high-quality data is inherently challenging due to limitations in acquisition, cost, and patient variability. Unlike natural image datasets, where large-scale labeled collections are readily available, medical datasets are often small, imbalanced, and difficult to annotate due to the requirement of expert clinical input. This scarcity of annotated data creates a significant barrier to training robust and generalizable deep learning models. Semi-supervised and unsupervised learning techniques have emerged as powerful solutions to mitigate these challenges by leveraging unlabeled data to enhance model performance.

Recent contrastive learning approaches, such as SimCLR (Chen et al., 2020) and BYOL (Grill et al., 2020), encourage representations of augmented views of the same sample to be similar in a shared latent space, reducing reliance on labeled data. In the medical domain, MedCLIP (Wang et al., 2022) extends contrastive learning by aligning medical images with textual descriptions, capturing richer semantic relationships.

Extending these capabilities to 3D is even more challenging due to the pronounced scarcity of valuable (labeled or unlabeled) 3D medical data. Authors of SLiViT (Avram et al., 2023) make this leap by leveraging 2D data from 3D OCT volumes via a 2.5D approach that enables robust performance across multiple tasks in three imaging modalities, even with fewer than 700 annotated volumes. Lee and colleagues do this too via their OCTCube approach (Liu et al., 2024b), where a 3D foundation model trained on over 26,000 OCT volumes is extended with contrastive learning to achieve state-of-the-art retinal disease prediction. Similarly, Swin UNETR (Tang et al., 2022) integrates multiple self-supervised strategies, such as inpainting, contrastive learning, and rotation correction, to learn more robust feature representations for 3D medical volumes. Our approach goes beyond these past studies by combining 3D OCT volumes, cross-attention mechanisms, and multi-task (supervised plus unsupervised) fine-tuning that enforces visualization consistency to improve model generalization, reduce annotation burdens, and enable AI-driven diagnostic tools that are both data-efficient and clinically reliable.

Volumetric OCT data contains key information for disease diagnosis oriented in a specific manner based on anatomy and pathophysiology. These features can be harnessed to create a more efficient and meaningful attention mechanism. The superior and inferior hemiretinas are the two parts of the retina that are divided by a horizontal line that runs through the fovea and ONH. Within each hemiretina, the ganglion cells send their axonal projections towards the ONH, forming the RNFL. To address the shortfalls of the conventional practice of condensing 3D OCT volume information into 2D reports, which results in the loss of key structural details for glaucoma classification, we propose a novel hybrid deep learning model. This model integrates cross-attention mechanisms into a 3D convolutional neural network (CNN), enabling the extraction of critical features from the superior and inferior hemiretinas, as well as from the optic nerve head (ONH) and macula, within OCT volumes. Additionally, as a fine-tuning and regularization step, we enforce consistency between visualizations from convolutional and attention layers, ensuring alignment between spatial feature extraction and transformer-based representations. Our contributions are as follows:

- We show the added benefit of our model, an Anatomically-

Informed Cross-attention Network operating on 3D data, AI-CNet3D (AI-‘See’-Net3D), achieved through the hybrid use of 3D CNNs with cross attention; by dividing the 3D volume and applying cross-attention, our model enhances glaucoma classification by capturing asymmetries between the hemiretinal regions while integrating information from the ONH and macula.

- We introduce a novel Channel Attention REpresentation (CARE), which provides direct visualization of channel attention outputs, offering a more precise and interpretable alternative to conventional Grad-CAM-based methods.
- By enforcing consistency between attention and convolutional visualizations, our model bridges the gap between these complementary feature extraction methods, improving robustness and interpretability.
- We validate our approach on two datasets (one proprietary and one publicly available), showing that it outperforms state-of-the-art attention models across all key metrics and conduct ablation studies which highlight the optimal positioning of attention within the 3D CNN architecture pipeline.
- Finally, our model is computationally efficient, reducing the parameter count by one-hundred-fold compared to other attention mechanisms while maintaining high diagnostic performance and comparable GFLOPS.

By leveraging anatomical priors, integrating CNNs with cross-attention, and enforcing consistency between feature representations, AI-CNet3D provides a more interpretable, efficient, and clinically relevant approach to 3D OCT analysis.

2. Related Works

2.1 Glaucoma classification

Glaucoma classification has progressed significantly with the integration of computational methods and deep learning-based approaches. Early work, such as that by Bock et al. (2010), applied appearance-based dimension reduction to color fundus images to develop a glaucoma risk index for classification. The introduction of deep learning marked a pivotal shift, with CNN-based methods becoming a standard for leveraging large datasets to enhance classification accuracy (Mehta et al., 2021; Barros et al., 2020). Chen et al. (2015) pioneered one of the first deep CNN architectures specifically for glaucoma classification from fundus images. Later, Li et al. (2019) extended CNN architectures by incorporating attention maps, enabling more focused analyses of critical retinal regions for improved interpretability.

Recent advances have shifted towards the utilization of RNFL data from OCT reports, supported by the findings of

Hood et al. (2022). Thakoor et al. (2020) developed a CNN model specifically for RNFL analysis from OCT images, employing concept activation vectors to compare model outputs with clinician eye fixations, adding a layer of clinical relevance. Additionally, Luo et al. (2023) introduced a large-scale OCT dataset to support semi-supervised learning, using a generalization-reinforced pseudo-labeling model to improve classification in cases with limited labeled data.

While the majority of existing research focuses on 2D imaging, recent progress by Maetschke et al. (2019) and George et al. (2020) has led to the development of CNN approaches for 3D OCT-based glaucoma classification, laying the groundwork for further innovation in 3D imaging modalities. However, these approaches have been validated on OCT volumes from only a single device manufacturer, leaving their generalizability across different imaging systems untested. Ensuring cross-manufacturer robustness is essential for the widespread clinical adoption of such methods, regardless of the OCT device used. We outline an approach for cross-manufacturer training as future work in Appendix A.5.

2.2 Cross-Attention and 3D Attention

In self-attention, the keys and values are derived from the same source as the queries, whereas in cross-attention, the keys and values come from a different source than the queries, allowing the model to focus on external information during processing (Vaswani, 2017; Lin et al., 2022). Chen et al. (2021) introduced CrossViT, that proposes a dual-branch transformer that processes image patches of varying sizes through separate branches, using multiple attention layers to fuse the tokens and enhance image features. To improve computational efficiency, they introduce a cross-attention-based token fusion module, where a single token from each branch serves as a query to exchange information between branches.

Modeling 3D attention is essential for tasks involving volumetric data, such as medical imaging, where spatial relationships extend across three dimensions. By capturing these interactions, 3D attention allows models to learn more complex spatial features with long-range anatomical dependencies, improving the accuracy and robustness of tasks like segmentation or classification (Islam et al., 2020; Wang et al., 2019b). Shaker et al. (2024) introduced Efficient Paired Attention (EPA), a computationally efficient method for calculating both spatial and channel self-attention in 3D volumes for segmentation tasks by using shared weights. They integrated EPA into a transformer block, utilizing it during both the downsampling and upsampling stages of a convolutional UNet, significantly enhancing performance while reducing computational costs. However, this reduction in computational costs introduced a significant bottleneck,

as a large feature volume is condensed into a small vector for spatial attention, creating a substantial constraint. As such, striking a balance between parameter efficiency and anatomical sensitivity remains a key challenge in designing attention modules for 3D medical tasks (Xie et al., 2023; Cao et al., 2022).

While standard 3D transformers and attention-augmented CNNs apply self-attention or spatial attention across entire volumes in a data-driven manner, our cross-attention mechanism is specifically constrained by retinal anatomy, computing attention only between anatomically meaningful regions (superior-inferior hemiretinas, macula-ONH pairs). Unlike generic attention mechanisms that must learn spatial relationships from scratch like EPA (Shaker et al., 2024), our approach embeds established medical knowledge directly into the architecture, enabling more efficient and clinically relevant feature learning.

2.3 3D Visualization

Maetschke et al. (2019) and George et al. (2020) achieved visualization of a 3D CNN model using 3D Grad-CAM, which highlights important regions of the input by back-propagating gradients from the class score to the final convolutional layer, computing the significance of feature maps, and generating a heatmap to identify key areas influencing the model's prediction. While convolutions excel at extracting local features, they may struggle to capture global context effectively, often resulting in sparse Grad-CAM heatmaps. For a clinician, it may be difficult to interpret the model's decision-making process using only 3D Grad-CAMs, as they are only applicable to convolutional layers and thus may not provide insights into the mechanisms of hybrid CNN-attention models.

Attention rollout was introduced as a post hoc method to trace how information flows from the input layer to the embeddings in higher layers of a transformer by calculating attention across multiple paths between nodes in different layers (Abnar and Zuidema, 2020). This is done by recursively multiplying attention weight matrices across layers, allowing for the total amount of information transferred between any two nodes to be computed. Chefer et al. (2021) extended attention visualization for computer vision classification tasks by developing a method that applies Deep Taylor Decomposition to assign local relevance scores, which are then propagated through the attention layers to improve interpretability.

There has been limited work on visualizing the components of 3D attention mechanisms. Typically, 3D attention is applied within a single layer, which restricts the use of advanced visualization methods such as attention rollout or Chefer et al. (2021), both of which require multi-layer attention propagation. This lack of suitable visualization

tools makes it challenging to interpret and analyze how attention mechanisms function in 3D models, limiting their transparency and usability in clinical applications.

2.4 Consistency-based Learning

Given the scarcity of labeled 3D medical imaging data, regularization techniques have been explored as a means of improving model generalization without relying on large-scale annotations. One particularly effective approach is visualization consistency, which enforces stability in the activations a model produces for a given input. By ensuring that different transformations of the same data yield consistent feature representations, these methods help reduce sensitivity to spurious variations, ultimately enhancing model robustness.

Prior work has predominantly focused on enforcing consistency within convolutional layers using natural images. For example, Guo et al. (2019) encourage consistency between the class activation maps (CAMs) from the last convolutional layer for both the original and augmented views of the same data. Similarly, Li et al. (2020) enforce consistency between images that share similar features, while Wang et al. (2019a) extend this idea by maintaining Grad-CAM consistency across different CNN layers for a single input. Xu et al. (2020) further generalize these ideas by enforcing consistency between learned attention maps across augmented images and multiple layers. More recently, Mirzazadeh et al. (2023) introduced an alternative approach by enforcing consistency between two different visualization techniques, Grad-CAM and Guided Backpropagation, to improve the quality of generated attention maps.

Despite these advancements, existing visualization consistency techniques remain fundamentally limited in scope, as they are primarily constrained to convolutional layers. With the increasing adoption of hybrid models that integrate convolutional layers with attention mechanisms, there is a pressing need to extend consistency-based learning beyond traditional CNNs. Unlike convolutional models, attention mechanisms dynamically reweight feature importance across an entire image or volume, making them susceptible to different types of instability. By enforcing consistency not only within CNN-based feature maps but also across attention outputs and hybrid representations, we can ensure that the model maintains spatial and anatomical coherence potentially even in low-data/low-labeled-data regimes. In the context of 3D medical imaging, where anatomical structures and pathological regions vary significantly, consistency-based learning offers a promising pathway to improve model interpretability, reduce sensitivity to noise, and enhance generalization across unseen cases.

2.5 Resource-Efficient Networks

Reducing parameter count and memory usage in 3D medical imaging models is critical due to the high computational demands imposed by volumetric data, which are often orders of magnitude greater than those of 2D images. While recent approaches have improved model speed during training and inference (Shaker et al., 2024; Pang et al., 2023; Liu et al., 2024a), they often overlook parameter efficiency, which is equally important for deployment. In portable or point-of-care settings, many applications require real-time inference on devices with limited computational resources (Shaker et al., 2023). Moreover, large models not only hinder deployment in such environments but also increase the risk of overfitting, especially when trained on small annotated medical datasets (Shaikhina and Khovanova, 2017). To address these challenges, efficient architectures that compress spatial and channel information while preserving critical anatomical context are essential for practical, scalable, and clinically viable solutions.

3. Datasets

Dataset 1 is comprised of 4,932 non-glaucomatous and 272 glaucomatous widefield OCT volumes obtained from Topcon Healthcare, Inc. (Tokyo, Japan) and labeled by OCT experts at Columbia University Irving Medical Center. To create a balanced dataset, we randomly sampled 272 non-glaucomatous volumes during each training trial, resulting in a total of 544 volumes with an equal split of 50% non-glaucomatous and 50% glaucomatous cases. This design choice was motivated by early observations that all models tended to overfit to the majority class when trained on imbalanced data even with resampling. For completeness, we provide comparative results of alternative sampling strategies attempted in Appendix A.4 and Table 8. The original volume dimensions were 128x885x512 (z, y, x) pixels; however, to reduce computational complexity while preserving the y-x aspect ratio, we downsampled the volumes using a uniform scaling affine transformation to 128x192x112 pixels. To ensure a robust evaluation, we divided the dataset into 65% for training, 15% for validation, and 20% for testing, providing a well-balanced selection of samples for model assessment.

Dataset 2 is a publicly available dataset provided by Maetschke et al. (2019) consisting of OCT scans centered on the optic nerve head (ONH), acquired from 624 patients using a Zeiss Cirrus SD-OCT scanner (Jena, Germany). After excluding scans with a signal strength below 7, 1,110 high-quality scans were retained for analysis, with 263 scans labeled as healthy and 847 diagnosed with primary open-angle glaucoma (POAG). We will refer to the healthy as non-glaucomatous and the POAG as glaucomatous. Glau-

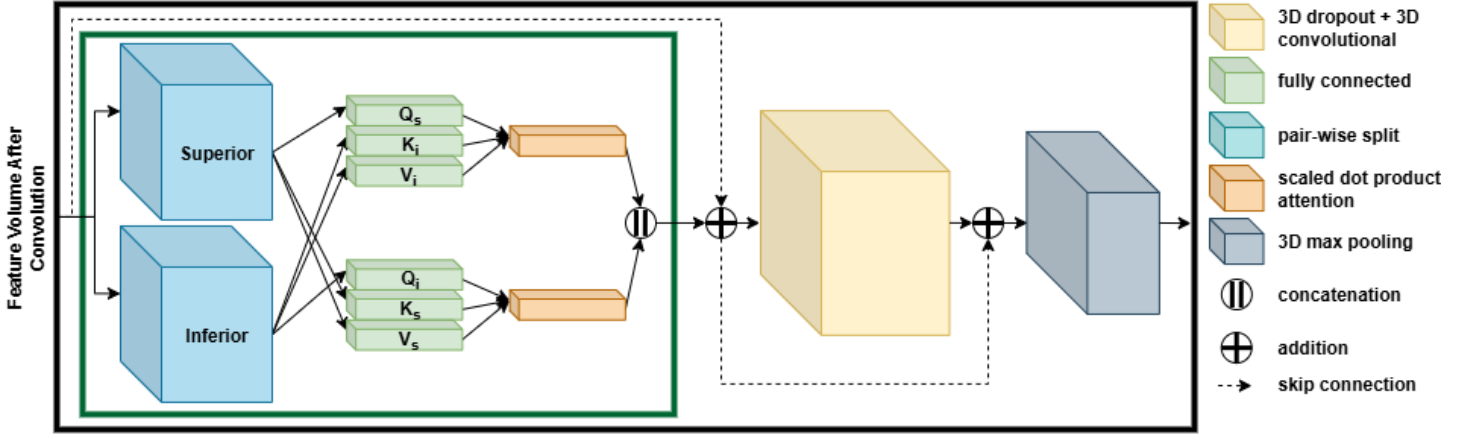


Figure 2: Our cross-attention mechanism operates between two pairs of subsections from the feature volume, as highlighted in the **green** box. In this example, we are computing cross-attention between the superior (S) and inferior (I) hemiretina split used for CA_H . For CA_{H-NA} (not visualized here), we would repeat the calculation performed in the **green** box for each pair of quarter-volumes and then concatenate the results before performing the skip connection addition.

comatous eyes were defined by the presence of visual field defects, confirmed by at least two consecutive abnormal test results. Just as with Dataset 1, we randomly sampled 263 glaucomatous volumes during each training trial, resulting in a total of 526 volumes with an equal split of 50% non-glaucomatous and 50% glaucomatous cases. We utilized the original volume dimensions of $64 \times 128 \times 64$ (z, y, x) pixels. The dataset was split into 65% training, 15% validation, and 20% testing, ensuring that scans from the same patient were not split across different sets and allowing more test data for evaluation.

4. Methods

4.1 Anatomically Informed Cross-attention

4.1.1 3D Superior-Inferior Cross-attention

When referring to the retina, the superior hemiretina denotes the nerve fibers originating from the upper portion of the retina, while the inferior hemiretina refers to those arising from the lower part. Together, they represent distinct sections of the retina, each responsible for transmitting visual information from the top and bottom halves of the eye, respectively. In cases of asymmetric glaucoma, either the superior or inferior hemiretina is typically affected. Despite this, no existing glaucoma classification models, to our knowledge, have leveraged this asymmetry to enhance classification. We introduce a novel 3D cross-attention mechanism that leverages the distinct information found in the superior and inferior hemiretinas within an OCT volume to improve glaucoma classification. A standard attention mechanism might struggle to fully capture the nuanced differences between these two regions. To address this, the feature volume can be split along the z-axis,

enabling cross-attention between the superior and inferior hemiretinas. This approach allows for relative comparison of the regions (e.g., a healthy inferior hemiretina serves as a reference for a superior hemiretina with disease or vice versa), thereby capturing the asymmetry and enhancing classification capabilities.

In Shaker et al. (2024), an efficient approach to computing spatial and channel attention for volumetric data is presented, leveraging shared key and query weights across the two separate self-attention calculations. For channel attention, the method applies a standard linear projection on the input volume, generating a channel value vector, followed by self-attention. Spatial attention, designed to minimize complexity, first applies a linear projection to form a spatial value vector. This vector is then projected down to a lower dimension, p , which is significantly smaller than the number of tokens, n . This adjustment reduces the computational complexity from $O(n^2)$ to $O(np)$, yet introduces a notable bottleneck. Specifically, projecting a volume of dimensions $C \times D \times H \times W$ into $C \times p$ entails substantial information loss, impacting the model's ability to capture critical spatial details. Additionally, this second projection layer introduces a high parameter count, as illustrated in Figure 6, further complicating model efficiency and potentially affecting scalability.

Therefore, to increase scalability while still enforcing our anatomical prior, we focus on using only channel attention. We found that projecting the entire spatial dimension of the feature volume into a small vector was not beneficial for model training (ablation studies in Appendix A.1). Instead of directly projecting the entire feature volume of size $C \times D \times H \times W$ into query, key, and value vectors for self-attention, we split the input feature volume, I_v , along the

z-axis into two feature volumes of size $C \times \frac{D}{2} \times H \times W$, representing the superior and inferior hemiretinas as shown in Figure 1b. In the first cross-attention step, the superior feature volume is projected into a query vector Q_s , while the inferior feature volume forms the key K_i and value V_i vectors. We then apply scaled dot product attention across the channel dimension, computed as:

$$A_{SI}(Q_s, K_i, V_i) = \text{softmax} \left(\frac{Q_s K_i^T}{\sqrt{d_k}} \right) V_i \quad (1)$$

where d_k is the dimensionality of the key vectors. In the second step, the roles are reversed: the inferior feature volume is projected as the query Q_i , while the superior feature volume forms the key K_s and value V_s vectors. Another round of scaled dot product attention is applied:

$$A_{IS}(Q_i, K_s, V_s) = \text{softmax} \left(\frac{Q_i K_s^T}{\sqrt{d_k}} \right) V_s \quad (2)$$

These two attention outputs are then concatenated together and reshaped to the original input size of $C \times D \times H \times W$ as shown in Fig. 2. A skip connection is then used to add the original I_v and results in our hemiretinal cross-attention, CA_H , as calculated in Equation 3. This bidirectional attention mechanism allows the model to capture critical interactions between the superior and inferior hemiretinas, improving the representation of the retinal structure. We call this model AI-CNet3D_H, illustrating its combination of information learned between the superior and inferior hemiretinas within the volume.

$$CA_H = (A_{SI} \parallel A_{IS}) \oplus I_v \quad (3)$$

4.1.2 Widefield (4-Way) Cross-attention

The majority of OCT datasets do not include a widefield view, but Dataset 1 used in this study and described above does. This means that it includes both the optic nerve head (ONH) and the macula within the volumetric scan of the retina. The distinction between ONH and macula is made by splitting the volume along the x-axis. As shown in Figure 1d, combining this with our split along the z-axis for the superior and inferior hemiretinas allows us to obtain four subvolumes from the input: the superior ONH (*SO*), superior macula (*SM*), inferior ONH (*IO*), and inferior macula (*IM*). We know from Section 4.1.1 that superior and inferior hemiretinas can be used to compute cross-attention. Furthermore, ONH and macular half-volumes are anatomically related, while opposing ONH and macular quarter volumes (e.g., superior ONH and inferior macula) are not anatomically related.

Leveraging this insight, we can enhance our 3D cross-attention mechanism for widefield OCT volumes. In Section 4.1.1, we computed cross-attention between the superior

and inferior hemiretinas. With widefield volumes, we extend this to compute cross-attention specifically between anatomically related pairs within *SO*, *SM*, *IO*, and *IM*. Given a quarter-volume, x , we project it to a query vector Q_x , while the other paired quarter-volume, y , forms the key K_y and value V_y vectors. We compute the scaled dot product attention, A_{xy} , just as in Equation 1, and then reverse roles using projected vectors Q_y , K_x , and V_x for our second scaled dot product attention calculation, A_{yx} , just as in Equation 2. We apply this same sequence of steps to the other two quarter-volumes, w and z , and obtain A_{wz} and A_{zw} . We can then compute the cross-attention for these two pairs, xy and wz , on our volume by concatenating the results as:

$$CA_{xywz} = (A_{xy} \parallel A_{yx}) \parallel (A_{wz} \parallel A_{zw}) \quad (4)$$

In this case, we compute cross-attention twice. Our first set of quarter volumes x , y , z , and w are represented by *SO*, *SM*, *IO*, and *IM* respectively and correspond to computing cross-attention, CA_{SupInf} within each hemiretina. Our second set of quarter volumes x , y , z , and w are represented by *SO*, *IO*, *SM*, and *IM* respectively and correspond to computing cross attention, CA_{MacONH} , within each macula and ONH half-volume. These operations are added together using a skip connection along with the original volume, I_v , as:

$$CA_{H-NA} = CA_{SupInf} \oplus CA_{MacONH} \oplus I_v \quad (5)$$

where $H - NA$ indicates using the hemiretinas and ONH and macula. We call this model AI-CNet3D_{H-NA}, illustrating its combination of information learned between the superior and inferior hemiretinas within the volume and between the macula and ONH.

4.1.3 3D Macula-Optic Nerve Head Cross-attention

For completeness, using our widefield volumes, we follow the same steps used in Section 4.1.1, but split our volume only once along the width axis (x-axis) into macula and optic nerve head (ONH) halves as shown in Figure 1c. This enables us to compute the cross-attention between the macula (neurons) and ONH (axons) as CA_{NA} , leveraging the distinct information presented in each. We call this model AI-CNet3D_{NA}, illustrating its combination of information learned between the macula and ONH within the volume.

4.2 3D CNN

We adopt the 3D CNN as described in Maetschke et al. (2019) and modify it to include our cross-attention mechanism. The original network consists of five 3D convolutional layers, each with ReLU activation and batch normalization.

The layers use filter banks of size 32-32-32-32-32, with filter dimensions of 7-5-3-3-3 and strides of 2-1-1-1-1. After the final convolutional layer, a global average pooling (GAP) layer is applied with a kernel size set to the smallest spatial dimension after downsampling, and a stride matching the final feature map shape, ensuring the entire spatial volume is reduced to a single value per channel. This adaptive configuration ensures spatially aware aggregation of features before classification. This is followed by a dense layer connected to the softmax output.

After the second and fourth convolutional layers (optimal position verified in Table 1), we introduce a cross-attention-based feature extraction block. The feature vector is first split into halves or quarters, which are processed through one of the three cross-attention mechanisms described previously, while maintaining a skip connection with the full feature vector. Inspired by Shaker et al. (2024), we apply a 3D dropout layer, followed by a 1x1x1 convolution with another skip connection to refine the feature representation. Finally, we employ 3D max pooling to downsample the attention maps, selectively retaining the highest activation values within each pooling region, which correspond to the most significant features.

4.3 Channel Attention REpresentation

To complement the 3D Grad-CAM output of the convolutional layers of our network, we introduce a new method called Channel Attention REpresentation (CARE) to visualize the 3D attention layers. This approach translates the attention output, denoted as CA_0 , into a volumetric representation aligned with the original input volume, thereby enabling interpretable visualization of critical features. The cross-attention output CA_0 is obtained post max-pooling (see Fig. 2), yielding the layer's most influential attention features. From this, we specifically isolate the channel dimension to capture attention features from multiple perspectives within the same spatial location. To condense this channel information, we compute the mean attention map by averaging CA_0 across the channel dimension C :

$$CA_1 = \frac{1}{C} \sum_{c'=1}^C CA_0(c', d, h, w) \quad (6)$$

The resulting CA_1 reduces dimensional complexity while preserving channel-wise aggregated attention weights. Next, to focus on features with positive contributions, analogous to Grad-CAM, we apply a rectified linear unit (ReLU) activation function to CA_1 , ensuring that only positive activations contribute to the visualization. For interpretability, the rectified attention map is max normalized to scale its values between 0 and 1, facilitating its use as a heatmap:

$$H_{CARE} = \frac{ReLU(CA_1)}{\max(ReLU(CA_1)) + \epsilon} \quad (7)$$

where ϵ is a small constant to prevent division by zero. This simple computation is performed at the last attention layer within the network and can be used with any of the mechanisms explained above. Since it is computed using a deeper network layer, the resulting $D \times H \times W$ dimensions may be smaller than those of the original volume. To address this, we apply 3D interpolation to rescale the H_{CARE} to match the input's original size only for visualizations (see Fig. 4), allowing us to overlay it and generate a clear, interpretable heatmap. We ensured that this operation was fully differentiable so that it could be utilized to train our model.

4.4 3D Grad-CAM

3D Gradient Weighted Class Activation Maps (3D Grad-CAMs) extend the original 2D Grad-CAM method (Selvaraju et al., 2020) to identify influential regions within a 3D volume, highlighting areas that significantly contribute to the model's class prediction. The gradient of the score y^c for class $c \in \{0, 1\}$ is computed with respect to a 3D feature map A^k at each voxel (i, j, d) . This gradient is globally averaged over the spatial dimensions to compute class-specific weights a_k^c for each feature map k :

$$a_k^c = \frac{1}{Z} \sum_i \sum_j \sum_d \frac{\partial y^c}{\partial A_{i,j,d}^k} \quad (8)$$

where Z is a normalization constant representing the total number of voxels. These weights are then used to scale the feature maps, emphasizing the most influential regions. The final 3D Grad-CAM heatmap $H_{Grad-CAM}$ is generated by summing the weighted feature maps across all channels and applying a rectified linear unit (ReLU) to ensure non-negative values. To normalize the heatmap, each voxel intensity is divided by the maximum value across the entire volume:

$$H_{Grad-CAM} = \frac{ReLU(\sum_k a_k^c A^k)}{\max(ReLU(\sum_k a_k^c A^k)) + \epsilon} \quad (9)$$

where ϵ is a small constant to prevent division by zero. This process ensures that the heatmap intensities are scaled between 0 and 1, making them in a visualizable range. The resulting heatmap highlights the regions within the 3D volume that contribute most strongly to predicting glaucoma presence, with warmer colors indicating greater influence on the model's decision (yellow represents highest influence).

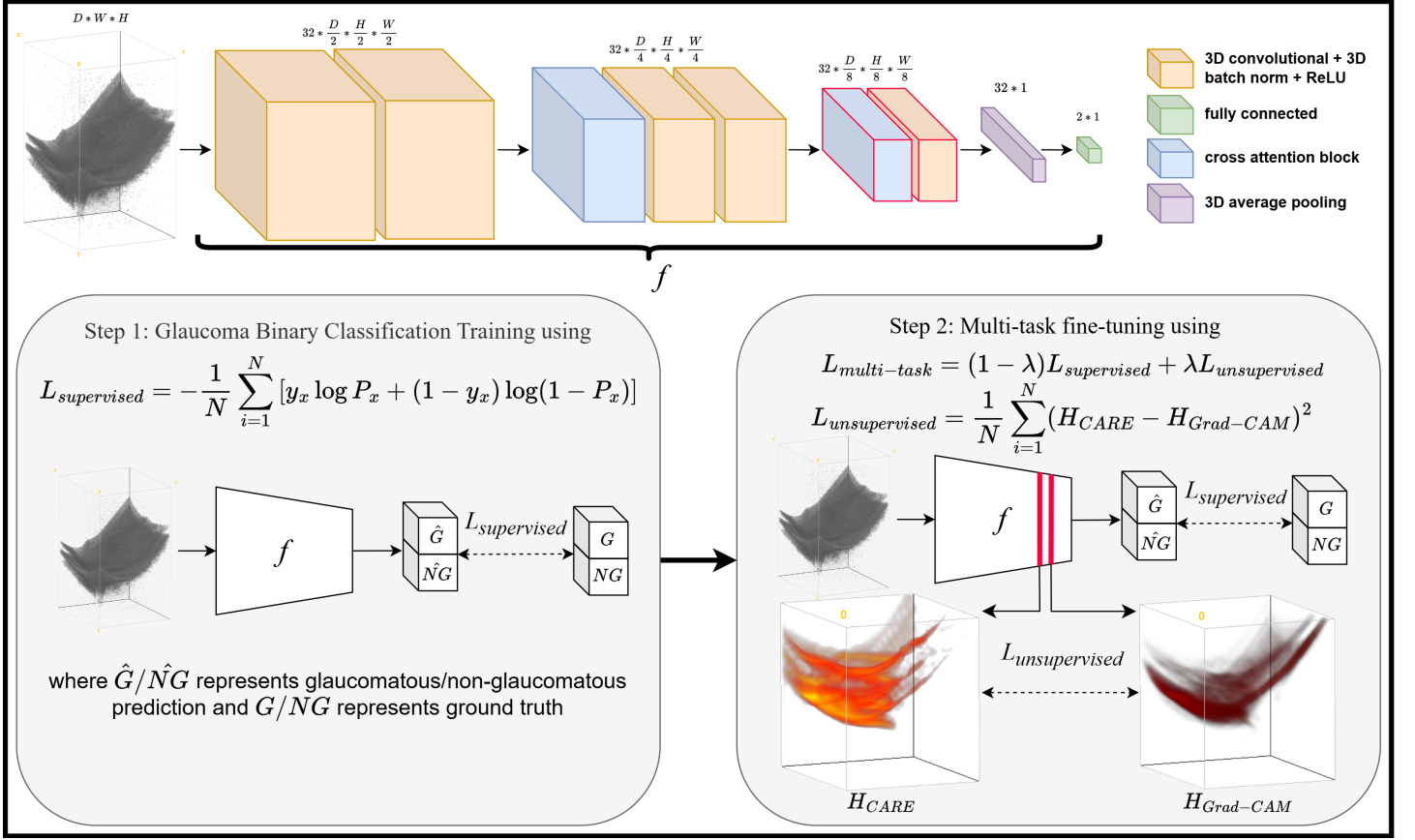


Figure 3: Visualization of our AI-CNet3D architecture (with the channel dimension omitted from visualization). We apply multiple layers of convolution along with two cross-attention blocks. Filter banks of size 32 are used consistently across the model. When training with multi-task fine-tuning, we utilize the last cross-attention and convolutional layers highlighted in red for alignment.

4.5 Multi-task Fine-tuning to Enforce Visualization-Based Consistency

We now present our method to enforce consistency between the attention and convolutional layers of our network. It is known from (Selvaraju et al., 2020) that the last convolutional layer of a CNN captures the most class-discriminative properties compared to earlier layers and therefore is best to use for visualization. We follow this trend and obtain the last convolutional layer’s output and the last attention layer output from our model. These outputs are of a much smaller dimension compared to the original input volume (due to downsampling for the convolutional layer and max-pooling for the attention layer), and thus they contain the most relevant information the model uses for classification.

Our goal is to enforce consistency between the hidden features extracted from the cross-attention module and the final convolutional layer, ensuring that features learned by one can be effectively shared with the other. Previous consistency-based loss methods have utilized Pearson correlation, Structural Similarity Index (SSIM), Kullback-Leibler (KL) divergence, and Mean Squared Error (MSE) to compare features. In practice, we found that MSE loss worked

best (ablation studies in Appendix A.3) to enforce feature alignment and improve model performance.

For a given input x , let H_{CARE} denote the final cross-attention hidden feature heatmap and $H_{Grad-CAM}$ represent the final convolutional layer hidden feature heatmap. We define the unsupervised consistency loss as:

$$L_{unsupervised} = \frac{1}{N} \sum_{i=1}^N (H_{CARE} - H_{Grad-CAM})^2 \quad (10)$$

Additionally, we compute the binary cross-entropy (BCE) loss between the model’s predicted glaucoma probability P_x and the ground truth y_x :

$$L_{supervised} = -\frac{1}{N} \sum_{i=1}^N [y_x \log P_x + (1 - y_x) \log(1 - P_x)] \quad (11)$$

We found that utilizing only the unsupervised loss while fine-tuning (ablation studies in Appendix A.2) leads to model degeneration in terms of classification performance. To jointly optimize for accurate classification and feature

Table 1: Results of ablation study with AI-CNet3D to determine the optimal placement of cross-attention blocks within the 3D CNN over three trials. Integrating cross-attention after the initial convolutions and before the final convolution yielded the best performance.

Cross-attention Placement	Avg. Acc. \pm Std.	Avg. AUROC \pm Std.
After conv 1 & 2	0.7982 \pm 0.0327	0.7980 \pm 0.0328
After conv 1 & 3	0.7951 \pm 0.0216	0.7945 \pm 0.022
After conv 2 & 3	0.7859 \pm 0.0312	0.7856 \pm 0.0294
After conv 2 & 4	0.8165 \pm 0.0375	0.8174 \pm 0.0353
After conv 2 & 5	0.7584 \pm 0.0189	0.7602 \pm 0.0154
After conv 3 & 4	0.7706 \pm 0.0259	0.7717 \pm 0.0246
After conv 3 & 5	0.8073 \pm 0.0417	0.8072 \pm 0.0405

consistency, we combine these losses into a multi-task objective function:

$$L_{multi-task} = (1 - \lambda)L_{supervised} + \lambda L_{unsupervised} \quad (12)$$

where λ is a weighting factor that controls the influence of the unsupervised consistency loss. This combined loss encourages the network to learn from labeled data while also enforcing consistency between the cross-attention module and the final convolutional layer without any labels, leading to improved feature robustness and interpretability.

4.6 Model Training and Data Augmentation

For both Dataset 1 and Dataset 2, the data is randomly shuffled and divided into training, validation, and test sets. Our model is trained for 250 epochs using a batch size of 4, a learning rate of 0.0001, and an early stopping patience of 25 epochs. We use the NAdam optimizer and Binary Cross-Entropy (BCE) loss highlighted in Equation 11. Each experiment conducted, for our model and the baseline models, is repeated five times, with the results for accuracy, specificity, sensitivity, AUROC, and F1-Score averaged across these runs.

Once we train our model, it can produce clear and meaningful Grad-CAMs and CAREs. We then fine-tune it for another 250 epochs using the multi-task unsupervised and supervised loss in Equation 12 to make these two visualizations consistent. We observed that if we attempted joint training from scratch, early incorrect visualizations were learned and propagated through the network for the rest of training. We utilized the same hyperparameters and data splits as our regular BCE training. We found through a hyperparameter search (ablation studies in Appendix A.2) that $\lambda = 0.75$ and 0.5 lead to the best results for Dataset 1 and 2, respectively.

During each epoch of model training and fine-tuning, we implement standard random grayscale augmentation and rescale the volumes by up to 1.25x. A widefield OCT scan contains information about both the macula and the

optic nerve head (ONH) in the retina. To leverage the symmetry between the right and left eyes in relation to the macula and nerve, we apply random reflections across the yz-plane. Additionally, to enhance the model's robustness to variations in the superior and inferior hemiretinas, we also apply random reflections across the xy-plane.

5. Results

To optimize attention placement, we conducted an ablation study by positioning our superior inferior cross-attention mechanism after various convolutional layers within the network. As shown in Table 3, the best performance occurred when cross-attention was applied after the second and fourth convolutions. This configuration leverages the convolutional layers' role in feature extraction, with early layers capturing small, local patterns like edges and circles, and later layers integrating these into more complex structures (Zeiler, 2014).

We benchmark against the TimeSFormer model from Bertasius et al. (2021), originally designed for video classification but adapted here for 3D volumes using a joint space-time self-attention mechanism. For a simpler self-attention baseline, we compare against a ViT (Dosovitskiy et al., 2020) adapted for 3D volumes. To evaluate against a strong convolutional architecture, we use SEResNeXt (Hu et al., 2018), which integrates Squeeze-and-Excitation blocks within a ResNeXt framework. We also compare our approach with the baseline 3D CNN from Maetschke et al. (2019). We further compare against M3T (Jang and Hwang, 2022), a multi-plane and multi-slice transformer network that combines 3D CNN, 2D CNN, and state-of-the-art 3D transformer architectures to leverage both local inductive biases and global attention relationships across axial, coronal, and sagittal planes for classification. We lastly evaluate against Med3D (Chen et al., 2019) with ResNet34 backbone, a heterogeneous 3D network pre-trained on the diverse 3DSeg-8 dataset that demonstrates superior transfer learning capabilities for 3D medical imaging tasks compared to models pre-trained on natural image datasets.

For comparison with a standard spatial and channel attention methods we replace our 3D cross-attention mechanism with the Efficient Paired Attention (EPA) algorithm from Shaker et al. (2024). Since this architecture is also a hybrid CNN-attention method, we can follow the same steps detailed in Section 4.3 to compute attention representations for the joint spatial and channel attention mechanisms used in EPA. These representations can then be used to perform consistency-based fine-tuning, allowing us to compare performance against our AI-CNet3D models in Table 3.

Table 2: Performance evaluation of the baseline 3D CNN, the hybrid EPA CNN model, TimeSformer, ViT, SE-ResNeXt, M3T, Med3D, and our proposed hybrid cross-attention CNN models on Dataset 1 (Topcon) and Dataset 2 (Zeiss). We report p-values for Mann Whitney U tests with our AI-CNet3D_H model in parentheses after the standard deviations.

	Model Type	Avg. Test Acc. \pm Std.	Avg. Test Spec. \pm Std.	Avg. Test Sens. \pm Std.	Avg. Test AUROC \pm Std.	Avg. Test F1 Score \pm Std.
Topcon	ViT (Dosovitskiy et al., 2020)	0.6422 \pm 0.0682 (0.008)	0.7241 \pm 0.1803 (0.151)	0.5556 \pm 0.2843 (0.032)	0.6398 \pm 0.0772 (0.008)	0.5503 \pm 0.2755 (0.008)
	TimeSformer (Bertasius et al., 2021)	0.7670 \pm 0.0336 (0.094)	0.7726 \pm 0.0639 (0.421)	0.7614 \pm 0.0729 (0.151)	0.7670 \pm 0.0313 (0.095)	0.7693 \pm 0.0300 (0.032)
	SEResNeXt50 (Hu et al., 2018)	0.7908 \pm 0.0220 (0.248)	0.8264 \pm 0.0246 (0.841)	0.7571 \pm 0.0451 (0.151)	0.7917 \pm 0.0193 (0.222)	0.7871 \pm 0.0208 (0.209)
	M3T (Jang and Hwang, 2022)	0.6532 \pm 0.0927 (0.016)	0.5942 \pm 0.2967 (0.056)	0.7143 \pm 0.1492 (0.151)	0.6542 \pm 0.0919 (0.016)	0.6785 \pm 0.0585 (0.008)
	Med3D (Chen et al., 2019)	0.6220 \pm 0.0825 (0.008)	0.4282 \pm 0.2220 (0.008)	0.8179 \pm 0.1288 (0.691)	0.6230 \pm 0.0728 (0.008)	0.6886 \pm 0.0452 (0.016)
	Base 3D CNN (Maetschke et al., 2019)	0.6073 \pm 0.0875 (0.008)	0.6062 \pm 0.3395 (0.421)	0.6122 \pm 0.3539 (1.000)	0.6092 \pm 0.0907 (0.008)	0.5453 \pm 0.2526 (0.008)
	EPA (Shaker et al., 2024)	0.7872 \pm 0.0404 (0.346)	0.8235 \pm 0.0372 (0.917)	0.7539 \pm 0.0593 (0.310)	0.7887 \pm 0.0385 (0.421)	0.7831 \pm 0.0429 (0.463)
	AI-CNet3D _{NA} (Ours)	0.8037 \pm 0.0356	0.7992 \pm 0.0669	0.8076 \pm 0.0253	0.8034 \pm 0.0373	0.8090 \pm 0.0249
	AI-CNet3D _{H-NA} (Ours)	0.7945 \pm 0.0356	0.7936 \pm 0.0820	0.7921 \pm 0.0457	0.7928 \pm 0.0359	0.7987 \pm 0.0233
	AI-CNet3D _H (Ours)	0.8183 \pm 0.0340	0.8290 \pm 0.0604	0.8063 \pm 0.0203	0.8176 \pm 0.0339	0.8204 \pm 0.0288
Zeiss	ViT (Dosovitskiy et al., 2020)	0.7520 \pm 0.1167 (0.222)	0.8077 \pm 0.1190 (0.600)	0.6797 \pm 0.3416 (1.000)	0.7437 \pm 0.1269 (0.151)	0.6531 \pm 0.3276 (0.421)
	TimeSformer (Bertasius et al., 2021)	0.8179 \pm 0.0479 (0.691)	0.8440 \pm 0.0714 (0.463)	0.7963 \pm 0.1073 (0.600)	0.8202 \pm 0.0455 (0.548)	0.8102 \pm 0.0545 (1.000)
	SEResNeXt50 (Hu et al., 2018)	0.8143 \pm 0.0411 (0.691)	0.8611 \pm 0.0830 (0.548)	0.7794 \pm 0.1009 (0.173)	0.8203 \pm 0.0398 (0.841)	0.8048 \pm 0.0414 (0.421)
	M3T (Jang and Hwang, 2022)	0.7920 \pm 0.0124 (0.008)	0.8561 \pm 0.0601 (0.310)	0.7295 \pm 0.0516 (0.016)	0.7928 \pm 0.0078 (0.008)	0.7756 \pm 0.0117 (0.008)
	Med3D (Chen et al., 2019)	0.8183 \pm 0.0249 (0.841)	0.8694 \pm 0.0572 (0.346)	0.7692 \pm 0.0887 (0.310)	0.8193 \pm 0.0250 (0.548)	0.8054 \pm 0.0337 (0.310)
	Base 3D CNN (Maetschke et al., 2019)	0.8300 \pm 0.0359 (0.600)	0.8730 \pm 0.0896 (0.463)	0.7891 \pm 0.0695 (0.346)	0.8310 \pm 0.0311 (0.310)	0.8222 \pm 0.0280 (0.173)
	EPA (Shaker et al., 2024)	0.8162 \pm 0.0429 (0.691)	0.8379 \pm 0.0753 (0.917)	0.8038 \pm 0.0810 (0.151)	0.8209 \pm 0.0438 (0.841)	0.8122 \pm 0.0398 (0.421)
	AI-CNet3D _H (Ours)	0.8315 \pm 0.0105	0.8246 \pm 0.0336	0.8425 \pm 0.0388	0.8336 \pm 0.0130	0.8311 \pm 0.0134

5.1 Anatomically Informed Cross-attention

In Table 2, we compare our 3D channel-wise cross-attention approaches to other attention approaches along with the base CNN presented in Maetschke et al. (2019). Examining Table 2 performance on our widefield Topcon Dataset 1, we can see our AI-CNet3D_{H-NA}, AI-CNet3D_H, and AI-CNet3D_{NA} models perform comparably or better than other non-cross-attention baseline approaches across all metrics. Our AI-CNet3D_H model performs comparably or significantly better on all metrics (except sensitivity) than all other models on Dataset 1, including the specialized medical models M3T and Med3D, which achieve lower performance across most metrics despite being designed for medical imaging tasks. While Med3D achieves a slightly higher sensitivity, it exhibits much greater variability, and notably, our AI-CNet3D_H model surpasses this performance when fine-tuned, as demonstrated in Table 3.

Our AI-CNet3D_H cross-attention mechanism consistently outperforms other methods across all metrics other than specificity on our ONH-only Zeiss Dataset 2. (Note: Since Dataset 2 contains ONH-only OCT scans without the macula, AI-CNet3D_{H-NA} and AI-CNet3D_{NA} cannot be evaluated.) The Base 3D CNN, SEResNeXt50, EPA, TimeSformer, and the medical-specific models M3T and Med3D demonstrate high specificity on Dataset 2, likely due to learning more conservative decision boundaries from overfitting to simpler non-glaucomatous cases, although often at the cost of lower sensitivity (the Base 3D CNN’s high specificity here can also be attributed to the fact that this model was optimized for Dataset 2 volumes). In contrast, our cross-attention approach achieves comparable specificity (and surpasses the other 4 models when fine-tuned, as shown in Table 3) while also improving sensitivity. This indicates a superior ability to reduce both false positives and

false negatives. Striking this balance is critical in medical applications, where accurately detecting disease (sensitivity) must be balanced with minimizing false alarms (specificity) to ensure reliable diagnostics.

5.2 Multi-task Fine-tuning to Enforce Visualization-Based Consistency

Integrating the cross-attention mechanism into the CNN framework yields a performance improvement. In addition, fine-tuning the model with a combination of unsupervised visualization consistency and BCE losses further enhances its effectiveness. As shown in Table 3, we compare our models against a hybrid EPA CNN model both before and after fine-tuning. Across both datasets, fine-tuning consistently improves generalization for all models, with our AI-CNet3D_H approach achieving the highest overall performance. These results show even greater performance compared to those in Table 2 (F1-score of AI-CNet3D_H with fine-tuning is significantly better than that of EPA with fine-tuning, $p = 0.03$, on Dataset 1). Furthermore, we can see multi-task fine-tuning models applied to Dataset 2 (Zeiss) exhibit improved specificity compared to their non-fine-tuned counterparts, while maintaining sensitivity within each respective model pair. Due to the lower resolution of Dataset 2, the model may struggle to capture anatomical features effectively before fine-tuning. However, after fine-tuning and enforcing layer-wise consistency, it is better able to capture these features, leading to more significant improvements compared to Dataset 1. This shows that the combination of our 3D cross-attention mechanism with consistency enforced between CARE attention representations and Grad-CAMs from convolutional layers within our model helps in precise feature extraction and increased regularization.

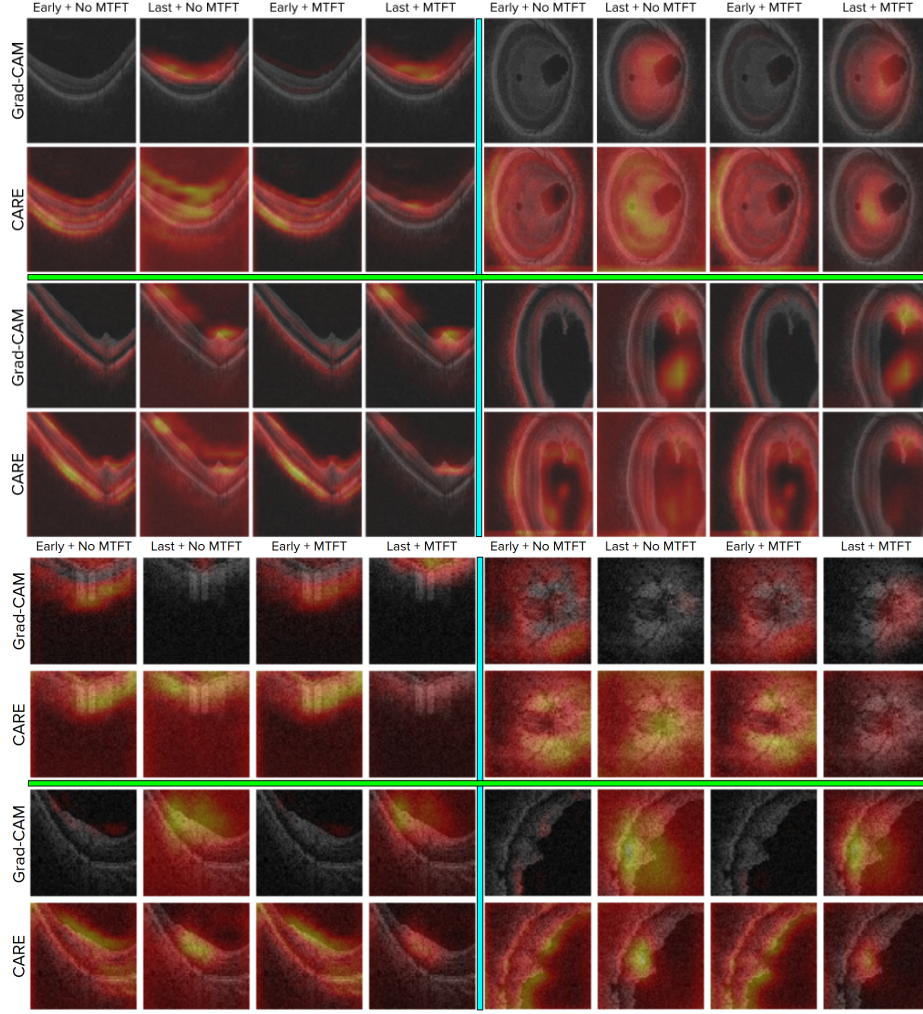


Figure 4: Comparison of CARE and Grad-CAM visualizations from our top-performing **AI-CNet3D** model before and after multi-task fine-tuning (MTFT) on **Dataset 1 (Topcon)** for the first four rows and **Dataset 2 (Zeiss)** for the last four rows. The heatmaps use a scale of intensities to represent importance, with **yellow** indicating the highest relevance, **red** indicating moderate importance, and the absence of activation representing the least relevance. The **No MTFT** model was trained with standard BCE loss for 250 epochs, while the **MTFT** model was fine-tuned for an additional 250 epochs using a combination of unsupervised MSE loss and supervised BCE loss. **Early** indicates attention from the second convolutional and first attention layer, whereas **Last** corresponds to the final convolutional and attention layers. **True positive** examples are shown above the **green** line, and **true negatives** are displayed below. **Axial slices** appear to the left of the **cyan** line, while **coronal slices** are to the right. After MTFT, attention maps from the **Last** layers exhibit greater consistency, highlighting more stable and interpretable representations.

5.3 3D Visualization

When using CARE and 3D Grad-CAM with our model, we found that extracting visualizations from earlier layers yielded the most aesthetically-pleasing results, as these layers retained the volume’s features at a higher resolution most effectively. In Fig. 4, we compare the CARE from the first cross-attention block with the 3D Grad-CAM from the second convolutional block, using the same model (AI-CNet3D_H) since it was trained on both datasets. Solely for visualization purposes, we apply one round of dilation to the Grad-CAM outputs to enhance under-highlighted regions and one round of erosion to the CARE outputs to refine over-highlighted areas. We can see similar regions highlighted by both methods, but CARE extracts informa-

tion from deeper layers within the retina which we discuss in Section 6.4. We can also see that after fine-tuning, the CARE and Grad-CAM visualizations are more consistent in the last layer since we only utilize deeper layers for alignment training. This means that earlier layers can still capture information using their unique advantages: convolutions for local feature extraction and attention for capturing long-range dependencies. Then, in the deeper parts of the network, these layers synergistically share information learned to make a prediction.

6. Discussion

Our targeted approach offers several key advantages over whole-volume spatial self-attention: (1) it directly models

Table 3: Performance evaluation of the hybrid EPA-CNN model and our proposed hybrid cross-attention CNN models on Dataset 1 (Topcon) and Dataset 2 (Zeiss), with and without multi-task fine-tuning. The unsupervised weighting in Equation 12 is set to $\lambda = 0.75$ for Dataset 1 and $\lambda = 0.5$ for Dataset 2. We report p-values for Mann Whitney U tests between models trained without and with fine-tuning in parentheses after the standard deviations.

	Model Type	Avg. Test Acc. \pm Std.	Avg. Test Spec. \pm Std.	Avg. Test Sens. \pm Std.	Avg. Test AUROC \pm Std.	Avg. Test F1 Score \pm Std.
Topcon	EPA (Shaker et al., 2024)	0.7872 \pm 0.0404 (0.461)	0.8235 \pm 0.0372 (0.600)	0.7539 \pm 0.0593 (0.834)	0.7887 \pm 0.0385 (0.548)	0.7831 \pm 0.0429 (0.841)
	EPA (Shaker et al., 2024) w/ fine-tuning	0.7963 \pm 0.0242	0.8187 \pm 0.0869	0.7738 \pm 0.0703	0.7962 \pm 0.0239	0.7948 \pm 0.0256
	AI-CNet3D _{NA} (Ours)	0.8037 \pm 0.0356 (0.462)	0.7992 \pm 0.0669 (0.548)	0.8076 \pm 0.0253 (0.674)	0.8034 \pm 0.0373 (0.548)	0.8090 \pm 0.0249 (0.421)
	AI-CNet3D _{NA} (Ours) w/ fine-tuning	0.8128 \pm 0.0263	0.8205 \pm 0.0535	0.8029 \pm 0.0191	0.8117 \pm 0.0274	0.8152 \pm 0.0144
	AI-CNet3D _{H-NA} (Ours)	0.7945 \pm 0.0356 (0.246)	0.7936 \pm 0.0820 (0.310)	0.7921 \pm 0.0457 (0.917)	0.7928 \pm 0.0359 (0.310)	0.7987 \pm 0.0233 (0.310)
	AI-CNet3D _{H-NA} (Ours) w/ fine-tuning	0.8220 \pm 0.0250	0.8659 \pm 0.0765	0.7772 \pm 0.0646	0.8216 \pm 0.0239	0.8170 \pm 0.0204
	AI-CNet3D _H (Ours)	0.8183 \pm 0.0340 (0.462)	0.8290 \pm 0.0604 (0.917)	0.8063 \pm 0.0203 (0.293)	0.8176 \pm 0.0339 (0.548)	0.8204 \pm 0.0288 (0.421)
	AI-CNet3D _H (Ours) w/ fine-tuning	0.8312 \pm 0.0137	0.8468 \pm 0.0482	0.8181 \pm 0.0272	0.8325 \pm 0.0135	0.8323 \pm 0.0092
Zeiss	EPA (Shaker et al., 2024)	0.8162 \pm 0.0429 (1.000)	0.8379 \pm 0.0753 (0.674)	0.8038 \pm 0.0810 (0.691)	0.8209 \pm 0.0438 (0.841)	0.8122 \pm 0.0398 (1.000)
	EPA (Shaker et al., 2024) w/ fine-tuning	0.8252 \pm 0.0406	0.8666 \pm 0.0514	0.7922 \pm 0.0977	0.8294 \pm 0.0426	0.8163 \pm 0.0434
	AI-CNet3D _H (Ours)	0.8315 \pm 0.0105 (0.095)	0.8246 \pm 0.0336 (0.047)	0.8425 \pm 0.0388 (1.000)	0.8336 \pm 0.0130 (0.095)	0.8311 \pm 0.0134 (0.222)
	AI-CNet3D _H (Ours) w/ fine-tuning	0.8573 \pm 0.0252	0.8771 \pm 0.0237	0.8392 \pm 0.0503	0.8582 \pm 0.0249	0.8534 \pm 0.0228

known pathophysiological relationships in glaucoma, where asymmetric damage between superior and inferior regions is clinically significant, (2) it dramatically reduces computational complexity by eliminating the need for costly volume projections required in spatial attention (achieving 241k-291k parameters versus 63-88M in standard transformers), and (3) it provides inherent interpretability aligned with clinical understanding. Unlike sequence-based approaches such as State Space Models (Gu et al., 2021) or Mamba (Gu and Dao, 2023) that process data linearly, our method explicitly captures the spatial relationships critical for understanding retinal pathology. Our CARE visualization reveals that the model leverages deeper retinal structures beyond conventional RNFL analysis, demonstrating how anatomically-constrained attention can uncover clinically relevant features that generic spatial attention might miss.

6.1 Anatomically Informed Cross-attention

The introduction of our 3D cross-attention mechanism provides new insights into capturing anatomically-informed structural variations in volumetric OCT data to enhance glaucoma classification performance. The AI-CNet3D_{H-NA}, AI-CNet3D_H, and AI-CNet3D_{NA} rows in Table 2 demonstrate that applying cross-attention yielded consistently better performance compared to models without cross-attention. The improvement seen in our AI-CNet3D models is grounded in current understanding of glaucoma and its pathophysiology. Given that damage to the ganglion cells (GCL) and their axons (RNFL) should be geographically correlated based on anatomy, cross-attention between the macula (where ganglion cell bodies reside) and the optic nerve (where their axons pass) is effective in isolating this correlation over models that do not incorporate cross-attention. Even more pronounced, the anatomical separation of supe-

rior and inferior hemiretina allows the AI-CNet3D_H model to detect early asymmetries through internal control and perform better than its other anatomically-informed variants.

Our results also highlight the versatility and robustness of AI-CNet3D, confirming its generalizability across both ONH-only and widefield OCT imaging formats. Table 2 illustrates the performance of our AI-CNet3D_H on both widefield (Dataset 1) and ONH-only (Dataset 2) volumes. The 3D CNN proposed by Maetschke et al. (2019) was optimized for Dataset 2 volumes. However, its performance degrades significantly when evaluated on Dataset 1, indicating limited generalizability across datasets. Our improvement in performance indicates that the advantage of our approach is achieved by embedding our understanding of disease pathology through modeling and thus is not limited to particular scans and orientations. Our approach removes the spatial constraints that may limit models' ability to analyze 3D images and allows for computation of biologically-meaningful cross-correlations between regions that are tailored toward a specific clinical question. This is a key advantage, as it offers a customization approach that can be applied to multiple imaging modalities and disease settings.

6.2 Multi-task Fine-tuning to Enforce Visualization-Based Consistency

Qualitative results in Fig. 4 demonstrate strong alignment and consistency between the last attention and convolutional layers, suggesting effective information sharing between them. This interaction enables the model to leverage the strengths of both mechanisms: convolutions excel at capturing fine-grained local structures due to their limited receptive fields, while attention mechanisms provide a com-

plementary ability to model long-range dependencies. By aligning their outputs, the convolutional layer gains access to broader contextual information, while the attention mechanism benefits from enhanced local feature sensitivity. This synergy improves feature representation, allowing the model to integrate both fine-detail spatial structures and global contextual cues, ultimately enhancing interpretability and performance.

Fine-tuning the model with a combination of unsupervised visualization consistency and BCE losses plays a crucial role in enhancing its generalization ability. By enforcing consistency in model visualizations across attention and convolutional layers, the model learns more stable and robust feature representations, reducing susceptibility to spurious correlations in the data. The inclusion of BCE loss during fine-tuning ensures that the model retains its learned features for classification while preventing degeneration. We can also notice that regardless of the attention mechanism used in Table 3, multi-task fine-tuning improves individual results. This shows how our fine-tuning strategy can be used for general improvement after a model has been trained. This improvement is particularly significant for medical imaging tasks: by leveraging consistency-based multi-task fine-tuning, we offer a dependable AI-assisted analysis framework that is agnostic to attention model backbone, enabling consistent performance improvement and interpretability enhancement, critical for clinical adoption.

6.3 Computational Complexity

Model efficiency is a crucial consideration in developing 3D models, particularly for applications in low-resource settings where hardware limitations and inference speed are major constraints, as noted in prior work such as Shaker et al. (2024). In designing our 3D cross-attention mechanism, we prioritized a lightweight architecture that balances robust performance with computational efficiency, allowing for fast deployment and real-time processing on resource-limited devices.

Our 3D cross-attention mechanism achieves these efficiency gains primarily by eliminating the need for volume projections for spatial attention computation, which drastically reduces parameter count without compromising model accuracy. As shown in Figure 6, when evaluated on $128 \times 192 \times 112$ pixel volumes with a single-batch input, traditional models exhibit significant variation in computational demands. The Base 3D CNN operates with approximately 222k parameters and 152.941 GFLOPS, while SEResNeXt50, EPA, and M3T require 28.36M, 25.01M, and 27.05M parameters, with computational costs of 72.064, 108.801, and 29.90 GFLOPS, respectively. More resource-intensive models like TimeSformer, ViT, and Med3D demand 63.15M, 88.18M, 63.30M parameters, with signif-

icantly higher computational costs of 1357.11, 135.34, 745.47 GFLOPS, respectively. In contrast, our proposed *AI-CNet3D_H*, *AI-CNet3D_{H-NA}*, and *AI-CNet3D_{NA}* achieve competitive performance with just 241k to 291k parameters while maintaining computational efficiency at 104.837 to 108.008 GFLOPS, underscoring our model's optimized architecture and computational efficiency.

Our method's ability to outperform models like TimeSformer in terms of computational efficiency position it as a highly-scalable approach, especially as medical imaging datasets grow larger and more complex. The reduced resource demand also makes AI-CNet3D ideal for integration into portable medical devices and remote healthcare, offering real-time, reliable diagnostics in resource-constrained settings.

6.4 3D Visualization

CARE visualizations suggest that leveraging 3D OCT volumes for glaucoma diagnosis enables deep learning models incorporating attention to learn information from deeper features in the retina beyond the typical RNFL or GCC relied on clinically in 2D OCT reports. While we have leveraged this advantage in the X-Z plane to correlate nerve to macula or superior to inferior as discussed above, such advantage may equally apply in the Y direction.

Past work has shown that Grad-CAM successfully highlights the RNFL layer (Thakoor et al., 2020), consistent with our current understanding of glaucoma pathophysiology. Both Grad-CAM and CARE demonstrated the importance of RNFL and GCC layers, as showcased by involvement of these superficial layers in the model's decision making. However, only CARE demonstrated the use of features from deeper retinal layers for classification of glaucoma, which deviates from the traditional understanding of glaucoma. Previous literature has suggested possible anatomical changes in the photoreceptor layers in patients with glaucoma (Fan et al., 2011; Trolli et al., 2024), although this has not been validated or used in clinical settings. Our results are consistent with these studies, indicating that deeper retinal structures, such as the photoreceptor layer, may have subtle changes that can be leveraged by AI models to detect glaucoma.

The power of multi-task fine-tuning (MTFT) in allowing convolutional layers (as visualized by Grad-CAM) and attention layers (as visualized by CARE) to share key information is also well demonstrated by our visualization. The consistency between the visualizations of the last convolutional and attention layers illustrates the successful integration of the shared information, as discussed in 6.2. The interpretability of this visualization is further supported by increased signals originating from the superior and inferior rims of the optic nerve head, following an arcuate pattern

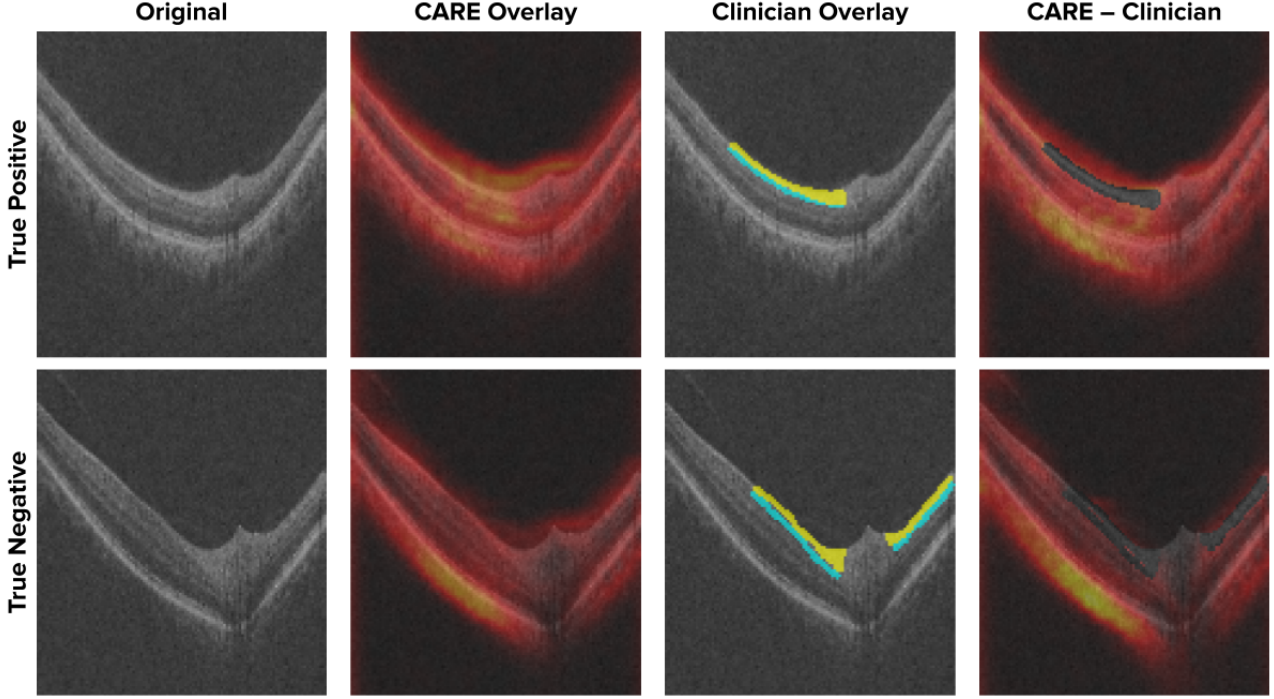


Figure 5: A comparison of what our CARE highlights versus the important regions identified by the clinician. In the final column, we subtract the clinical regions of interest (ROIs) from the CARE heatmap to highlight how our method captures additional features within the volumes beyond the clinical ROIs.

into the macula (especially evident in Topcon true-negative coronal slices in Fig. 4), consistent with our clinical understanding of glaucoma.

Metric	Avg. \pm Std.	Value
Mask Coverage	0.9337 ± 0.1004	
CARE Coverage	0.0293 ± 0.0125	
Enrichment	3.0915 ± 0.6144	

Table 4: Quantitative evaluation of CARE attention alignment with clinician-annotated anatomical regions. Mask coverage measures the fraction of clinician-defined regions that receive high CARE attention, CARE coverage indicates the fraction of high-attention pixels that fall within anatomical boundaries, and enrichment quantifies how much more concentrated CARE is within versus outside the expert-annotated RNFL/GCL regions of interest.

To validate our method’s ability to capture clinically relevant anatomical regions, we obtained expert segmentation annotations from a clinician at Columbia University Irving Medical Center. The clinician manually segmented the middle slice of each test volume from the Topcon dataset, identifying RNFL and GCL Regions of Interest (ROIs). These

annotations were converted into binary masks and compared against the corresponding middle slice of our CARE attention maps generated using the first attention layer of *AI-CNet3D_H* after multi-task fine-tuning. We converted each CARE slice into a binary segmentation using the 75th percentile as a threshold (τ) to identify the most intense regions. We then computed mask coverage as $\frac{|M \cap \text{CARE}_\tau|}{|M|}$, CARE coverage as $\frac{|M \cap \text{CARE}_\tau|}{|\text{CARE}_\tau|}$, and enrichment as $\frac{\text{CARE}_M}{\text{CARE}_{V \setminus M}}$, where M represents the clinician-annotated mask, CARE_τ represents the binarized CARE map above threshold τ , $M \cap \text{CARE}_\tau$ represents their intersection, V represents the full slice, and CARE_M and $\text{CARE}_{V \setminus M}$ denote the mean CARE intensity inside and outside the annotation mask area, respectively. Values of 1.0 for enrichment indicate the same CARE density inside and outside the mask, while values greater than 1.0 indicate that CARE is more enriched inside the annotation mask area.

Table 4 reveals that our model demonstrates exceptional annotation mask coverage in targeting anatomical structures, with 93.37% of anatomical pixels receiving significant attention. As seen with CARE coverage, only 2.93% of the highest attention pixels fall within the clinician-defined anatomical boundaries. This reflects our model’s broader analytical scope, with potential to enable novel biomarker

discovery by capturing extensive retinal features beyond conventionally-defined clinical RoIs. Additionally, the CARE coverage is likely slightly higher in reality, as the physician annotation of RoIs were limited by resolution, and the true RoIs (RNFL and GCL layers) are anatomically known to continue beyond the annotated regions, nasally and temporally. At the same time, the fact that the enrichment value for all test cases is greater than 1.0, indicates that 100% of the time, CARE is more concentrated within the clinically defined regions of interest. This comprehensive attention distribution is visualized in Figure 5, which demonstrates that our model encompasses the clinician’s primary areas of interest while extending analysis to capture additional diagnostically-relevant features throughout the retinal volume.

There are some caveats to using CARE and Grad-CAM for consistency and 3D visualization. Notably, as shown in Fig. 4, consistency after multi-task fine-tuning (MTFT) is significantly stronger in the last layers for true negative cases. This may explain the larger increases in specificity observed in Table 3, as the model becomes more adept at learning from negative cases. Additionally, a qualitative comparison of Zeiss and Topcon examples reveals that Zeiss visualizations are less precise. This is likely due to the lower resolution of Zeiss volumes ($64 \times 128 \times 64$), which leads to further degradation in heatmap quality when downsampled within our model.

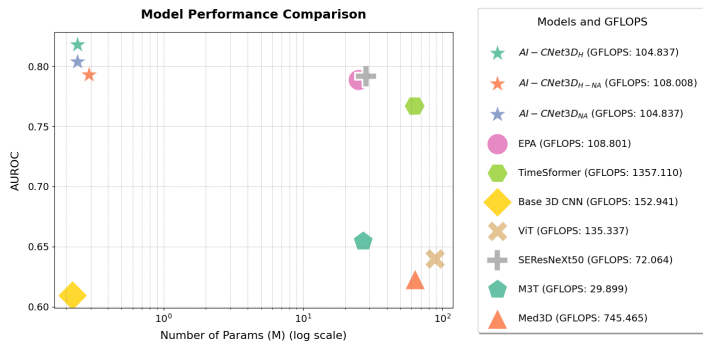


Figure 6: Efficiency analysis of our models and baseline models with parameters displayed on a log scale for enhanced visualization. Our models achieve the highest AUROC, matching EPA’s GFLOP performance while significantly reducing parameter count, highlighting their efficiency and accuracy.

7. Conclusions and Future Directions

In this paper, we introduced a novel 3D cross-attention mechanism for widefield OCT reports, demonstrating its potential to enhance glaucoma classification by incorporating anatomical priors and inter-region correlations within 3D volumes. Our approach, implemented as the AI-CNet3D

model, outperforms traditional 3D CNN-based models, as well as other state-of-the-art spatial and channel attention methods, in both classification performance and computational efficiency. The use of superior-inferior and macula-nerve cross-attention mechanisms allows our model to leverage the inherent anatomical relationships in OCT scans, which leads to improved sensitivity, specificity, and overall performance, which may be beneficial in the detection of subtle or early-stage glaucoma. The computational efficiency of our model, achieved by eliminating the need for volume projections and reducing parameter count, makes it particularly suitable for resource-constrained environments, such as portable imaging devices and point-of-care diagnostics.

Our newly-introduced CARE technique enables visualization of 3D attention mechanisms, which are inherently challenging to portray visually. Compared to Grad-CAM, CARE visualizations highlighted the RNFL and GCC as well as deeper structures, such as photoreceptors, as critical for the model’s decision-making. We enhanced model performance by training with a joint unsupervised visualization consistency loss and regular BCE loss as a fine-tuning step. This process allows the model to generalize better while also increasing its interpretability with alignment between layers. Furthermore, CARE specifically elucidated the contribution from deeper retinal structures for glaucoma classification. By leveraging our technique, both CARE and 3D Grad-CAM can be applied to hybrid CNN-attention models, empowering clinicians to interpret contributions of every layer in the model, including attention layers, rather than limiting their insights to convolutional layers alone.

Our work also contributes to the growing body of research on 3D deep learning models for medical imaging, highlighting the importance of attention-based and consistency-based multi-task fine-tuning approaches to improve interpretability and model performance in clinically relevant tasks. To ensure class balance, we used equal amounts of glaucomatous and non-glaucomatous data in our training pipeline, both with and without fine-tuning. As a direction for future work, the full set of available samples from both datasets, or additional data from external domains, could be incorporated into a pre-training task to improve model robustness and generalization. Additionally, we aim to further investigate the clinical impact of our model by leveraging CARE outputs to create pseudo-segmentation masks that can be corrected by minimal annotation of a clinician. We can then use our cross-attention network to identify and segment novel biomarkers for retinal diseases, thereby advancing AI-driven diagnostics in ophthalmology.

Acknowledgments

The authors are grateful to George A. Cioffi, Jeffrey M. Liebmann, Aakriti G. Shukla, and Ives A. Valenzuela for their guidance and insights as glaucoma specialists. Many thanks go to Mary Durbin and Reena Chopra (Topcon Healthcare, Inc.) for collaborative data sharing as well as Emmanouil Tsamis and Donald C. Hood for providing ground truth labels for the volumetric OCT data. We also acknowledge our funding sources: Columbia University's Data Science Institute Seed Fund and an Unrestricted Grant from Research to Prevent Blindness, New York, NY, USA.

Ethical Standards

Our Topcon Dataset 1 was from a retrospective study conducted in accordance with the Declaration of Helsinki and approved on 14 February 2023 by Advarra Institutional Review Board (MOD01564217).

Conflicts of Interest

Author K.A.T. has received research funding from Topcon Healthcare for a study unrelated to the topic of this paper. All other authors declare no further conflicts of interest.

Data availability

Given the sensitive nature of the human subject data used in our study, it is essential to maintain confidentiality and adhere to ethical guidelines. While our data collection received approval from the Institutional Review Board (IRB), any further access to Dataset 1 must be carefully regulated. To ensure appropriate use, we would need to assess data requests to verify that the intended purpose aligns with the submitted request. For the journal's internal evaluation, we have already provided a minimal de-identified dataset through article figures and can furnish additional datasets if further review is necessary. Dataset 2 is available publicly as published by Maetschke et al. (2019) at <https://zenodo.org/records/1481223>.

References

- Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. *arXiv preprint arXiv:2005.00928*, 2020.
- Oren Avram, Berkin Durmus, Nadav Rakocz, Giulia Corradetti, Ulzee An, Muneeswar G Nitalla, Ákos Rudas, Yu Wakatsuki, Kazutaka Hirabayashi, Swetha Velaga, et al. Slivit: a general ai framework for clinical-feature diagnosis from limited 3d biomedical-imaging data. *Research Square*, pages rs–3, 2023.
- Daniele MS Barros, Julio CC Moura, Cefas R Freire, Alexandre C Taleb, Ricardo AM Valentim, and Philippi SG Morais. Machine learning applied to retinal image processing for glaucoma detection: review and perspective. *Biomedical engineering online*, 19:1–21, 2020.
- Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, volume 2, page 4, 2021.
- Rüdiger Bock, Jörg Meier, László G Nyúl, Joachim Hornegger, and Georg Michelson. Glaucoma risk index: automated glaucoma detection from color fundus images. *Medical image analysis*, 14(3):471–481, 2010.
- Igor I Bussel, Gadi Wollstein, and Joel S Schuman. Oct for glaucoma diagnosis, screening and detection of glaucoma progression. *British Journal of Ophthalmology*, 98(Suppl 2):ii15–ii19, 2014.
- Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European conference on computer vision*, pages 205–218. Springer, 2022.
- Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 782–791, 2021.
- Chun-Fu Richard Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 357–366, 2021.
- Sihong Chen, Kai Ma, and Yefeng Zheng. Med3d: Transfer learning for 3d medical image analysis. *arXiv preprint arXiv:1904.00625*, 2019.
- Teresa C Chen, Ambika Hogue, Anna K Junk, Kouros Nouri-Mahdavi, Sunita Radhakrishnan, Hana L Takusagawa, and Philip P Chen. Spectral-domain oct: helping the clinician diagnose glaucoma: a report by the american academy of ophthalmology. *Ophthalmology*, 125(11):1817–1827, 2018.
- Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

- Xiangyu Chen, Yanwu Xu, Damon Wing Kee Wong, Tien Yin Wong, and Jiang Liu. Glaucoma detection based on deep convolutional neural network. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 715–718, 2015. .
- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929, 2020. URL <https://arxiv.org/abs/2010.11929>.
- Ning Fan, Nina Huang, Dennis Shun Chiu Lam, and Christopher Kai-shun Leung. Measurement of photoreceptor layer in glaucoma: a spectral-domain optical coherence tomography study. *Journal of ophthalmology*, 2011(1): 264803, 2011.
- Alexi Geevarghese, Gadi Wollstein, Hiroshi Ishikawa, and Joel S Schuman. Optical coherence tomography and glaucoma. *Annual review of vision science*, 7(1):693–726, 2021.
- Yasmeen George, Bhavna J. Antony, Hiroshi Ishikawa, Gadi Wollstein, Joel S. Schuman, and Rahil Garnavi. Attention-guided 3d-cnn framework for glaucoma detection and structural-functional association using volumetric images. *IEEE Journal of Biomedical and Health Informatics*, 24(12):3421–3430, 2020. .
- Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Dohersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.
- Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.
- Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*, 2021.
- Hao Guo, Kang Zheng, Xiaochuan Fan, Hongkai Yu, and Song Wang. Visual attention consistency under image transforms for multi-label image classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 729–739, 2019.
- Donald C Hood, Sol La Bruna, Emmanouil Tsamis, Kaveri A Thakoor, Anvit Rai, Ari Leshno, Carlos GV de Moraes, George A Cioffi, and Jeffrey M Liebmann. Detecting glaucoma with only oct: Implications for the clinic, research, screening, and ai development. *Progress in Retinal and Eye Research*, 90:101052, 2022.
- Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018. .
- Mobarakol Islam, VS Vibashan, V Jeya Maria Jose, Navodini Wijethilake, Uppal Utkarsh, and Hongliang Ren. Brain tumor segmentation and survival prediction using 3d attention unet. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 5th International Workshop, BrainLes 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 17, 2019, Revised Selected Papers, Part I 5*, pages 262–272. Springer, 2020.
- Jinseong Jang and Dosik Hwang. M3t: three-dimensional medical image classifier using multi-plane and multi-slice transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20718–20729, 2022.
- Liu Li, Mai Xu, Xiaofei Wang, Lai Jiang, and Hanruo Liu. Attention based glaucoma detection: A large-scale database and cnn model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10571–10580, 2019.
- Yang Li, Shichao Kan, and Zhihai He. Unsupervised deep metric learning with transformed attention consistency and contrastive clustering loss. In *European Conference on Computer Vision*, pages 141–157. Springer, 2020.
- Hezheng Lin, Xing Cheng, Xiangyu Wu, and Dong Shen. Cat: Cross attention in vision transformer. In *2022 IEEE international conference on multimedia and expo (ICME)*, pages 1–6. IEEE, 2022.
- Jiarun Liu, Hao Yang, Hong-Yu Zhou, Lequan Yu, Yong Liang, Yizhou Yu, Shaoting Zhang, Hairong Zheng, and Shanshan Wang. Swin-umamba†: Adapting mamba-based vision foundation models for medical image segmentation. *IEEE Transactions on Medical Imaging*, 2024a.
- Zixuan Liu, Hanwen Xu, Addie Woicik, Linda G. Shapiro, Marian Blazes, Yue Wu, Verena Steffen, Catherine Cukras, Cecilia S. Lee, Miao Zhang, Aaron Y. Lee, and Sheng Wang. Octcube-m: A 3d multimodal optical coherence tomography foundation model for retinal and systemic diseases with cross-cohort and cross-device validation, 2024b. URL <https://arxiv.org/abs/2408.11227>.

- Yan Luo, Min Shi, Yu Tian, Tobias Elze, and Mengyu Wang. Harvard glaucoma detection and progression: A multimodal multitask dataset and generalization-reinforced semi-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20471–20482, 2023.
- Stefan Maetschke, Bhavna Antony, Hiroshi Ishikawa, Gadi Wollstein, Joel Schuman, and Rahil Garnavi. A feature agnostic approach for glaucoma detection in oct volumes. *PLOS ONE*, 14(7):e0219126, July 2019. ISSN 1932-6203. . URL <http://dx.doi.org/10.1371/journal.pone.0219126>.
- Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.
- Parmita Mehta, Christine A Petersen, Joanne C Wen, Michael R Banitt, Philip P Chen, Karine D Bojikian, Catherine Egan, Su-In Lee, Magdalena Balazinska, Aaron Y Lee, et al. Automated detection of glaucoma with interpretable machine learning using clinical data and multimodal retinal images. *American Journal of Ophthalmology*, 231:154–169, 2021.
- Ali Mirzazadeh, Florian Dubost, Maxwell Pike, Krish Maniar, Max Zuo, Christopher Lee-Messer, and Daniel Rubin. Atcon: Attention consistency for vision models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1880–1889, 2023.
- Yan Pang, Jiaming Liang, Teng Huang, Hao Chen, Yunhao Li, Dan Li, Lin Huang, and Qiong Wang. Slim unetr: scale hybrid transformers to efficient 3d medical image segmentation under limited computational resources. *IEEE Transactions on Medical Imaging*, 43(3):994–1005, 2023.
- Harry A Quigley and Aimee T Broman. The number of people with glaucoma worldwide in 2010 and 2020. *British journal of ophthalmology*, 90(3):262–267, 2006.
- Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: visual explanations from deep networks via gradient-based localization. *International journal of computer vision*, 128:336–359, 2020.
- Torgyn Shaikhina and Natalia A Khovanova. Handling limited datasets with neural networks in medical applications: A small-data approach. *Artificial intelligence in medicine*, 75:51–63, 2017.
- Abdelrahman Shaker, Muhammad Maaz, Hanoona Rasheed, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Swiftformer: Efficient additive attention for transformer-based real-time mobile vision applications. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 17425–17436, 2023.
- Abdelrahman M. Shaker, Muhammad Maaz, Hanoona Rasheed, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Unetr++: Delving into efficient and accurate 3d medical image segmentation. *IEEE Transactions on Medical Imaging*, 2024. .
- Jaimie D Steinmetz, Rupert RA Bourne, Paul Svitil Briant, Seth R Flaxman, Hugh RB Taylor, Jost B Jonas, Amir Aberhe Abdoli, Woldu Aberhe Abrha, Ahmed Abualhasan, Eman Girum Abu-Gharbieh, et al. Causes of blindness and vision impairment in 2020 and trends over 30 years, and prevalence of avoidable blindness in relation to vision 2020: the right to sight: an analysis for the global burden of disease study. *The Lancet Global Health*, 9(2):e144–e160, 2021.
- Yucheng Tang, Dong Yang, Wenqi Li, Holger R Roth, Bennett Landman, Daguang Xu, Vishwesh Nath, and Ali Hatamizadeh. Self-supervised pre-training of swin transformers for 3d medical image analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20730–20740, 2022.
- Kaveri A Thakoor, Sharath C Koorathota, Donald C Hood, and Paul Sajda. Robust and interpretable convolutional neural networks to detect glaucoma in optical coherence tomography images. *IEEE Transactions on Biomedical Engineering*, 68(8):2456–2466, 2020.
- Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.
- Eleonora Trolli, Matilde Roda, Nicola Valsecchi, Davide Cacciatore, Elena Nardi, Valentina Della Pasqua, Andrea Mercanti, and Luigi Fontana. A parafoveal retinal cones analysis using adaptive-optics retinal camera in patients with primary open angle glaucoma. *Eye*, pages 1–7, 2024.
- A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.
- Lezi Wang, Ziyang Wu, Srikrishna Karanam, Kuan-Chuan Peng, Rajat Vikram Singh, Bo Liu, and Dimitris N Metaxas. Sharpen focus: Learning with attention separability and consistency. In *Proceedings of the IEEE/CVF*

- International Conference on Computer Vision*, pages 512–521, 2019a.
- Xudong Wang, Shizhong Han, Yunqiang Chen, Dashan Gao, and Nuno Vasconcelos. Volumetric attention for 3d medical image segmentation and detection. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part VI 22*, pages 175–184. Springer, 2019b.
- Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. *arXiv preprint arXiv:2210.10163*, 2022.
- Yutong Xie, Bing Yang, Qingbiao Guan, Jianpeng Zhang, Qi Wu, and Yong Xia. Attention mechanisms in medical image segmentation: A survey. *arXiv preprint arXiv:2305.17937*, 2023.
- Haotian Xu, Xiaobo Jin, Qiufeng Wang, and Kaizhu Huang. Multi-scale attention consistency for multi-label image classification. In *Neural Information Processing: 27th International Conference, ICONIP 2020, Bangkok, Thailand, November 18–22, 2020, Proceedings, Part IV 27*, pages 815–823. Springer, 2020.
- Jieping Ye and Jun Liu. Sparse methods for biomedical data. *ACM Sigkdd Explorations Newsletter*, 14(1):4–15, 2012.
- Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. In Yoshua Bengio and Yann LeCun, editors, *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2–4, 2016, Conference Track Proceedings*, 2016. URL <http://arxiv.org/abs/1511.07122>.
- MD Zeiler. Visualizing and understanding convolutional networks. In *European conference on computer vision/arXiv*, volume 1311, 2014.
- Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.

Appendix A. Ablation Experiments

A.1 Removing spatial attention

Efficient Paired Attention (EPA), introduced by Shaker et al. (2024), was designed to efficiently compute spatial and channel self-attention for 3D feature volumes. However, its reliance on projections to condense spatial dimensions into smaller vectors introduces a significant bottleneck and increases the parameter count. In our cross-attention method, we address this limitation by computing attention exclusively along the channel dimension. As demonstrated in Table 5, incorporating cross-attention across both spatial and channel dimensions provides no tangible benefit beyond an increased parameter count.

A.2 Selecting λ for Multi-Task Fine-Tuning

In our multi-task fine-tuning framework, the overall loss function is defined as:

$$L_{\text{multi-task}} = (1 - \lambda)L_{\text{supervised}} + \lambda L_{\text{unsupervised}}. \quad (13)$$

The parameter λ controls the balance between the supervised loss $L_{\text{supervised}}$ and the unsupervised loss $L_{\text{unsupervised}}$. A higher λ increases the influence of unsupervised learning, while a lower λ prioritizes supervised learning. To determine the optimal λ value, we conducted ablation studies on two datasets: Topcon and Zeiss. The results, shown in Table 6, indicate that the impact of λ varies across datasets.

For the Topcon dataset, we observe that $\lambda = 0.75$ achieves the highest test accuracy (0.8104), specificity (0.8209), and AUROC (0.8110). The F1 score (0.8150) is also close to the highest value obtained. This suggests that emphasizing the unsupervised component at this weight contributes to improved model performance, likely by leveraging additional structure in the data to refine decision boundaries. For the Zeiss dataset, $\lambda = 0.50$ performs best, yielding the highest test accuracy (0.8681), sensitivity (0.8761), AUROC (0.8693), and F1 score (0.8627). Interestingly, while $\lambda = 0.90$ achieves high specificity (0.8926), it does not lead to better performance across the other metrics. This indicates that the optimal λ value depends on dataset characteristics, and a moderate balance between supervised and unsupervised learning is preferable to avoid overfitting (Mirzazadeh et al., 2023) to patterns that do not generalize well.

Across both datasets, we observe that relying solely on the unsupervised loss for fine-tuning ($\lambda = 1.0$) leads to model degeneration. This suggests that excessive dependence on unsupervised learning can degrade performance by overemphasizing features that do not align well with the primary classification task, ultimately reducing the model's effectiveness.

A.3 Selecting the Unsupervised Loss Function

Given that $\lambda = 0.75$ provided strong performance in the Topcon dataset, we further examined the effect of different loss functions for $L_{\text{unsupervised}}$ at this setting. Table 7 presents the results of this ablation study.

Among the four loss functions tested (MSE, SSIM, Pearson, and Gaussian Pearson), MSE consistently outperformed the others in terms of accuracy (0.8104), specificity (0.8209), AUROC (0.8110), and F1 score (0.8150). This suggests that minimizing mean squared error in the unsupervised loss effectively preserves useful feature representations while avoiding excessive penalization of small variations in data distributions.

SSIM and Pearson correlation loss yield suboptimal performance compared to MSE. SSIM, designed for structural similarity, performs worse across all metrics, indicating that it may not sufficiently preserve relevant feature distributions in the context of medical imaging classification. Pearson correlation, while improving over SSIM, does not reach the accuracy or F1-score achieved with MSE, possibly due to its focus on linear relationships rather than absolute differences.

Interestingly, Gaussian Pearson loss achieves the highest sensitivity (0.8332), indicating strong recall for positive cases. However, its low specificity (0.6724) and relatively lower AUROC (0.7528) suggest an overemphasis on certain patterns that may not generalize well. This highlights a trade-off when using loss functions that heavily favor recall over precision.

Based on these findings, we conclude that MSE is the most effective choice for $L_{\text{unsupervised}}$ in our multi-task fine-tuning framework, providing a balance between sensitivity and specificity while maximizing overall performance. Future work may explore hybrid loss formulations to further optimize model generalization across datasets.

A.4 Sampling Strategy

Since Dataset 1 (Topcon) contains 4932 non-glaucomatous and 272 glaucomatous samples, and Dataset 2 (Zeiss) contains 263 non-glaucomatous and 847 glaucomatous samples, we applied resampling strategies to mitigate class imbalance and prevent the model from overfitting to the majority class.

The first method we used was a class-weighted version of the binary cross-entropy (BCE) loss for supervised training. The weights were computed separately for each dataset based on the proportion of glaucomatous ($y_x = 1$) and non-glaucomatous ($y_x = 0$) samples. Let the predicted glaucoma probability be $P_x \in [0, 1]$, and the ground truth label be $y_x \in \{0, 1\}$. The class-weighted BCE loss is defined as:

Table 5: Ablation study for deciding how to compute cross-attention. Results are based on 5 trial averages.

	Attention Computation	Avg. Test Acc. \pm Std.	Avg. Test Spec. \pm Std.	Avg. Test Sens. \pm Std.	Avg. Test AUROC \pm Std.	Avg. Test F1 Score \pm Std.
Topcon	Only Channel Cross-attention	0.8183 \pm 0.0340	0.8290 \pm 0.0604	0.8063 \pm 0.0203	0.8176 \pm 0.0339	0.8204 \pm 0.0288
	Spatial + Channel Cross-attention	0.8018 \pm 0.0206	0.8176 \pm 0.0648	0.7848 \pm 0.0558	0.8012 \pm 0.0211	0.8018 \pm 0.0171
	EPA (Shaker et al., 2024)	0.7872 \pm 0.0404	0.8235 \pm 0.0372	0.7539 \pm 0.0593	0.7887 \pm 0.0385	0.7831 \pm 0.0429

Table 6: A comparison for selecting which λ to utilize for each dataset. These results are based on 3 trial averages.

	λ Value	Avg. Test Acc. \pm Std.	Avg. Test Spec. \pm Std.	Avg. Test Sens. \pm Std.	Avg. Test AUROC \pm Std.	Avg. Test F1 Score \pm Std.
Topcon	0.25	0.8073 \pm 0.0198	0.7730 \pm 0.0469	0.8360 \pm 0.0173	0.8045 \pm 0.0231	0.8197 \pm 0.0109
	0.50	0.7859 \pm 0.0114	0.7687 \pm 0.0177	0.8003 \pm 0.0173	0.7845 \pm 0.0111	0.7957 \pm 0.0182
	0.75	0.8104 \pm 0.0086	0.8209 \pm 0.0606	0.8011 \pm 0.0412	0.8110 \pm 0.0104	0.8150 \pm 0.0097
	0.90	0.7951 \pm 0.0043	0.7429 \pm 0.0646	0.8420 \pm 0.0636	0.7924 \pm 0.0056	0.8103 \pm 0.0136
	1.0	0.7951 \pm 0.0043	0.7429 \pm 0.0646	0.8420 \pm 0.0636	0.7924 \pm 0.0056	0.8103 \pm 0.0136
Zeiss	0.25	0.8617 \pm 0.0157	0.8987 \pm 0.0588	0.8240 \pm 0.0567	0.8613 \pm 0.0161	0.8496 \pm 0.0140
	0.50	0.8681 \pm 0.0115	0.8625 \pm 0.0207	0.8761 \pm 0.0461	0.8693 \pm 0.0129	0.8627 \pm 0.0141
	0.75	0.8519 \pm 0.0183	0.8634 \pm 0.0647	0.8426 \pm 0.0377	0.8530 \pm 0.0153	0.8439 \pm 0.0160
	0.90	0.8680 \pm 0.0070	0.8926 \pm 0.0276	0.8428 \pm 0.0363	0.8677 \pm 0.0080	0.8582 \pm 0.0073
	1.0	0.7332 \pm 0.0340	0.8801 \pm 0.0749	0.5700 \pm 0.0291	0.7250 \pm 0.0295	0.6707 \pm 0.0188

Table 7: Ablation study for selecting the loss function with $\lambda = 0.75$. Results are based on 3 trial averages.

	Loss Function	Avg. Test Acc. \pm Std.	Avg. Test Spec. \pm Std.	Avg. Test Sens. \pm Std.	Avg. Test AUROC \pm Std.	Avg. Test F1 Score \pm Std.
Topcon	MSE	0.8104 \pm 0.0086	0.8209 \pm 0.0606	0.8011 \pm 0.0412	0.8110 \pm 0.0104	0.8150 \pm 0.0097
	SSIM	0.7737 \pm 0.0189	0.7826 \pm 0.0134	0.7676 \pm 0.0344	0.7751 \pm 0.0172	0.7798 \pm 0.0169
	Pearson	0.7920 \pm 0.0114	0.8014 \pm 0.0198	0.7842 \pm 0.0390	0.7928 \pm 0.0099	0.7971 \pm 0.0151
	Gaussian Pearson	0.7523 \pm 0.0899	0.6724 \pm 0.1843	0.8332 \pm 0.0559	0.7528 \pm 0.0889	0.7824 \pm 0.0658

Table 8: Ablation study for deciding on sampling strategy. Results are based on 5 trial averages.

	Data Sampling Method	Avg. Test Acc. \pm Std.	Avg. Test Spec. \pm Std.	Avg. Test Sens. \pm Std.	Avg. Test AUROC \pm Std.	Avg. Test F1 Score \pm Std.
Topcon	No Sampling	0.9500 \pm 0.0186	0.9636 \pm 0.0204	0.7139 \pm 0.0283	0.8388 \pm 0.0143	0.6214 \pm 0.1005
	Weighted Loss	0.9299 \pm 0.0260	0.9409 \pm 0.0263	0.7368 \pm 0.0177	0.8388 \pm 0.0172	0.5509 \pm 0.0819
	Random Undersampling	0.8183 \pm 0.0340	0.8290 \pm 0.0604	0.8063 \pm 0.0203	0.8176 \pm 0.0339	0.8204 \pm 0.0288
Zeiss	No Sampling	0.8753 \pm 0.0207	0.7313 \pm 0.0643	0.9245 \pm 0.0266	0.8279 \pm 0.0299	0.9178 \pm 0.0139
	Weighted Loss	0.8723 \pm 0.0206	0.7237 \pm 0.0653	0.9225 \pm 0.0298	0.8231 \pm 0.0262	0.9155 \pm 0.0156
	Random Undersampling	0.8315 \pm 0.0105	0.8246 \pm 0.0336	0.8425 \pm 0.0388	0.8336 \pm 0.0130	0.8311 \pm 0.0134

Table 9: Federated 3D CNN Performance on Topcon and Zeiss Test Sets (AUROC and F1-Score, Mean \pm Std)

Data Used	Topcon AUROC	Topcon F1	Zeiss AUROC	Zeiss F1
Original Topcon	0.7658 \pm 0.0476	0.7829 \pm 0.0335	N/A	N/A
Original Zeiss	N/A	N/A	0.8438 \pm 0.0495	0.8439 \pm 0.0392
Original Topcon + Original Zeiss	0.7875 \pm 0.0230	0.8061 \pm 0.0225	0.8382 \pm 0.0436	0.8311 \pm 0.0374

$$L_{\text{weighted}} = -\frac{1}{N} \sum_{i=1}^N [w_1 y_x \log P_x + w_0 (1 - y_x) \log (1 - P_x)] \quad (14)$$

To emphasize the contribution of the minority class, we used the following weights:

$$w_0^{(1)} = \frac{272}{4932 + 272}, \quad w_1^{(1)} = 1 - w_0^{(1)} \quad (\text{Topcon})$$

$$w_0^{(2)} = \frac{847}{847 + 263}, \quad w_1^{(2)} = 1 - w_0^{(2)} \quad (\text{Zeiss})$$

These weights were applied per sample based on the ground truth label. This ensures that glaucomatous or non-glaucomatous cases that are underrepresented have a proportionately greater influence during optimization.

The second method involved random undersampling of the majority class to construct a balanced dataset during training. For each training trial, we randomly sampled an equal number of glaucomatous and non-glaucomatous volumes to form a balanced training batch. This approach guarantees equal representation of both classes during each optimization step. While it reduces the total number of available samples, it helps prevent the model from becoming biased toward the majority class and often improves performance on the minority class.

In Table 8, we report the results of three sampling strategies: No Sampling, Weighted Loss, and Random Undersampling, applied during training of the base AI-CNet3D_H model in Step 1 (Figure 3). These results show that while No Sampling and Weighted Loss improve some metrics, they fall short in others. Sensitivity is low when training on Topcon data, which has more negative examples. Conversely, specificity is low for Zeiss data, which contains more positive examples. Random undersampling results in the most balanced performance across all metrics, suggesting that neither class dominates the learning process.

A.5 Federated Training

We also evaluate a scenario where the model is trained using combined data from both datasets. Since Dataset 1 (Topcon) contains both macula and ONH regions while Dataset 2 (Zeiss) contains only ONH, we standardize the data by cropping Topcon volumes to include only the ONH region and resizing them to 64×128×64 to match the Zeiss volume dimensions.

Table 9 presents the performance comparison between the base 3D CNN trained on individual datasets versus a combined approach using the FedAvg protocol (McMahan et al., 2017). The results indicate that combining datasets does not yield performance improvements when evaluated on their respective test sets. Notably, the combined training

approach shows modest improvements on the Topcon test set but slightly reduced performance on the Zeiss test set compared to training exclusively on Zeiss data.

This lack of substantial improvement can be attributed to the fundamental differences between the two OCT acquisition systems, which produce volumes with distinct noise characteristics and imaging artifacts. Future research will focus on domain adaptation techniques to harmonize these different volume types and achieve more consistent OCT data representation across platforms.