Robust AMD Stage Grading with Exclusively OCTA Modality Leveraging 3D Volume

Haochen Zhang[1], Anna Heinke[2], Carlo Miguel B. Galang[2], Daniel N. Deussen[2], Bo Wen[1], Dirk-Uwe G. Bartsch[2], William R. Freeman[2], Truong Q. Nguyen[1], Cheolhong An[1]
[1]Electrical and Computer Engineering Department, UC San Diego
[2]Jacobs Retina Center, Shiley Eye Institute, UC San Diego
{haz035, tqn001, chan}@ucsd.edu {aheinke, dbartsch, wrfreeman}@health.ucsd.edu

Abstract

Age-related Macular Degeneration (AMD) is a degenerative eye disease that causes central vision loss. Optical Coherence Tomography Angiography (OCTA) is an emerging imaging modality that aids in the diagnosis of AMD by displaying the pathogenic vessels in the subretinal space. In this paper, we investigate the effectiveness of OCTA from the view of deep classifiers. To the best of our knowledge, this is the first study that solely uses OCTA for AMD stage grading. By developing a 2D classifier based on OCTA projections, we identify that segmentation errors in retinal layers significantly affect the accuracy of classification. To address this issue, we propose analyzing 3D OCTA volumes directly using a 2D convolutional neural network trained with additional projection supervision. Our experimental results show that we achieve over 80% accuracy on a four-stage grading task on both error-free and error-prone test sets, which is significantly higher than 60%, the accuracy of human experts. This demonstrates that OCTA provides sufficient information for AMD stage grading and the proposed 3D volume analyzer is more robust when dealing with OCTA data with segmentation errors.

1. Introduction

Age-related Macular Degeneration (AMD), one of the leading causes of severe irreversible vision impairment, is a progressive eye disease associated with abnormal vascular alteration and growth originating from the choroid. Starting from an early non-exudative stage, AMD can progress to an exudative stage where 90% of patients may lose vision [5]. Since the progression of AMD has manifestations associated most commonly with the choroidal neovascular (CNV), early detection of pathological vessels is crucial in optimal treatment management and maintaining vision for AMD patients.

However, imaging vessels within different retina layers is not supported by typical retinal imaging techniques. For example, fundus imaging can only reveal large retinal vessels, drusens, and areas of atrophy, which may indicate the presence of AMD, but make it difficult to determine the stage of the disease. Fluorescein Angiography (FA) can show CNV only at a specific time point, which is often short and challenging to capture. Optical Coherence Tomography (OCT) can display retinal layers and fluid but lacks the ability to visualize vessels. In contrast, OCT Angiography (OCTA), as an emerging imaging modality, has the capability to display vascular networks in different retinal layers [8, 25, 12], as depicted in Fig. 1 and Fig. 3. It shows superficial and deep vascular complex (SVC and DVC), avascular layer and choriocapillaris (CC). By visualizing the pathological CNV vessels directly, it enables not only an earlier detection, but also a way to monitor the clinical response to treatment. In Fig. 1, we provide a comparison between fundus and OCTA w.r.t. different AMD stages.

Unfortunately, even with the above-mentioned benefits, OCTA has not been regarded as the gold standard in clinical decision making yet, because the correlation between vessels in OCTA and AMD stages is not strictly proven. On the clinical side, ophthalmologists are actively searching biomarkers for AMD diagnosis from OCTA, mainly based on manual analysis and their own experience. In this work, we present experimental evidence of the informativeness of OCTA from the perspective of data-driven classifiers. We believe that deep learning is capable of this task with two advantages. Firstly, some deep learning algorithms have been proven to surpass human-level performance on natural image classification [14]. Moreover, it is more efficient for computer to handle 3D data or multiple projections than human. Consequently, we expect that deep learning classifiers would identify hidden patterns imperceptible to human

This work is supported by grant NIH R01 EY033847-02.

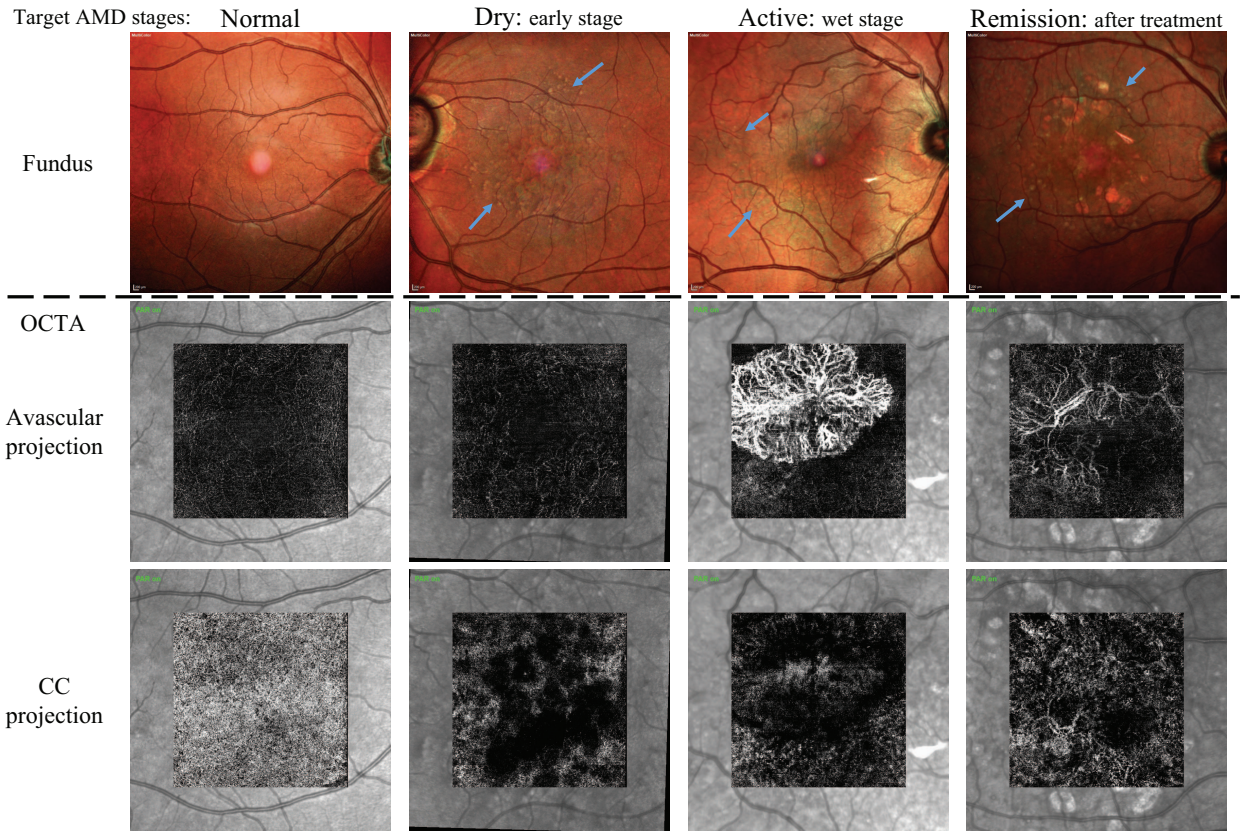

Figure 1. Comparison between fundus and OCTA w.r.t. AMD stages. As shown by the blue arrows, all AMD stages exhibit drusens and it is difficult to differentiate each stage based on the pattern of drusens. For instance, in the provided example, the early stage (dry) displays clearer drusens than the progressive stage (active). In contrast, OCTA allows for a distinction between dry and normal stages using the hollows in CC projection, and between active and dry stages with the presence of CNV in avascular projection. It is still an ongoing challenge to tell active stage from remission for human experts, yet this paper demonstrates it is achievable with the proposed deep classifiers in both 2D and 3D cases.

eyes and improve AMD diagnosis.

In this paper, we focus on OCTA modality only and build a series of deep learning based AMD stage graders. We summarize our contributions as follows:

- We experimentally verified that the OCTA projections, which ophthalmologists usually use for diagnosis, are easily affected by layer segmentation errors. Those errors degrade the classification performance.
- We propose to use 3D raw OCTA volume to avoid the impacts of those errors. To achieve this, we modify a pretrained 2D network to perform volume classification. We also adopt an additional projection supervision to facilitate training of shallow feature extractor.
- Experimental results show that the proposed classifier can achieve the accuracy of more than 80%, regardless of the presence of layer segmentation errors. These results prove the effectiveness of our methods and suggest that OCTA is a promising modality to distinguish various stages of AMD disease.

2. Related Works

OCTA analysis in computer vision. In recent years, OCTA has emerged as a valuable tool in ophthalmology, offering a non-invasive way to visualize and analyze the vascular network of the retina. Therefore in the realm of computer vision, most OCTA-based works have focused on segmentation tasks. Alam et al [1] used U-Net to perform artery-vein classification and adopted transfer learning to compensate for the small dataset. In [13], the avascular area was detected in OCTA projections with a multi-scaled encoder-decoder neural network. Li et al [19] proposed to segment vessels with 3D OCTA inputs to get rid of projection images and retinal layer segmentation. In addition to segmentation, there are also some deep learning-based OCTA classification works. For example, Le et al [18] adopted the VGG16 network to classify diabetic retinopathy stages. Lin et al [20] went further and performed classification and segmentation simultaneously using boundary shape and distance map as additional supervision to im-

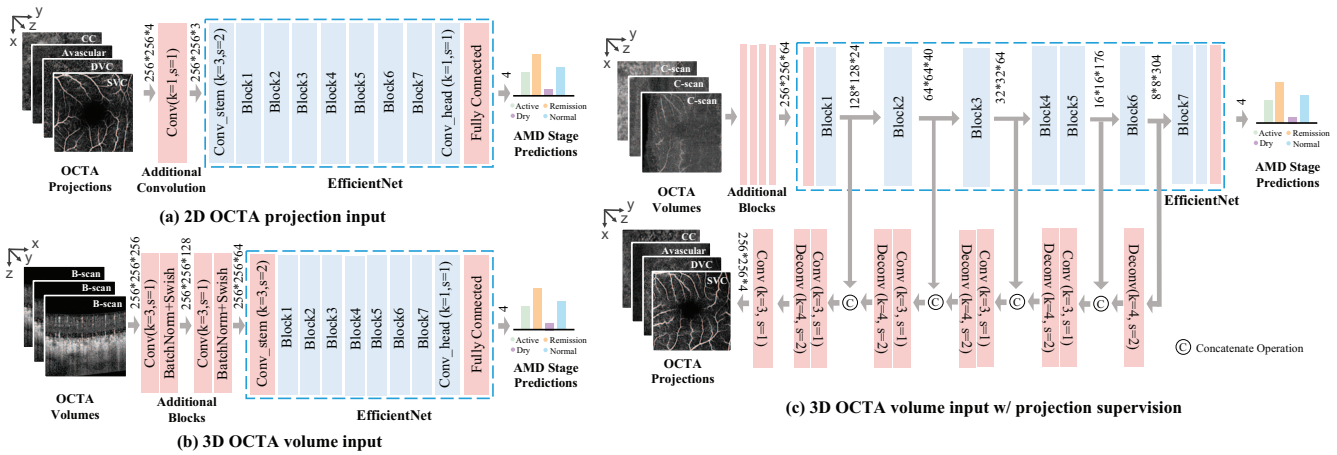

Figure 2. The proposed network structures for (a) 2D projections, (b) 3D volumes and (c) 3D volumes with 2D projection supervision. The layers in blue have pretrained weights while those in red are trained from scratch.

prove accuracy. Apart from classification and segmentation, some researchers have focused on 3D vessel reconstruction [35], projection quality assessments [33] and improving the en face OCTA generation [36]. Although these works have shown promising results, none of them have considered the grading of AMD stage, which is a critical task in the clinical management of AMD patients.

AMD diagnosis with deep learning. To the best of our knowledge, there is no existing AMD diagnosis work using OCTA modality only. Instead, they usually use color fundus, FA and most recently OCT modality. Alqudah et al [2] trained a customized CNN to classify retina into five distinct stages of AMD based on OCT B-scans. Motozawa et al [21] first classified AMD/no AMD and then identified the presence of exudative changes. Das et al [9] integrated multi-scale deep image features to enhance OCT classification. He et al [16] leveraged GANs to generate synthetic images in order to increase training data size. In addition to stage classification, Banerjee et al [3] combined hand-craft and CNN features in a LSTM to predict AMD progression. Rakocz et al [23] designed a SLIVER-net to classify risk factors of AMD progression which could operate on both 2D B-scans and 3D volumes. Russakoff et al [27] predicted the likelihood of converting from early/intermediate to advanced AMD. Furthermore, there have been several recent works [31, 17, 30] that employ multimodal images such as fundus photographs, OCT B-scans, and OCTA projections to grade AMD. In this paper, we focus on the latest work [30] in Sec. 4.2 for comparison, which utilizes OCT B-scans, OCT projections and OCTA projections.

OCTA datasets. The advancement in deep learning has led to significant progress in the field of retinal disease diagnosis and management. Various challenges have been organized to evaluate the performance of computer-aided diagnosis systems on different retinal diseases, such as glaucoma and AMD. The GAMMA challenge [34] is one such challenge, which provides 2D fundus image and 3D OCT volume, focusing on glaucoma diagnosis. The ADAM challenge [11] evaluates the performance of automated AMD diagnosis based on fundus image. Although these challenges have provided valuable insights into the development of automated diagnosis systems, they do not include OCTA information in their datasets, which is the key objective of this paper. Consequently, it is impossible for us to experiment on those datasets. In the supplementary material, we report the detailed information about existing OCTA datasets to show their limitation in OCTA based AMD stage grading. In this paper, we experiment with an OCTA dataset collected by ourselves, which has the largest number of AMD samples available and is specifically curated for AMD stage grading task.

3. Methods

3.1. 2D Classifier based on OCTA projection images

In clinical practice, ophthalmologists usually refer to OCTA projections for diagnosis, inspiring related classifiers [18, 30] using the same inputs. In this section, we also develop a baseline classifier with 2D OCTA projection inputs, for analysis and comparison.

Classifier structure. Different from existing method [30], which used a custom CNN without pretraining, our approach utilizes a well established image classification network as backbone. Moreover, we pretrain the backbone with ImageNet [10], and subsequently fine-tune it with our OCTA projections. As shown in Fig. 2 (a), we adopt the EfficientNet in our network, because it is reported to achieve the best trade-off between performance and model size [29]. Since we set up four input channels to take four OCTA projections, we include an additional convolution layer with

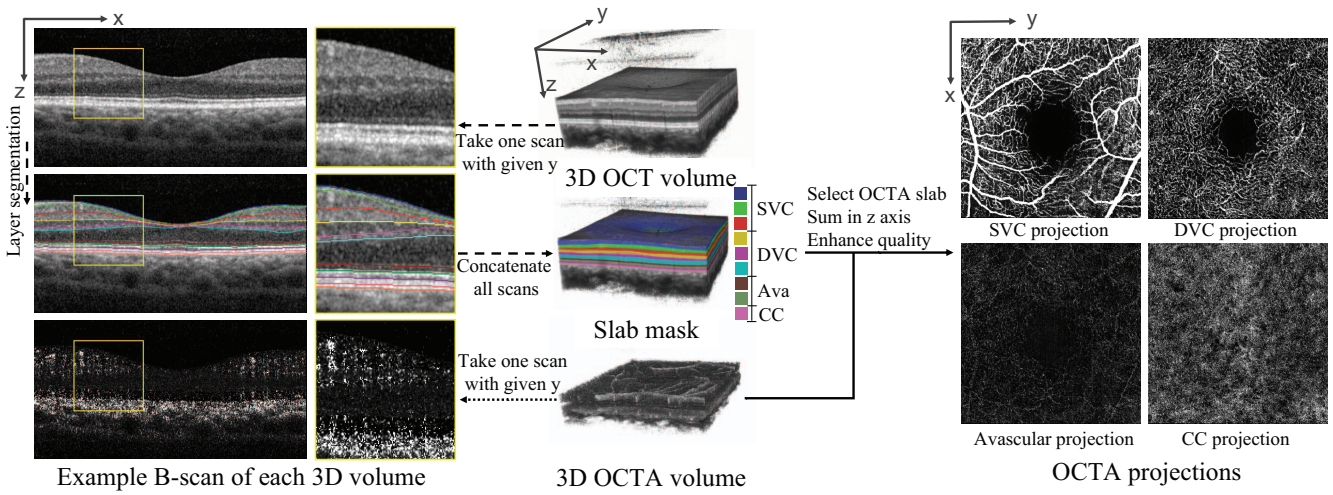

Figure 3. Illustration of the interrelationships among OCT and OCTA raw volume, B-scans, and OCTA projection. A single B-scan is a cross-sectional slice of the 3D volume with a specific y-axis value. Retinal slab masks are derived from retinal layer segmentation in each B-scan of 3D OCT volume. OCTA projections are generated by summing up the motion responses in selected OCTA slabs followed by quality enhancement.

kernel size 1 before the EfficientNet to address channel mismatching. Additionally, since we only have four target categories, we adjusted the output of the last fully-connected layer to match the number of categories.

Warmup strategy. Consequently as shown in Fig. 2, the layers are divided into two groups: the red layers with no pretrained weights and the blue layers pretrained with Imagenet [10]. Since different layers have different initialization weights, the red layers could disrupt the tuning of the blue ones if fine-tuned all the layers together. So we use a warmup strategy as follows. We first freeze all the blue layers and train only the red ones for 600 epochs. During this step, we also train all the BatchNorm layers to better transfer from natural images distribution to OCTA projections distribution. Then we finetune all the layers together for another 900 epochs with a smaller learning rate.

3.2. Presence of layer segmentation errors

During the development of our 2D classifier, we find that OCTA projections are not always reliable due to their sensitivity to the quality of retinal layer segmentation, which plays an important role in OCTA projection generation. This problem is common but often overlooked in most published literature [18, 30]. It is worth noting that previous research [19] has also reported that failures in layer segmentation can lead to difficulties in OCTA vessel segmentation. In this section, we aim to investigate the prevalence of layer segmentation errors and their impact in context of AMD stages grading.

OCTA projection generation. Raw OCTA data capture the movements of blood in a 3D retinal space which are difficult to interpret by humans. Therefore, OCTA imaging machines commonly project raw OCTA volumes onto 2D images to enhance their visual interpretation. The projection process may differ among commercial instruments. Here, we consider the image taken by Heidelberg[1] as an example [24]. As illustrated in Fig. 3, the Heidelberg software estimates the boundaries of different retinal layers to divide the 3D space into several slabs. Within selected slabs, which are determined by anatomical criteria, it calculates the summation of OCTA responses along z-axis to generate a 2D image. Additionally, the software employs a contrast function and a projection artifact removal algorithm to enhance the image quality. When executed successfully, these steps produce highly informative and visually appealing 2D images that are easily interpretable by doctors.

Influence of segmentation errors. Unfortunately, the estimated layer boundaries in the first step are not always accurate, resulting in segmentation errors that significantly impact the quality of OCTA projections. Since most commercial instruments usually estimate those boundaries based on image gradient and graphcut algorithm [28], which is not robust enough, the layer segmentation errors are actually prevailing, especially for distorted retina with AMD disease. To gain a better understanding of the magnitude of the problem, we conduct a manual check of 530 OCTA samples from different AMD stages and report the results in Table 1. Not surprisingly, we find that almost three-fourths of samples in the active stage have layer segmentation errors. The overall error rate among 530 samples is as high as 54.3% and, more accurately, we can calculate the balanced overall error rate by averaging the last row

[1]The Heidelberg HRA+OCT Spectralis System, version 1.11.2.0 (Heidelberg Engineering, Heidelberg, Germany)

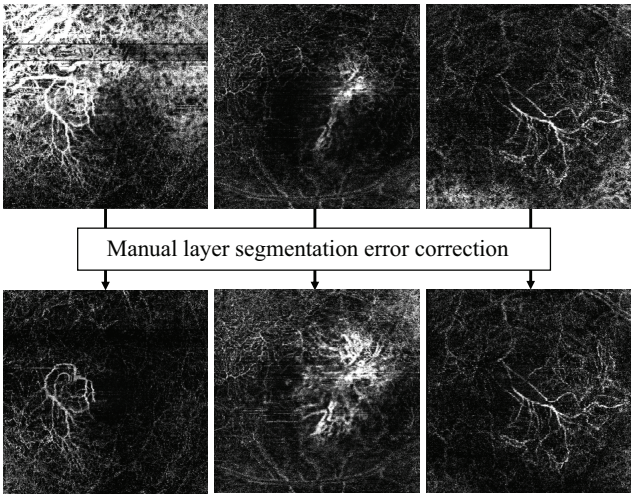

Figure 4. Examples of avascular projection w/o and w/ manual layer segmentation error correction by human experts. Layer segmentation errors lead to incorrect vascular networks, missing vessels and noise in OCTA projections, which complicates the classification for both ophthalmologists and neural networks.

of Table 1, which is 46.2%. These findings indicate that the problem of layer segmentation error is pervasive and requires urgent attention. As shown in Fig. 4, layer segmentation errors lead to incorrect vascular networks or missing vessels in OCTA projections, which complicates the classification for both ophthalmologists and neural networks. The influence of layer segmentation errors on deep classifier is quantified and discussed in Sec. 4.1.

3.3. Avoid segmentation errors with 3D input

Since the errors in layer segmentation significantly affect the quality of OCTA projections, we propose to directly apply raw OCTA volume[2] for classification. In this section, we provide a detailed description of our method, which utilizes a 2D convolutional neural network to analyze 3D OCTA data. We then delve into the reasons behind our choice of channel dimension and how we further improve the training process to achieve optimal performance.

2D backbone for volume classification. Considering that there is no available large-scale 3D dataset for pretraining a 3D classifier, we use a 2D network with pretrained weights to analyze 3D data. It means that we take one dimension of 3D as channel and the other two as spatial. As shown in Fig. 2 (b), we gradually reduce the input channel by extending the additional convolution to two Conv-BN-Swish blocks with kernel size 3. Each block divides the channels by 2 and the input channel of the EfficientNet is ultimately revised to a desired number, i.e. 64 in our experiments. Based on the ablation experiments reported in Table 5, we find that better accuracy is achieved by treating the dimensions of B-scan as spatial and incorporating different B-scans in the channel dimension, i.e. taking the y-axis as the channel.

[2]OCTA volume in this paper represents for the raw blood motion responses in 3D space before projection. No structural OCT B-scan is used in this method.

Table 1. Distribution of error-free and error-prone samples and associated error rates. Clearly, samples with more severe AMD have larger error rate. The overall error rate shows layer segmentation error is a common problem in OCTA projections. Please refer to Fig. 4 for visual indication of the detrimental effects of segmentation errors on the quality of the OCTA projections.

sample type	Active	Remission	Dry	Normal	Total
# w/ seg. error	138	91	57	2	288
# w/o seg. error	52	39	90	61	242
error percentage	72.6%	70%	38.8%	3.2%	54.3%

Why y-axis is better. This result is not in line with our expectations, because the 2D network, which takes OCTA projections as input, is equivalent to treating the z-axis as a channel. So we investigated this issue and identified an explanation. In typical convolutional networks, the first convolution layer reduces the spatial resolution by a factor of 2 while significantly increases the number of channels, for instance, from 3 to 64. Consequently, there is no significant loss of information in this layer. In contrast, our additional convolution blocks drastically reduce the number of channels, from 256 to 64, resulting in a loss of information if they are not appropriately trained. When considering the z-axis as a channel, this loss of information is especially significant. However, it is less pronounced when using the y-axis as the channel because consecutive B-scans are often similar to each other and contain a lot of redundancy.

Projection supervision. This analysis leads to a method further enhancing the performance, whose key idea is to improve the training of shallow feature extractor. To achieve this, we propose to add another branch onto the EfficientNet backbone, as illustrated in Fig. 2 (c). This newly added branch functions in a similar way to the decoder of the Unet [26] and is capable of generating OCTA projections from the 3D OCTA volume. By doing so, the additional convolutional blocks, along with some shallow layers in EfficientNet, can better preserve the information necessary for displaying vessel patterns and aiding in AMD grading. It is worth noting that this branch serves only for loss calculation and can be discarded during the inference stage. As a result, we improve accuracy without requiring additional inputs or incurring extra inference time costs.

4. Experiment Results

Dataset. Because there is no public OCTA dataset suitable for AMD stage grading, we use our own dataset collected from Jacobs Retina Center at Shiley Eye Institute in experiments. The dataset consists of 889 raw OCTA

volumes with corresponding projections belonging to four AMD stages: active, remission, dry and normal. Please refer to Fig. 1 for examples. 'Active' means the pathogenic vessels are leaking fluid while 'remission' means the pathological vessels were once active but recovered after treatment and showing no fluid. 'Dry' represents an early stage of AMD which is not exudative and 'normal', as name implies, is obtained by imaging healthy retina. For dataset division, we firstly choose a predetermined number of samples from each category to form the testing set. Then, we randomly select validation set from the remaining samples to conduct a 5-fold validation experiment. Following this strategy, we created two sub-datasets: an easier subset which only had samples with no layer segmentation errors in its testing set, indicated as 'error-free' and a harder subset containing numerous samples with errors in its testing set, indicated as 'error-prone'. Please refer to the supplementary material for more details about the dataset design.

Implementation. We implement all our deep classifiers on PyTorch platform. To save GPU memory, we downsample OCTA projections and volumes to 256×256 and 256×256×256, respectively. Then we adopt several data augmentations to increase their diversity. In detail, we use random flipping, rotation and cropping with resizing. We randomly apply gamma transformation and Gaussian smooth to increase the diversity of intensity. For projections, we also use grid distortion to augment the shapes. For both 2D and 3D data, we adopt a sample-wise normalization to whiten the sample intensity. Oversampling training data in each category is used to balance their distribution. The networks are trained by Adam optimizer with 10^{-5} weight decay. The initial learning rate is 10^{-3} and decreases via a cosine scheduler with minimum value 10^{-5}. The cosine loss serves as our optimization target, which is proven to be effective with small data amounts [4]. For projection supervision branch, we employ MSE to compute projection differences, and the ratio between cosine loss and MSE is decided by ablation experiments shown in Table 5. We have released our models and codes on this website.

4.1. Influence of layer segmentation errors

We create two datasets to assess the impact of layer segmentation errors: 'clean' and 'mixed'. They have the same size but the 'clean' set only includes error-free samples, while the 'mixed' set includes data with and without segmentation errors. For 'clean' dataset, we randomly selected 14 samples from each category for testing and used the remaining samples as training. Then we considered samples with errors. For the three categories except 'normal', we replaced 7 testing samples with randomly selected 7 samples with errors. Consequently, we obtained a testing set that has the same scale as 'clean' but includes data both with and without errors. We generated a training set with same properties by running the same process and name this dataset as 'mixed'. By considering both "mixed" and "clean" dataset, we plan to simulate the process in which we correct layer segmentation errors in 'mixed' dataset.

Table 2. Classification accuracy with different training/testing datasets. 'Clean' means a set with no segmentation errors and 'Mixed' means a set mixed with samples with and without errors.

Train on	Test on	Accuracy
Clean set	Clean set	69.64%
Clean set	Mixed set	53.57%
Mixed set	Clean set	64.29%
Mixed set	Mixed set	57.14%

We conduct 5-fold validation experiments using Resnet18 [15] on both datasets and use the ensemble prediction as the final result by averaging the predictions of 5 classifiers trained in each fold. Note that we can choose to train and test with either 'clean' or 'mixed' set, resulting in 4 different combinations, shown in Table 2. The first two rows of Table 2 show that the classifier struggles to generalize from clean samples to those with errors, indicating data with and without errors follow different distributions. Taking the last row into account, we find adding samples with errors to the training set benefits, showing that the classifier may learn the joint distribution of samples with and without errors if given enough training data. The accuracy in the last two rows shows that, even trained on samples with errors, the clean test still works better, implying that samples with errors are hard to learn.

This experiment suggests two ways for improving the performance: 1) collecting enough data to cover the joint distribution of samples with and without errors; 2) avoiding layer segmentation errors and reducing the gap between each distribution. We focus on the second option, as it is not practical to collect sufficient data in a short time.

4.2. Performance of deep classifier

In this section, we experiment mainly on two datasets, namely error-free and error-prone. For the error-free test set, we utilized the clean test set from Sec. 4.1. However, as indicated in Table 2, the training set must be cleaned to enhance its performance. Therefore, we integrated error-free training samples along with samples without error annotations, referred to as 'unknown' samples, while eliminating all known error-prone training samples. By adopting this approach, we can effectively cleanse the training set while keeping its size. In contrast, the error-prone test set comprises solely of samples containing errors in all AMD stages, and all samples except those designated for testing were utilized to construct the error-prone training set. More detail about relevant datasets can be found in the supplementary material.

Table 3. Ensemble accuracy (%) and RoC-AUC performance of different AMD graders with 2D inputs. Error-free and Error-prone are two testing sets w/ and w/o segmentation errors, respectively. MM: Multimodal information (including OCT B-scan, OCT and OCTA projections), PT: Pretraining.

2D Input	Setting		Error-free		Error-prone	
	MM	PT	Accuracy	AUC	Accuracy	AUC
Thakoor et. al. [30]	✗	✗	55.36	0.8159	57	0.8176
	✓	✗	62.5	0.8512	66	0.8428
ours(2D)	✗	✗	73.21	0.8565	62	0.8065
	✗	✓	**80.36**	**0.9264**	**72**	**0.8697**
Human	-	-	58.92	-	60	-

Table 4. Ensemble accuracy (%) and RoC-AUC performance of different AMD graders with 3D inputs. PT: Pretraining, PS: Projection Supervision.

3D Input	Setting		Error-free		Error-prone	
	PT	PS	Accuracy	AUC	Accuracy	AUC
Effic.Net 3D	✗	✗	75	0.9489	69	0.8841
Med.Net34 [7]	✓	✗	73.21	0.9238	73	0.9009
ours(3D)	✓	✗	82.14	**0.9524**	74	0.9055
	✓	✓	**83.93**	0.9298	**80**	**0.912**

For baseline method, as far as we know, there is no deep learning based AMD stage grader using OCTA only. Therefore, we use a multimodal AMD grader [30] for performance comparison. We train their networks on our dataset for fair comparison since their dataset is not publicly available. We implement two classifiers based on their official codes which use OCTA information only and use multimodal information from OCT B-scan, OCT and OCTA projection. Note that there are two differences between their task and ours: 1) they do not have 'remission' in their target categories, and 2) we do not have high-definition OCT B-scans in our dataset, so we use common B-scans as an alternative. We also replace ORCC projection used in their experiments with SVC projection. We conduct 5-fold validation experiments on two sub-datasets: an easier subset which only has samples with no layer segmentation errors in its testing set (error-free), and a harder subset containing numerous samples with errors (error-prone). In Table 3 and 4, we report the ensemble accuracy, AUC of RoC in 'one v.s. rest' manner, and the performance of human experts on the same test sets. Please refer to supplement material for details of human expert evaluation.

As shown in Table 3, Ours-2D with pretrained weights significantly outperforms Thakoor et. al. [30] regardless of the use of multimodal information. This is due to the difference in network structure and training strategy. Note that [30] trained a customized network with four 3D convolution and three fully connected layers from scratch, which is much simpler than EfficientNet. The benefit of EfficientNet backbone is evident from the first and third rows and, as shown in the third and fourth rows, pretraining the model further improves its ability to identify useful patterns in OCTA projections. Note that Ours-2D demonstrates significant improvements compared to human experts, indicating the potential of OCTA as a diagnostic modality in AMD grading. These promising results call for further exploration of OCTA-derived biomarkers for accurate AMD diagnosis.

When considering Ours-3D in Table 4, we observed a notable improvement compared to Ours-2D. Since the structures of both networks are quite similar (Fig. 2), this gain demonstrates the advantages of directly grading 3D OCTA volumes and reducing the gap between data with and without errors. The advantage of Ours-3D method can be also substantiated by examining the performance differences of Ours-2D and Ours-3D in error-free and error-prone settings. In the error-free setting, where fewer samples are affected by errors, the improvement gained from using Ours-3D is relatively smaller. However, in the error-prone setting, where errors are more prevalent, the performance of Ours-2D experiences a significant decline, while Ours-3D maintains high performance levels. This differential behavior in error-free and error-prone settings serves as evidence that the proposed Ours-3D method is more robust in the presence of layer segmentation errors.

In comparing Ours-3D with 3D EfficientNet and MedicalNet34 [7], both of which utilize 3D convolutions, we find that 2D backbone is more effective. This finding is actually consistent with some early works in action recognition [32, 22, 6]. Their experiments verified that well-designed 2D convolution network is better than 3D, especially when training data is limited. Our result indicates that utilizing a pretrained 2D network is currently a promising method for analyzing 3D OCTA until a large-scale 3D OCTA dataset is available. Finally, the efficacy of our proposed projection supervision is demonstrated in the last two rows, where the accuracy is improved to over 80%. It also indicates that OCTA is an informative modality for AMD grading.

4.3. Ablation study

This section presents our ablation experiments, which aim to investigate the impact of different factors on the performance of our classifiers. Specifically, we examine the effects of different choices of channel axis, different ratios of loss weights, and the use of pretrained weights and projection supervision. The accuracy of different classifiers trained on the first validation fold are reported in Table 5.

Firstly, our results indicate that taking y-axis as the channel is more effective than z-axis when projection supervision is not used. The reason has been elaborated in Sec. 3.3. Then the use of projection supervision improves the z-axis inputs while negatively impacting y-axis channel inputs. This outcome is consistent with our expectations since taking z-axis as channel means taking x and y dimension as

Table 5. Ablation experiments w.r.t the choice of channel axis and the loss weight ratio. The accuracy here pertains to the performance of individual classifier trained on the first validation fold, instead of the outcome of ensemble.

Settings				Accuracy (%)
Channel axis	Pretrain	Proj. Supervision	Weight ratio	
y axis				54
y axis	✓			69
y axis	✓	✓	1:10	64
z axis				50
z axis	✓			64
z axis	✓	✓	$1{:}10^{-1}$	69
z axis	✓	✓	1:10	72
z axis	✓	✓	$1{:}10^{3}$	74
z axis	✓	✓	$1{:}10^{4}$	67

spatial which aligns with the spatial dimension of OCTA projections. It is unreasonable to generate OCTA projections from a stack of OCTA B-scans. Furthermore, our experiments on various weight ratios demonstrate that the ideal ratio between Cosine loss and MSE loss is approximately $1{:}10^{3}$ for z-axis channel inputs. Finally, our experiments also show that the use of pretrained weights improves the performance of the classifiers, regardless of which dimension is selected as the channel.

4.4. Detailed comparison with human expert

As described in Fig. 1, we expect our method outperforms human experts in this four-stage grading task. To compare the performance of our proposed method with that of human experts, in this section, we conducted a detailed analysis of the confusion matrix in different settings.

Firstly, we evaluated the matrix of the human expert on the error-free test set. It can be observed that the ophthalmologist who took this experiment performed well in distinguishing between the 'dry' and 'normal' categories but struggled in differentiating between the 'remission' and 'active' categories. This highlights the ongoing challenge in accurately determining the active stage of AMD for human experts, thereby emphasizing the significance of our work.

Subsequently, we analyzed the matrix of the human expert on an error-prone test set. It can be found that the human expert continued to face difficulty in distinguishing between the 'active' and 'remission' categories, but this time, the accuracy of the 'dry' category significantly decreased. This is exactly the consequence caused by layer segmentation errors, i.e. the incorrect vascular networks and missing vessels in the OCTA projections caused confusion for the human expert.

In contrast, our proposed method, termed Ours-3D, shows a significant improvement in the confusion matrix, accurately classifying the majority of test samples in each category. On the error-free test set, Ours-3D performed slightly worse in the 'remission' category, owing to the relatively fewer training samples in this category. On the error-prone test set, our method demonstrated greater robustness to segmentation errors by directly taking the raw OCTA volume as input and bypassing the impact of those errors. Overall, our proposed method not only outperforms human experts in this AMD grading task but also offers increased robustness to segmentation errors, which is a critical consideration in accurately detecting and grading AMD.

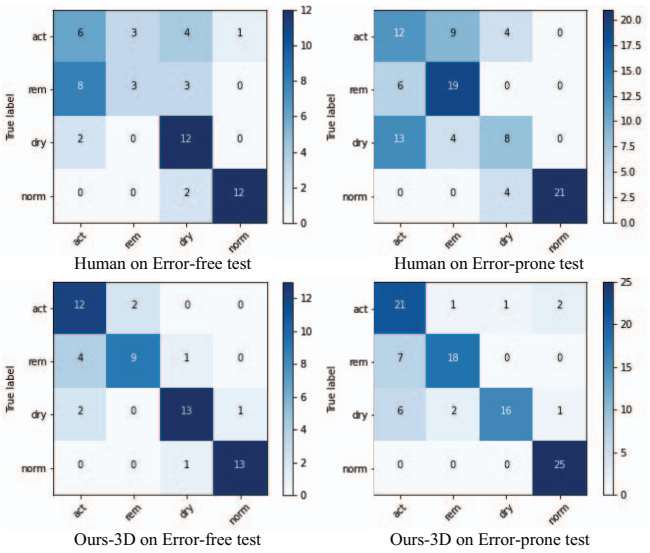

Figure 5. Confusion matrix comparison between our proposed method and human experts on different test sets. Ours-3D, outperforms human experts in accurately distinguishing between the 'active' and 'remission' categories. Also, as indicated by the smaller performance drop observed in the 'dry' category, our method demonstrates greater robustness to layer segmentation errors.

5. Conclusion

In this paper, we firstly elaborate the influence of layer segmentation errors in the context of AMD stage grading and propose to address it via analyzing the 3D OCTA volume directly. With the pretrained 2D EfficientNet backbone and projection supervision, we achieve an accuracy of over 80% on both error-free and -prone test sets, which significantly outperforms 60% accuracy of human experts. Our results suggest that OCTA modality alone can identify different AMD stages and encourage the exploration of OCTA-derived biomarkers for diagnosis. In future work, we plan to explain the decision-making of these well-performed classifiers so as to develop deep learning-based biomarkers for accurate AMD diagnosis.

6. Acknowledgment

We thank Prof. Kaveri A. Thakoor who shares the source codes and helps in implementing their multimodal methods.

References

[1] Minhaj Alam, David Le, Taeyoon Son, Jennifer I Lim, and Xincheng Yao. AV-Net: deep learning for fully automated artery-vein classification in optical coherence tomography angiography. *Biomedical optics express*, 11(9):5249–5257, 2020.

[2] Ali Mohammad Alqudah. AOCT-NET: A convolutional network automated classification of multiclass retinal diseases using spectral-domain optical coherence tomography images. *Medical & biological engineering & computing*, 58(1):41–53, 2020.

[3] Khaled Alsaih, Mohd Zuki Yusoff, Tong Boon Tang, Ibrahima Faye, and Fabrice Mériaudeau. Deep learning architectures analysis for age-related macular degeneration segmentation on optical coherence tomography scans. *Computer methods and programs in biomedicine*, 195:105566, 2020.

[4] Bjorn Barz and Joachim Denzler. Deep learning on small datasets without pre-training using cosine loss. In *WACV*, pages 1371–1380, 2020.

[5] Rupert RA Bourne, Jost B Jonas, Seth R Flaxman, Jill Keeffe, Janet Leasher, Kovin Naidoo, Maurizio B Parodi, Konrad Pesudovs, Holly Price, Richard A White, et al. Prevalence and causes of vision loss in high-income countries and in eastern and central europe: 1990–2010. *British Journal of Ophthalmology*, 98(5):629–638, 2014.

[6] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. In *CVPR*, pages 6299–6308, 2017.

[7] Sihong Chen, Kai Ma, and Yefeng Zheng. Med3D: Transfer learning for 3D medical image analysis. *arXiv preprint arXiv:1904.00625*, 2019.

[8] Gabriel J Coscas, Marco Lupidi, Florence Coscas, Carlo Cagini, and Eric H Souied. Optical coherence tomography angiography versus traditional multimodal imaging in assessing the activity of exudative age-related macular degeneration: a new diagnostic challenge. *Retina*, 35(11):2219–2228, 2015.

[9] Vineeta Das, Samarendra Dandapat, and Prabin Kumar Bora. Multi-scale deep feature fusion for automated classification of macular pathologies from OCT images. *Biomedical signal processing and Control*, 54:101605, 2019.

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009.

[11] Huihui Fang, Fei Li, Huazhu Fu, Xu Sun, Xingxing Cao, Fengbin Lin, Jaemin Son, Sunho Kim, Gwenole Quellec, Sarah Matta, et al. ADAM challenge: Detecting age-related macular degeneration from fundus images. *IEEE Transactions on Medical Imaging*, 41(10):2828–2847, 2022.

[12] Marie-Louise Farecki, Matthias Gutfleisch, Henrik Faatz, Kai Rothaus, Britta Heimes, Georg Spital, Albrecht Lommatzsch, and Daniel Pauleikhoff. Characteristics of type 1 and 2 CNV in exudative AMD in OCT-Angiography. *Graefe's Archive for Clinical and Experimental Ophthalmology*, 255:913–921, 2017.

[13] Yukun Guo, Acner Camino, Jie Wang, David Huang, Thomas S Hwang, and Yali Jia. MEDnet, a neural network for automated detection of avascular area in OCT angiography. *Biomedical optics express*, 9(11):5147–5158, 2018.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *ICCV*, pages 1026–1034, 2015.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[16] Xingxin He, Leyuan Fang, Hossein Rabbani, Xiangdong Chen, and Zhimin Liu. Retinal optical coherence tomography image classification with label smoothing generative adversarial network. *Neurocomputing*, 405:37–47, 2020.

[17] Kai Jin, Yan Yan, Menglu Chen, Jun Wang, Xiangji Pan, Xindi Liu, Mushui Liu, Lixia Lou, Yao Wang, and Juan Ye. Multimodal deep learning with feature level fusion for identification of choroidal neovascularization activity in age-related macular degeneration. *Acta Ophthalmologica*, 100(2):e512–e520, 2022.

[18] David Le, Minhaj Alam, Cham K Yao, Jennifer I Lim, Yi-Ting Hsieh, Robison VP Chan, Devrim Toslak, and Xincheng Yao. Transfer learning for automated OCTA detection of diabetic retinopathy. *Translational Vision Science & Technology*, 9(2):35–35, 2020.

[19] Mingchao Li, Yerui Chen, Zexuan Ji, Keren Xie, Songtao Yuan, Qiang Chen, and Shuo Li. Image projection network: 3D to 2D image segmentation in OCTA images. *IEEE Transactions on Medical Imaging*, 39(11):3343–3354, 2020.

[20] Li Lin, Zhonghua Wang, Jiewei Wu, Yijin Huang, Junyan Lyu, Pujin Cheng, Jiong Wu, and Xiaoying Tang. BSDA-net: A boundary shape and distance aware joint learning framework for segmenting and classifying OCTA images. In *MICCAI*, pages 65–75, 2021.

[21] Naohiro Motozawa, Guangzhou An, Seiji Takagi, Shohei Kitahata, Michiko Mandai, Yasuhiko Hirami, Hideo Yokota, Masahiro Akiba, Akitaka Tsujikawa, Masayo Takahashi, et al. Optical coherence tomography-based deep-learning models for classifying normal and age-related macular degeneration and exudative and non-exudative age-related macular degeneration changes. *Ophthalmology and therapy*, 8(4):527–539, 2019.

[22] Zhaofan Qiu, Ting Yao, and Tao Mei. Learning spatio-temporal representation with pseudo-3D residual networks. In *ICCV*, pages 5533–5541, 2017.

[23] Nadav Rakocz, Jeffrey N Chiang, Muneeswar G Nittala, Giulia Corradetti, Liran Tiosano, Swetha Velaga, Michael Thompson, Brian L Hill, Sriram Sankararaman, Jonathan L Haines, et al. Automated identification of clinical features from sparsely annotated 3-dimensional medical imaging. *NPJ digital medicine*, 4(1):1–13, 2021.

[24] Roland Rocholz, Michel M. Teussink, Rosa Dolz-Marco, Claudia Holzhey, Jan F. Dechent, Ali Tafreshi, and Stephan Schulz. SPECTRALIS Optical Coherence Tomography Angiography (OCTA): Principles and Clinical Applications. from https://www.heidelbergengineering.

com/media/e-learning/Totara/Dateien/pdf-tutorials/210111-001_SPECTRALIS%20OCTA%20-%20Principles%20and%20Clinical%20Applications_EN.pdf.

[25] Luiz Roisman, Qinqin Zhang, Ruikang K Wang, Giovanni Gregori, Anqi Zhang, Chieh-Li Chen, Mary K Durbin, Lin An, Paul F Stetson, Gillian Robbins, et al. Optical coherence tomography angiography of asymptomatic neovascularization in intermediate age-related macular degeneration. *Ophthalmology*, 123(6):1309–1319, 2016.

[26] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241, 2015.

[27] Daniel B Russakoff, Ali Lamin, Jonathan D Oakley, Adam M Dubis, and Sobha Sivaprasad. Deep learning for prediction of AMD progression: A pilot study. *Investigative ophthalmology & visual science*, 60(2):712–722, 2019.

[28] Julia Schottenhamml, Eric M Moult, Stefan B Ploner, Siyu Chen, Eduardo Novais, Lennart Husvogt, Jay S Duker, Nadia K Waheed, James G Fujimoto, and Andreas K Maier. OCT-OCTA segmentation: Combining structural and blood flow information to segment Bruch's membrane. *Biomedical Optics Express*, 12(1):84–99, 2021.

[29] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *ICML*, pages 6105–6114, 2019.

[30] Kaveri A Thakoor, Jiaang Yao, Darius Bordbar, Omar Moussa, Weijie Lin, Paul Sajda, and Royce WS Chen. A multimodal deep learning system to distinguish late stages of AMD and to compare expert vs. AI ocular biomarkers. *Scientific reports*, 12(1):1–11, 2022.

[31] Ehsan Vaghefi, Sophie Hill, Hannah M Kersten, and David Squirrell. Multimodal retinal image analysis via deep learning for the diagnosis of intermediate dry age-related macular degeneration: A feasibility study. *Journal of Ophthalmology*, 2020, 2020.

[32] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*, pages 20–36, 2016.

[33] Yufei Wang, Yiqing Shen, Meng Yuan, Jing Xu, Bin Yang, Chi Liu, Wenjia Cai, Weijing Cheng, and Wei Wang. A deep learning-based quality assessment and segmentation system with a large-scale benchmark dataset for optical coherence tomographic angiography image. *arXiv preprint arXiv:2107.10476*, 2021.

[34] Junde Wu, Huihui Fang, Fei Li, Huazhu Fu, Fengbin Lin, Jiongcheng Li, Lexing Huang, Qinji Yu, Sifan Song, Xingxing Xu, et al. Gamma challenge: Glaucoma grading from multi-modality images. *arXiv preprint arXiv:2202.06511*, 2022.

[35] Shuai Yu, Yonghuai Liu, Jiong Zhang, Jianyang Xie, Yalin Zheng, Jiang Liu, and Yitian Zhao. Cross-domain depth estimation network for 3D vessel reconstruction in OCT angiography. In *MICCAI*, pages 13–23, 2021.

[36] Yuhan Zhang, Chen Huang, Mingchao Li, Sha Xie, Keren Xie, Zexuan Ji, Songtao Yuan, and Qiang Chen. Robust layer segmentation against complex retinal abnormalities for en face OCTA generation. In *MICCAI*, pages 647–655, 2020.