Learning Two-Stream CNN for Multi-Modal Age-related Macular Degeneration Categorization

Weisen Wang, Xirong Li, Zhiyan Xu, Weihong Yu, Jianchun Zhao, Dayong Ding, Youxin Chen

***Abstract*—This paper tackles automated categorization of Age-related Macular Degeneration (AMD), a common macular disease among people over 50. Previous research efforts mainly focus on AMD categorization with a single-modal input, let it be a color fundus photograph (CFP) or an OCT B-scan image. By contrast, we consider AMD categorization given a multi-modal input, a direction that is clinically meaningful yet mostly unexplored. Contrary to the prior art that takes a traditional approach of feature extraction plus classifier training that cannot be jointly optimized, we opt for end-to-end multi-modal Convolutional Neural Networks (MM-CNN). Our MM-CNN is instantiated by a two-stream CNN, with spatially-invariant fusion to combine information from the CFP and OCT streams. In order to visually interpret the contribution of the individual modalities to the final prediction, we extend the class activation mapping (CAM) technique to the multi-modal scenario. For effective training of MM-CNN, we develop two data augmentation methods. One is GAN-based CFP/OCT image synthesis, with our novel use of CAMs as conditional input of a high-resolution image-to-image translation GAN. The other method is Loose Pairing, which pairs a CFP image and an OCT image on the basis of their classes instead of eye identities. Experiments on a clinical dataset consisting of 1,094 CFP images and 1,289 OCT images acquired from 1,093 distinct eyes show that the proposed solution obtains better F1 and Accuracy than multiple baselines for multi-modal AMD categorization. Code and data are available at https://github.com/li-xirong/mmc-amd.**

***Index Terms*—Multi-modal AMD categorization, multi-modal fundus imaging, two-stream CNN, data augmentation, loose pairing training, image synthesis**

Weisen Wang and Xirong Li are with the MoE Key Lab of Data Engineering and Knowledge Engineering, Renmin University of China and AIMC Lab, School of Information, Renmin University of China, Beijing 100872, China (e-mail: {weisen,xirong}@ruc.edu.cn)

Jianchun Zhao and Dayong Ding are with Vistel AI Lab, Visionary Intelligence Ltd., Beijing 100872, China (e-mail: {jianchun.zhao,dayong.ding}@vistel.cn)

Zhiyan Xu, Weihong Yu and Youxin Chen are with the Key Lab of Ocular Fundus Disease, Chinese Academy of Medical Sciences and Dept. of Ophthalmology, Peking Union Medical College Hospital, Beijing 100730, China (e-mail: xuzythu@163.com, 536273640@qq.com, chenyouxinpumch@163.com)

I. INTRODUCTION

AGE-related Macular Degeneration (AMD) is a common retinal disease among people over the age of 50. Without timely and proper treatment, patients with AMD may suffer from impaired vision or even blindness [1]. Subject to the presence of choroidal neovascularization in the retina, the medical profession classifies the disease into *dry* AMD (atrophic AMD and early stage AMD) and *wet* AMD (neovascular AMD) [2]. Moreover, when neovascularization occurs below the retinal pigment epithelium, such a type of wet AMD is known as Polypoidal Choroidal Vasculopathy (PCV) [3]–[5], typically characterized by polypoidal or aneurysmal dilations [6]. For the ease of description, we refer to wet AMD excluding PCV as wetAMD hereafter. Due to different treatments of dryAMD, PCV and wetAMD [5], we need to not only distinguish AMD from normal, but also perform a fine-grained classification of the three AMD subcategories. To perform an AMD examination for a specific eye, an ophthalmologist commonly uses two noninvasive fundus imaging techniques, *i.e.,* color fundus photography (CFP) and optical coherence tomography (OCT). Given the retina, a CFP image shows its en face, while an OCT B-scan image captures its longitudinal section. Examples of normal CFP and OCT images and those with specific AMD-related pathologies are shown in Fig. 1. A recent study [7] reports that when provided with multi-modal information (CFP / OCT images and clinical notes), experts make fewer diagnostic errors for referral recommendation. Unsurprisingly, due to the shortage of experienced ophthalmologists, there is a growing amount of research efforts towards automated AMD screening based on either color fundus images [8]–[10], OCT images [11]–[13] or both [14].

Existing works on AMD categorization mainly exploit one modality, let it be CFP images of the posterior pole [8]–[10], [15] or OCT images of the macular [11]–[13], [16], [17]. In [9], for example, Burlina *et al.* deal with CFP, training AlexNet [18], a deep convolutional neural network (CNN), from scratch. An OCT-based method is developed by Lee *et al.* [11], where a VGGNet [19] is trained to predict the presence of AMD in OCT images.

Given that CFP and OCT images reflect the state of the retina in distinct aspects, and thus complement each other, a joint utilization of both modalities appears to be natural. To our surprise, however, few attempt is made in this direction.
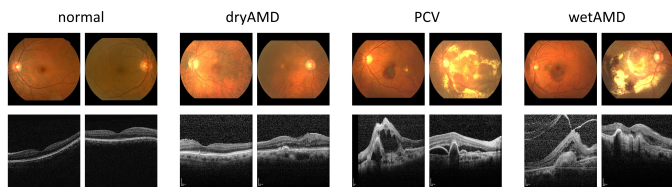
Fig. 1. Color fundus images (first row) and OCT B-scan images (second row) of normal eyes and eyes with specific diseases. Color fundus images show en face of the fundus, while OCT B-scans visualize cross-sectional information of the retina and choroid. Images per group are randomly selected from our dataset based on their classes.

To the best of our knowledge, the first work on multi-modal AMD categorization is by Yoo *et al.* [14]. The authors leverage a pre-trained VGGNet to extract features from CFP and OCT images. The features are concatenated and classified by a random forest classifier. Their preliminary experiment indicates that the multi-modal method outperforms its single-modal counterpart. Nevertheless, given the rapid progress of deep learning for visual recognition, the VGGNet feature and the random forest classifier are both suboptimal. We thus see multiple crucial questions left unanswered. First, suppose the single-modal baseline is re-implemented based on a state-of-the-art CNN trained end-to-end for the task, will the current multi-modal method [14] remain more effective? If the answer turns out to be "no", further questions naturally follow as what deep learning network is suited for the task and how to train it effectively? Training a multi-modal network is nontrivial, as paired multi-modal training instances are far fewer than single-modal training instances. Therefore, one cannot take for granted that a multi-modal CNN will be better than its single-modal competitor. Last but not least, how to visually interpret the contribution of each modality to the final prediction?

Towards answering the above questions, we make contributions as follows:

• We propose a novel end-to-end solution for multi-modal AMD categorization, the first of its kind to the best of our knowledge. The backbone of our solution is a two-stream CNN, see Fig. 2(a). While such an architecture has been applied for video action recognition [20], the fusion layer needs to be re-considered for the new task, not only for effectively fusing information from the individual modalities, but also for visualizing their contributions to the final prediction. To that end, we propose to use spatially-invariant fusion (SI-Fusion), which simply performs global average pooling on modal-specific feature maps followed by vector concatenation. SI-Fusion is not new by itself. Its novelty is due to its suitability for multi-modal AMD categorization. Compared to EarlyFusion which concatenates CFP and OCT together as a whole, LateFusion which averages predictions of CFP- / OCT- CNNs, and multi-head self-attention based fusion [21], SI-Fusion strikes a proper balance between a network's learning capacity (for exploiting the complementarity between CFP and OCT) and its training complexity (for being more data-efficient). SI-Fusion also allows us to extend the CAM visualization technique [22] for the first time to the multi-modal scenario for visual interpretation.

• Due to the natural shortage of multi-modal training data, learning two-stream CNN with conventional data augmentation methods is insufficient. For multi-modal training data augmentation, we introduce two methods, *i.e., CAM-conditioned Image Synthesis* and *Loose Pairing*. The first method, working at the image level, is to synthesize CFP / OCT images by a high-resolution generative adversarial network (GAN) [23] re-purposed in the new context, see Fig. 2(b). The second method pairs CFP and OCT images w.r.t their classes instead of eye identities. Consequently, we develop a training pipeline, as illustrated in Fig. 2(c), to effectively learn two-stream CNN from the augmented data.

• Experiments on a clinical dataset, with 1,094 CFP and 1,289 OCT images acquired from 1,093 eyes, verify the effectiveness of our solution. It surpasses the prior art [14] clearly, with 0.914 *vs* 0.792 in F1-score and 0.863 *vs* 0.690 in Accuracy, for four-class AMD categorization. Code, data and model links are available at github[1].

To sum up, the novelty of this paper is two-fold, *i.e.,* clinically, the end-to-end multi-modal approach to AMD detection and technically, the CAM-based data augmentation with loose pairing. A preliminary version of this work was published at MICCAI'19 [24]. We improve over the conference version in three aspects, *i.e.,* problem setup, method and evaluation. First, PCV is considered, making our setup more clinically meaningful, yet more challenging due to its high resemblance to other wet AMD. Second, we enhance our data augmentation process with GAN-based CFP / OCT image synthesis (Sec. III-B.1). For an effective use of both real and synthetic training data, a new training pipeline is developed (Sec. III-C). The above changes in problem setup and method naturally lead to a substantial extension of our evaluation.

The rest of the paper is organized as follows. We discuss related work in Sec. II, followed by the proposed solution in Sec. III, experiments in Sec. IV, and conclusions in Sec. V.

II. RELATED WORK

As this paper contributes to automated AMD categorization in terms of network designs and data augmentation methods, we review prior art for AMD categorization in Sec. II-A and data augmentation in Sec. II-B. Note that we focus on deep learning approaches. For tackling AMD categorization by non-deep learning approaches, we refer to [25].

A. Deep Learning based AMD categorization

Existing works are largely based on a single modality, either CFP images [8]–[10], [15] or OCT images [11]–[13], [16], see Table I. In [8], [15], for instance, Burlina *et al.* tackle AMD categorization via a classical image classification approach, *i.e.,* feature extraction plus classifier training. In particular, the authors employ a CNN pretrained on ImageNet to extract visual features from CFP images and then train a linear SVM classifier. In a follow-up work [9], Burlina *et al.* find that an end-to-end trained CNN performs better than the classical approach. Ensemble learning is exploited by Grassmann *et al.* to
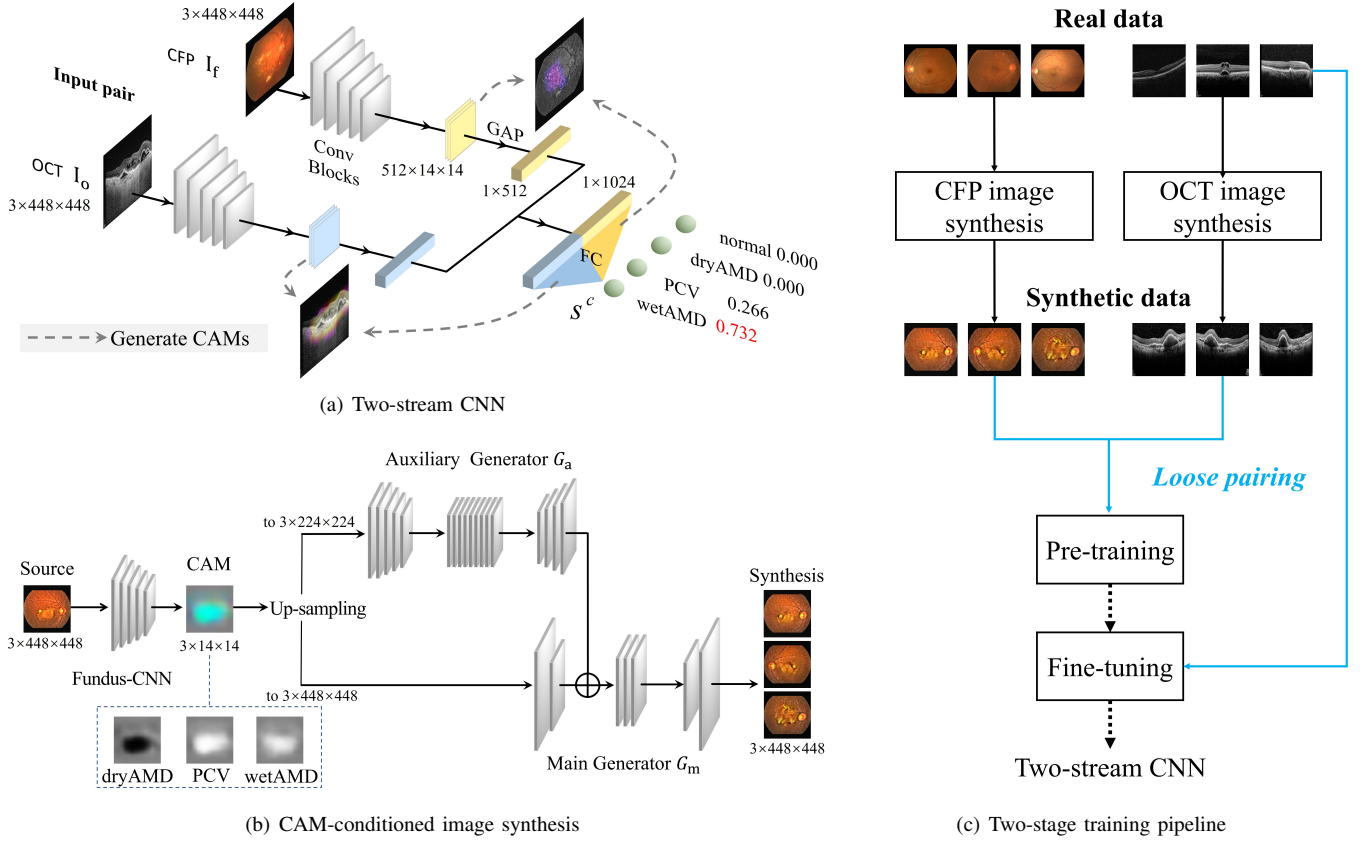
---

[1] https://github.com/li-xirong/mmc-amd

(a) Two-stream CNN

(b) CAM-conditioned image synthesis

(c) Two-stage training pipeline

Fig. 2. **Proposed end-to-end solution for multi-modal AMD categorization**. Given a pair of CFP and OCT images from a specific eye, our two-stream CNN makes a four-class prediction concerning the probability of the eye being *normal*, *dryAMD*, *PCV* and *wetAMD*, respectively. Extending class activation mapping (CAM) to the multi-modal scenario allows us to visualize contributions of each modality to final predictions. For effective training, we introduce two data augmentation methods. The first method is to synthesize CFP / OCT images by pix2pixHD, a high-resolution image-to-image translation network re-purposed in the new context. Given a $3 \times 448 \times 448$ source image, we pre-train the corresponding single-modal CNN to produce CAMs *w.r.t* each AMD class. The CAMs are stacked to form an three-channel image $[CAM^{dry}; CAM^{pcv}; CAM^{wet}]$ of $3 \times 14 \times 14$, which is then fed into pix2pixHD for image synthesis. A fully convolutional network G_a is used as an auxiliary generator to generate a $3 \times 224 \times 224$ image. With the help of G_a, another fully convolutional network G_m is then used as a main generator to generate a double-sized image. Manipulating the CAMs results in multiple synthesized images. The second method pairs CFP and OCT images based their classes instead of eye identities. A two-stage training is performed, where the two-stream CNN is first pre-trained on synthetic and loosely paired CFP and OCT images, and then fine-tuned on real and loosely paired multi-modal data.

TABLE I

STATE-OF-THE-ART FOR AUTOMATED AMD CATEGORIZATION. THIS PAPER PROPOSES TWO-STREAM CNN FOR MULTI-MODAL AMD CATEGORIZATION, AND DEVELOPS MULTI-MODAL DATA AUGMENTATION FOR EFFECTIVE TRAINING.

Modality	Paper	Categorization model	Data augmentation
CFP	Burlina *et al.* [8], [15]	OverFeat feature + linear SVM classifier	unmentioned
	Burlina *et al.* [9]	AlexNet trained from scratch	unmentioned
	Grassmann *et al.* [10]	Ensemble of AlexNet, GoogLeNet, Inception-V3, VGG11, ResNet-101, and Inception-ResNet-V2	◦ crop, flip, rotation
OCT	Lee *et al.* [11]	VGG16 trained from scratch	unmentioned
	Treder *et al.* [13]	Fine-tuned Inception-V3	◦ flip
	Karri *et al.* [12]	Fine-tuned GoogLeNet	unmentioned
	Russakoff *et al.* [16]	A customized CNN	◦ rotation, additive noise
CFP + OCT	Yoo *et al.* [14]	VGG19 features + random forest classifier	◦ translation, rotation, brightness change, additive noise
	This work	Two-Stream CNN	◦ crop, flip, rotation, changes in brightness / saturation / contrast ● *CAM-conditioned image synthesis* ● *Loose pairing*

combine multiple CNNs [10]. As for OCT-based methods, Lee *et al.* [11] train a VGG16 model from scratch to classify OCT images either as *normal* or as *AMD*. By contrast, Karri *et al.* [12] and Treder *et al.* [13] train their CNNs by fine tuning. Different from the above works, we exploit both CFP and OCT images as multi-modal input.

Few attempt has been made for multi-modal AMD categorization. To the best of our knowledge, Yoo *et al.* [14] make the first effort in this direction, where the authors perform three-class categorization, *i.e.,* normal, dryAMD and wetAMD, given CFP and OCT images. Similar to [8], [15] in the single-modal context, the authors follow the classical

approach, employing a pre-trained VGG19 model to extract visual features from both modalities. The features are then concatenated and used as input of a random forest classifier. Different from [14], our model is trained end-to-end.

B. Data Augmentation for CNN Training

Ever since the great success of AlexNet [18], data augmentation using low-level image processing techniques such as crop, flip, rotation and changes in brightness, saturation and contrast has been a rule of thumb for CNN training. Not surprisingly, we observe from Table I that data augmentation used by existing works for AMD categorization all follows this convention. We notice a very recent attempt by Burlina *et al.* [26] to train a binary AMD classifier using GAN-generated examples. In particular, the authors train two PGGANs [27], one for generating normal CFP images and the other for synthesizing CFP images with AMD. Their motivation, *i.e.,* to substitute GAN-generated data for manually labeled data, differs from ours. Moreover, according to their study, the classifier trained on synthetic data has lower performance than its counterpart trained on real data (0.8292 *vs* 0.9112 in Accuracy and 0.9235 *vs* 0.9706 in AUC). The result suggests that synthetic data alone is insufficient.

For synthesizing high-resolution training images, there is an increasing interest in leveraging image-to-image GANs [28]–[32]. For detection of exudates in CFP images, Zheng *et al.* [28] expand their training data by developing a pix2pix [33] based network that converts exudate segmentation maps into CFP images. As illustrated in Fig. 3(a), a pix2pix model converts an input image, *e.g.,* an exudate segmentation map in [28], to a synthetic CFP image by a fully convolutional

(a) pix2pix

(b) pix2pixHD

Fig. 3. **Conceptual illustrations of pix2pix and pix2pixHD for image-to-image translation**. A pix2pix model converts an input image to a synthetic image using a generator network G, the parameters of which are optimized during training together with a discriminator network D. Note that ⊕ indicates channel-wise concatenation. A pix2pixHD model generates higher-resolution images in a coarse-to-fine manner. An auxiliary generator G_a and an auxiliary discriminator D_a are first trained for lower-resolution synthesis. Then, a main generator G_m and a main discriminator D_m are used to double the resolution.

network known as a generator. The generator is trained to fool another convolutional network known as a discriminator, which is responsible for discriminating between real and fake samples. For tumor segmentation in brain MRI images, Shin *et al.* [29] train pix2pix to synthesize MRI images from tumor and tissue segmentation maps. For lesion detection in chest X-ray images, Xing *et al.* [30] also adopt pix2pix, where the input of pix2pix is obtained by masking out all regions of abnormal images except for lesion areas. In the context of diabetic retinopathy grading, Zhou *et al.* [31] synthesize CFP images by a pix2pixHD [23] based network, which translates vessel and lesion masks into the images. As shown in Fig. 3(b), the major improvement of pix2pixHD over pix2pix is a joint use of two generators and two discriminators to generate higher-resolution images in a coarse-to-fine manner.

Our proposed method uses pix2pixHD. However, in contrast to existing works where the input is manually annotated [28], [30], or automatically produced yet requiring manual annotation for training underlying semantic segmentation models [31], or both [29], our input, namely Class Activation Maps (CAMs), is automated extracted by image classifiers, with no need of any lesion annotation.

III. PROPOSED SOLUTION

Given a CFP image I_f and an OCT image I_o acquired from a specific eye, our goal is to categorize the condition of the eye into one of the four classes, *i.e.,* {*normal*, *dryAMD*, *PCV*, *wetAMD*}. We propose a multi-modal CNN (MM-CNN) to predict a specific class c based on the paired input $\{I_f, I_o\}$. MM-CNN is described in Sec. III-A, followed by our data augmentation methods in Sec. III-B. The training strategy for MM-CNN is detailed in Sec. III-C.

A. MM-CNN for AMD Categorization

1) Network Architecture: We instantiate MM-CNN with a two-stream CNN. As shown in Fig. 2(a), the network contains two symmetric branches, which process I_f and I_o in parallel. This type of network architecture bears some similarity to the two-stream network investigated in the context of human action recognition in videos [20]. The main distinction lies in at which layer multi-modal fusion is conducted. Recall that a single-modal CNN extracts information from a given image by producing arrays of 2-D feature maps layer-by-layer. While being continuously down-sized, these feature maps preserve the spatial information of the input image to some extent. As for video data, its RGB stream and Optical Flow stream are spatially correlated. It is this reason that fusion by combining feature maps from the distinct streams is preferred for video action recognition [20]. Contrary to video data, the CFP image and the OCT image lack such spatial correlation, as noted in Sec. I. We therefore opt for spatially invariant fusion, achieved by performing the fusion after the Global Average Pooling (GAP) layer, which compresses each feature map to a scalar and thus removes the spatial information.

For each branch of the two-stream CNN, we adopt ResNet-18 as their backbone. While other CNNs can also be used, our consideration of choosing ResNet-18 is as follows. This
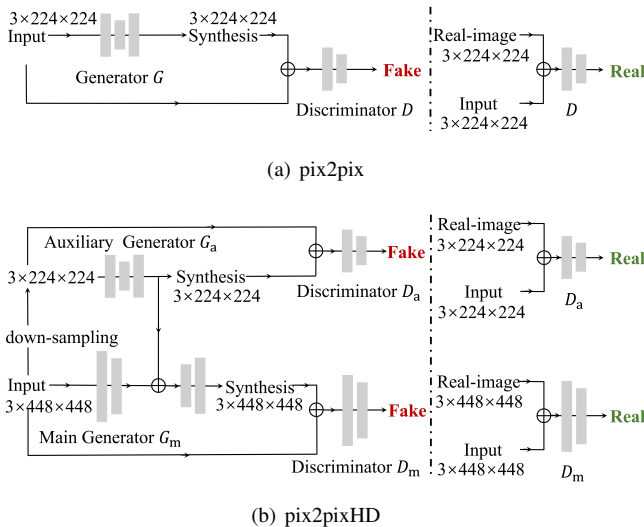
network has relatively fewer parameters and thus needs less data for training, meanwhile its effectiveness has been justified in other fundus image recognition tasks [34], [35]. Note that for the OCT image, we convert each of its pixels from grayscale to RGB by duplicating the intensity for each RGB component. The CFP and OCT images are normalized into the range $[-1, 1]$ with mean and std of 0.5, using a default PyTorch normalization operation[2]. As such, the same architecture and initialization are applied to both branches.

Spatially-invariant fusion. For the CFP branch, the last convolutional block of ResNet-18 produces an array of 512 feature maps, denoted as $\mathbf{F}_f = \{F_{f,1}, \ldots, F_{f,512}\}$. Each feature map has a size of $m \times m$, with m subject to the size of the input image. Instead of the commonly used input size of $3 \times 224 \times 224$, we choose a larger size of $3 \times 448 \times 448$, which shows better performance in our preliminary experiment[3]. Accordingly, the value of m is 14. For a specific feature map $F_{f,i}$, we use $F_{f,i}(x, y)$ to access the feature value at a specific position (x, y). In a similar style, the feature maps from the OCT branch are defined as $\mathbf{F}_o = \{F_{o,1}, \ldots, F_{o,512}\}$.

Our spatially-invariant fusion layer is simply implemented as follows. For $\mathbf{F}_f$ and $\mathbf{F}_o$, they are fed into a GAP layer in parallel, resulting in two 512-d feature vectors, indicated by $\bar{F}_f = (\bar{F}_{f,1}, \ldots, \bar{F}_{f,512})$ and $\bar{F}_o = (\bar{F}_{o,1}, \ldots, \bar{F}_{o,512})$, respectively. A concatenation operation merges $\bar{F}_f$ and $\bar{F}_o$ into a new 1024-d vector with semantic information from both modalities combined. For the four-class classification task, by feeding the merged vector into a fully connected layer of shape 1024×4, we obtain s^c as a decision score for a specific class:

$$s^c = \underbrace{\sum_{i=1}^{512} w_{f,i}^c \cdot \bar{F}_{f,i}}_{s_f^c} + \underbrace{\sum_{i=1}^{512} w_{o,i}^c \cdot \bar{F}_{o,i}}_{s_o^c}, \qquad (1)$$

where $\{w_{f,1}^c, \ldots, w_{f,512}^c\}$ and $\{w_{o,1}^c, \ldots, w_{o,512}^c\}$ denote class-specific parameters of the fully connected layer, while s_f^c and s_o^c indicate decision scores contributed by the CFP branch and the OCT branch, respectively. Classification is realized by selecting the class maximizing s^c.

2) Multi-Modal Class Activation Mapping for Visual Interpretation: We note from (1) that the class-specific score s^c is contributed by the two modalities in an additive manner. In order to visualize their contributions, we employ Class Activation Mapping (CAM) [22]. As a response-based visualization technique, CAM projects the classification score back to a $m \times m$ heatmap, and thus reveals which part of the input image contributes the most. To make the paper more self-contained, we first depict CAM developed initially for a single-modal CNN, with CFP images as a showcase. We then extend CAM for the multi-modal scenario.

To simplify our notation, we re-purpose s_f^c from (1) to indicate the decision score of class c made by a single-modal

CNN with respect to a test CFP image I_f. Recalling that $\bar{F}_{f,i}$ is equivalent to $\frac{1}{m^2} \sum_{x,y} F_{f,i}(x, y)$, we express s_f^c as

$$\begin{aligned} s_f^c &= \sum_{i=1}^{512} w_{f,i}^c \cdot \bar{F}_{f,i} \\ &= \sum_{i=1}^{512} w_{f,i}^c \cdot \left(\frac{1}{m^2} \sum_{x,y} \cdot F_{f,i}(x, y)\right) \\ &= \sum_{x,y} \frac{1}{m^2} \sum_{i=1}^{512} w_{f,i}^c \cdot F_{f,i}(x, y). \end{aligned} \qquad (2)$$

By defining CAM_f^c as a $m \times m$ feature map computed by

$$CAM_f^c(x, y) = \frac{1}{m^2} \sum_{i=1}^{512} w_{f,i}^c \cdot F_{f,i}(x, y), \qquad (3)$$

s_f^c is equivalently the sum of CAM_f^c, *i.e.,*

$$s_f^c = \sum_{x,y} CAM_f^c(x, y). \qquad (4)$$

We see from (4) that the classification score is spatially distributed to each position the CAM map. Consequently, by overlaying an input image with its corresponding (up-sampled) CAM map, salient regions are highlighted, see Fig. 4(b).

We derive a multi-modal version of CAM for class c as

$$\begin{cases} CAM_f^c(x, y) = \frac{1}{m^2} \sum_{i=1}^{512} w_{f,i}^c \cdot F_{f,i}(x, y), \\ \\ CAM_o^c(x, y) = \frac{1}{m^2} \sum_{i=1}^{512} w_{o,i}^c \cdot F_{o,i}(x, y). \end{cases} \qquad (5)$$

Putting (1), (2) and (5) together, the score s^c for class c can be reformulated using the CFP and OCT CAMs, namely

$$s^c = \sum_{x,y} CAM_f^c(x, y) + \sum_{x,y} CAM_o^c(x, y). \qquad (6)$$

The decision score is now distributed via $CAM_f^c(x, y)$ and $CAM_o^c(x, y)$ to each position on the CFP and OCT images, in a down-sampled size of $m \times m$. This enables us to visualize the contribution of each modality, see Fig. 4.

### B. Multi-Modal Data Augmentation

For effectively increasing the amount of multi-modal training instances, we propose two data augmentation methods, one for image-level (Sec. III-B.1) and the other for pair-level (Sec. III-B.2). Both methods are applied only on the training data.

1) CAM-conditioned Image Synthesis for Image-level Data Augmentation: In order to generate CFP / OCT images for a specific AMD class c, we re-purpose CAMs as input of pix2pixHD [23], a powerful network for image-to-image translation. The proposed CAM-conditioned image synthesis method is illustrated in Fig. 2(b).

From a source image to CAMs. Given a source image I from a specific AMD class c, we first use a single-modal CNN, which we have pre-trained, to produce CAMs *w.r.t.* the three AMD classes. We ignore the normal class, as its CAMs do not reflect abnormalities by definition. As described in Sec. III-A.2, for an input image of $3 \times 448 \times 448$, each CAM computed by (3) has a size of 14×14. The CAMs are stacked to form a three-channel image $I_{cam} = [CAM^{dry}; CAM^{pcv}; CAM^{wet}]$. With the CAMs treated as red, green and blue channels in an RGB image, I_{cam} is colorized naturally. As manifested by the second column of Fig. 5, dominant colors are better observed in CAMs of OCT images, suggesting this modality

---

[2]torchvision.transforms.Normalize(mean=(0.5, 0.5, 0.5), std=(0.5, 0.5, 0.5))
[3]Compared with the input size of $3 \times 224 \times 224$, the larger input size of $3 \times 448 \times 448$ leads to loss in accuracy (from 0.721 to 0.717) for CFP-CNN, yet improvements for both OCT-CNN (from 0.811 to 0.818) and MM-CNN (from 0.795 to 0.804).
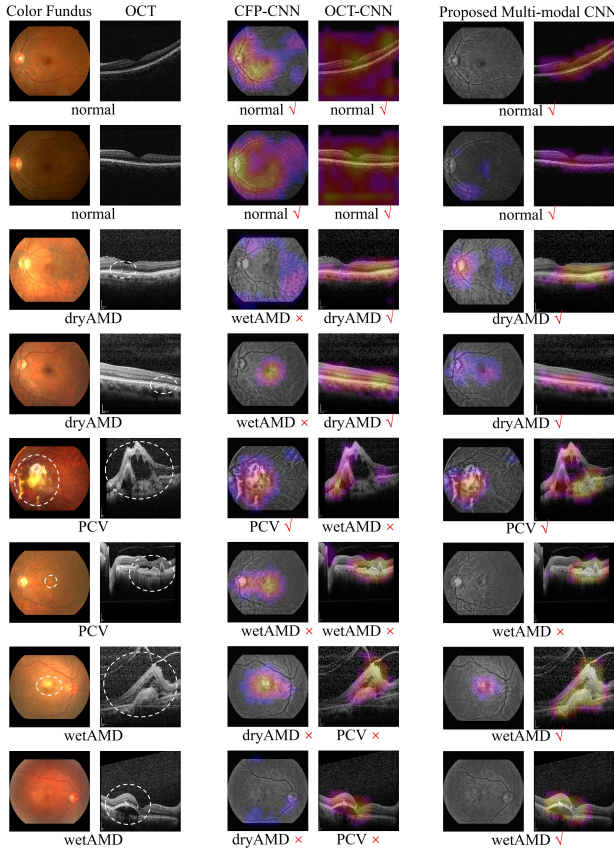
(a) Multi-modal input   (b) Single-modal CAMs   (c) Multi-modal CAMs

Fig. 4. **CAM-based visualization of which part of input images contribute to the final predictions made by single-modal / multi-modal CNNs**. Input images are shown in (a), with AMD-related regions marked out by dotted ellipses. In (a), labels under each pair of images are ground-truth, while labels in (b) and (c) are predictions made by corresponding models. Correct and incorrect predictions are marked out by ✓ and ×, respectively. For a better visualization, the images are overlaid with single-modal CAMs (b) and multi-modal CAMs (c), whilst the CFP images are converted to gray scale.

is more suited for AMD categorization. Next, we describe how to convert I_{cam} to a positive instance *w.r.t.* the given class.

From CAMs to synthesized images. Different from a classical GAN that has one generator and one discriminator, pix2pixHD consists of a main generator G_m and an auxiliary generator G_a that produce images at two different resolutions, which are $3 \times 448 \times 448$ and $3 \times 224 \times 224$ in this work, see Fig. 3(b). Accordingly, there are two fully convolutional network based discriminators D_m and D_a responsible for the two resolutions. Note that we do not draw the discriminators in Fig. 2(b) as they are used only in the training stage. A new image is generated by G_m with assistance from G_a. In our context, I_{cam} is first up-sampled by bicubic interpolation to two scales, *i.e.*, $3 \times 224 \times 224$ and $3 \times 448 \times 448$. The two enlarged CAMs are separately fed to G_a and G_m. As illustrated in Fig. 3(b), information extracted by G_a is injected into G_m by combing their feature maps by the matrix addition. For each modality, we train a pix2pixHD model, see Section IV-A for details of the training procedure.

We observe that the highest response region in the CAM

tends to indicate abnormal areas in the source image. Hence, in order to synthesize diverse yet meaningful images, we manipulate I_{cam} by moving around its highest response region. Specifically, we localize such a region by a sliding-window approach, with the window size empirically set to be 5×5. The window that has the maximal response with respect to the given class is selected. Then, the region is swapped with a randomly chosen region of the same size. In this manner we construct for the same source image a number of distinct CAMs, which we term *manipulated CAMs*, to distinguish them from the original CAM. Fig. 5 showcases some source images, their original and manipulated CAMs, and synthesized images.
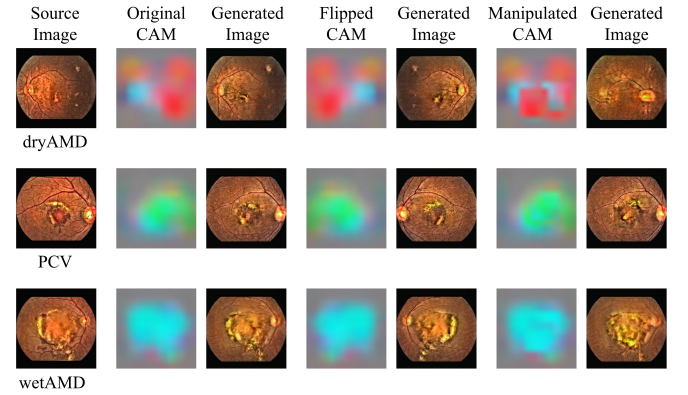
(a) Real CFP images, CAMs, and synthesized CFP images
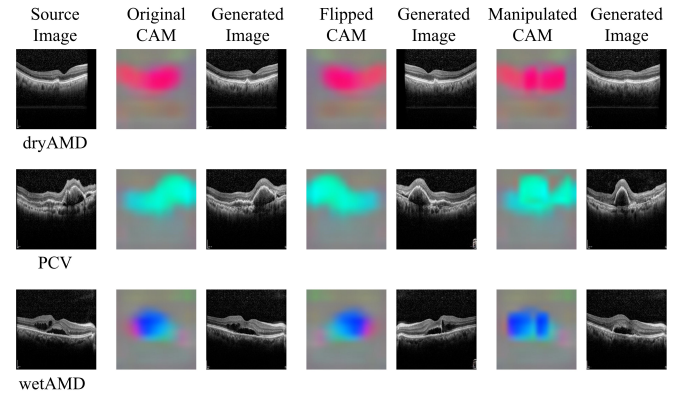
(b) Real OCT images, CAMs, and synthesized OCT images

Fig. 5. **Illustrations of images generated by the CAM-conditioned image synthesis** for (a) CFP and (b) OCT images. For better visualization, for each source image we stack its three CAMs, *i.e.*, CAM^{dry}, CAM^{pcv} and CAM^{wet} to form an RGB image. The fact that dominant colors are better observed in OCT images suggests this modality is more suited for AMD categorization. A flipped CAM indicates horizontal flip of an original CAM, while a manipulated CAM is obtained by moving the highest response region. Best viewed in digit.

When dealing with a novel image, it is possible that the CAM may mis-localize some AMD lesions and thus result in a wrong CFP / OCT image. To ensure the credibility of CAM visualization, our tactic here is to re-use the training images that are already been seen by CFP-CNN / OCT-CNN in their training stage. Such an "overfitting" produces good-quality CAM in general. In fact, we observe that feeding a CAM to pix2pixHD reconstructs its source image well. Manipulating

the CAM by swapping is mainly for diversifying the synthetic data. Moreover, the synthetic data will be used exclusively for pre-training MM-CNN. We expect to reduce the negative effect of wrongly generated images by fine-tuning.

2) Loose Pairing for Pair-Level Data Augmentation: To obtain a multi-modal training example, a natural strategy is to strictly select a CFP image and an OCT image based on their eye identities. Suppose we have two PCV eyes, eye^a and eye^b, in the training set. Each eye is associated with several images, *i.e.*, $set^a = \{f^a, o_1^a, o_2^a, o_3^a\}$ and $set^b = \{f^b, o_1^b, o_2^b\}$, where f and o stand for CFP and OCT images, respectively. There are five strict pairs only, *i.e.*, $\{(f^a, o_i^a), (f^b, o_j^b)\}$, $i = 1, 2, 3$ and $j = 1, 2$. In order to increase the number of multi-modal instances for training, we propose to construct input pairs based on *labels* instead of eyes. A color fundus image can be paired with an OCT image as long as their class labels are identical. We coin this method *Loose Pairing*. Consequently, we obtain five loose pairs from set^a and set^b and ten multi-modal instances in total. With loose pairing, the size of the training set is substantially increased.

C. A Two-stage Training Strategy for Multi-Modal CNN

Given a relatively small real-world dataset, the joint use of the proposed CAM-conditioned image synthesis and loose pairing enables us to construct a large amount of multi-modal training instances. However, directly combining the real dataset and the generated dataset is problematic, as the former will be easily outnumbered. In that regard, we train the MM-CNN in two stages, *i.e.,* pre-training and fine-tuning. In the first stage, a CFP image and an OCT image are loosely paired, with at least one of them sampled from the generated dataset. Given the MM-CNN trained on these fake pairs, we perform fine-tuning in the second stage, using the real dataset. Despite the simplicity, such a two-stage training strategy is found to be effective, as we will shortly show in Sec. IV-C.

For the backbone network, we start with ResNet-18 pre-trained on ImageNet [36]. Instead of using the original input size is $3 \times 224 \times 224$, we use a larger input size of $3 \times 448 \times 448$. The kernel size of the GAP layer is accordingly adjusted, from 7×7 to 14×14, to ensure the dimensionality of the last feature vector is invariant to this change. For both single-modal and multi-modal CNNs, we use cross-entropy, a common loss function for multi-class classification. SGD is used as the optimizer, with momentum of 0.9, weight decay of 1e-4 and batch size of 8. Per training process, a model is selected based on its performance on a held-out validation set.

Contrast-limited adaptive histogram equalization (CLAHE) [37] is used to enhance CFP images. For noise reduction in OCT images, median filtering with a 3×3 kernel is applied. Note that the proposed data augmentation methods are not meant for replacing conventional data augmentation strategies. In fact, the conventional strategies are needed to train better single-modal CNNs for CAM-based image synthesis. The proposed strategies and the conventional strategies shall be used in combination. In advance to the proposed data augmentation, we perform low-level common data augmentation operations including random crop, flip, rotation, and random changes in contrast, saturation and brightness on training images.

IV. EXPERIMENTS

A. Experimental Setup

Multi-modal dataset. We collected our experimental data from the outpatient clinic of the Dept. of Ophthalmology, Peking Union Medical College Hospital. The dataset initially contains 1,094 CFP images from 1,093 eyes of 829 subjects, acquired by a Topcon fundus camera. For 817 eyes, they are associated with one to five OCT images. The OCT images are central B-scans acquired by a Topcon OCT camera and a Heidelberg OCT camera and manually selected by technicians. We chose these cameras as they were the most frequently used in our outpatient clinic, allowing us to collect a decent amount of samples for this research. The condition of each eye is jointly assessed by two ophthalmologists based on its CFP, OCT, Fluorescein angiography (FA) or indocyanine green angiography (ICGA), when applicable. Accordingly, each eye is categorized into *normal*, *dryAMD*, *PCV* or *wetAMD*. Images associated with a specific eye gets the same label.

Eventually, we obtain an expert-labeled multi-modal dataset of 1,094 CFP and 1,289 OCT images. All these data are approved by the IRB of Peking Union Medical College Hospital. We obey the principles of the Declaration of Helsinki.

TABLE II

DATA USED IN OUR EXPERIMENTS. DATA SPLIT IS MADE BASED ON EYES. IN PARENTHESES ARE NUMBER OF EYES PER CLASS IN EACH SPLIT. AS THE NUMBER OF OCT IMAGES VARIES PER EYE, OCT IMAGES OF THE FOUR CLASSES ARE MORE UNBALANCED.

Class	Training set		Validation set		Test set	
	CFP	*OCT*	*CFP*	*OCT*	*CFP*	*OCT*
normal	155 (155)	156 (155)	20 (20)	20 (20)	20 (20)	20 (20)
dryAMD	67 (67)	33 (22)	20 (20)	35 (20)	20 (20)	38 (20)
PCV	259 (259)	289 (156)	20 (20)	44 (20)	20 (20)	47 (20)
wetAMD	453 (452)	531 (325)	20 (20)	38 (20)	20 (20)	38 (20)

In order to construct a class-balanced test set, for each class we randomly sample 20 eyes from the eyes that have both CFP and OCT images. This setting enables us to compare an multi-modal input with its single-modal counterpart. The setting also allows us to make a head-to-head comparison between the two single modalities, namely CFP versus OCT. In a similar manner, we build a multi-modal validation set for model selection. We use the remainder for training. Data split is based on eye identities, so images from a specific eye do not appear in more than one subset.

Note that the relatively small amount of training samples for dryAMD will adversely affect the stability of CFP image based models. Its impact on OCT-based models is minor, as dryAMD-related visual patterns such as drusen are easily recognizable in OCT images. The effect of data shuffle on the model performance is provided in the Appendix.

Performance metrics. For class-specific evaluation, we report *Sensitivity*, *Specificity* and the *F1* score, which is the harmonic mean between Sensitivity and Specificity. For overall performance evaluation, we report F1 averaged over the four classes. Besides, we report *Accuracy*, the ratio of correctly categorized instances, which are CFP or OCT images for single-modal CNNs and CFP-OCT pairs for MM-CNNs. Due
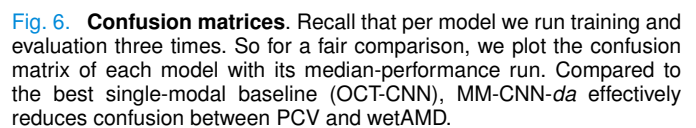
to randomness in SGD based training, the performance of a model obtained in a specific training round may vary. To reduce such a random effect, for each model we run training and evaluation three times, and report averaged score per metric. That is, for a given model, its reported Sensitive, Specificity and F1 scores per class are all averaged based on the three runs. Hence, the reported F1 is not the harmonic mean between the reported Sensitivity and Specificity.

Implementation. For fair comparisons, all CNN models assessed in our experiments use ResNet-18, pre-trained on ImageNet, as their backbones. For all the models, common data augmentation is applied for training. A model trained with the proposed data augmentation methods, when applicable, will be postfixed with -*da*. Per abnormal image in the original training data, we synthesize three new images, yielding 201 dryAMD / 777 PCV / 1,359 wetAMD for the CFP modality and 99 dryAMD / 867 PCV / 1,593 wetAMD for the OCT modality. Loose pairing these synthetic images results in nearly 2.9M multi-modal samples for pre-training. As for finetuning, loose pairing is conducted on the real set, resulting in 317.6K multi-modal samples. All deep models are implemented using PyTorch [38] on a Ubuntu 16.04 server with an NVIDIA® GeForce® RTX 2080 Ti GPU.

For the single-modal baselines, two ResNet-18 models are separately trained on the CFP and OCT images, referred to as CFP-CNN and OCT-CNN. For both models, their input dimension is $3 \times 448 \times 448$. They are trained in a similar manner as that of MM-CNN described in Section III-C, *i.e.,* SGD as the optimizer, momentum of 0.9, weight decay of 1e-4 and mini-batch size of 8. To deal with the imbalanced training samples, per mini-batch we select samples of the four classes in equal proportions.

Following [23], we train pix2pixHD in a coarse-to-fine manner. Specifically, for the first p epochs, only the auxiliary generator G_a and the auxiliary discriminate D_a are trained to generate images of $3 \times 224 \times 224$. For the next q epochs, the main and auxiliary networks are jointly trained to synthesize images of $3 \times 448 \times 448$. Horizontal flip with probability of 0.5 and random crop are utilized as data augmentation. Training a pix2pixHD model on our dataset (batch size 1) typically requires 150 epochs, with $p = 100$ and $q = 50$. This amounts to a training time of around 8 hours, exploiting about 120k augmented images in total. We did not observe model collapse. Artifacts, *e.g.,* synthesized CFP images with two optic discs, are occasionally observed, see Fig. 8.

B. Experiment 1. Multi-modal versus Single-modal

Baselines. Our single-modal baselines are CFP-CNN and OCT-CNN. To study if the synthetic images help the single-modal models, we train them using the two-stage strategy described in Sec. III-C. The resultant models are named as CFP-CNN-*da* and OCT-CNN-*da*. As for multi-modal baselines, we consider the following three methods, *i.e.,* EarlyFusion, LateFusion and MM-MHSA. EarlyFusion stacks the CFP and OCT images into a 4-channel input, while LateFusion combines CFP-CNN and OCT-CNN by averaging the output

of their softmax layers. MM-MHSA uses multi-modal multi-head self attention [21] to fuse CFP and OCT features[4].

Results. As Table III shows, MM-CNN, trained without using the proposed data augmentation, is less effective than OCT-CNN (0.872 *vs* 0.886 in F1 and 0.804 *vs* 0.818 in Accuracy). Similarly, EarlyFusion, LateFusion and MM-MHSA do not surpass OCT-CNN. These results suggest that exploit two modalities does not necessarily guarantee better performance.

Compared to OCT-CNN, our MM-CNN-*da* obtains relative improvements of 3.2% in F1 and 5.5% in Accuracy. As Fig. 6 shows, MM-CNN-da effectively reduces misclassification between PCV and wetAMD.

Concerning the two single-modal models, OCT-CNN outperforms CFP-CNN, 0.886 *vs* 0.774 in terms of the overall F1 score. The confusion matrices in Fig. 6 reveal more details. CFP-CNN recognizes the normal class as easy as OCT-CNN. It also has a good ability to find true PCV. However, CFP-CNN tends to incorrectly predict dryAMD as wetAMD and wetAMD as PCV. OCT-CNN also has difficulty in distinguishing PCV from wetAMD. Due to its relatively distinct visual patterns, dryAMD is relatively easier to be recognized than PCV and wetAMD, thus better separated in the learned CNN feature spaces, see Fig. 7. Hence, although dryAMD has fewer training samples, its result is better than that of the other two.

Fig. 6. **Confusion matrices**. Recall that per model we run training and evaluation three times. So for a fair comparison, we plot the confusion matrix of each model with its median-performance run. Compared to the best single-modal baseline (OCT-CNN), MM-CNN-*da* effectively reduces confusion between PCV and wetAMD.

The better overall performance of OCT-CNN-*da* against OCT-CNN (and CFP-CNN-*da* against CFP-CNN) suggests that the single-modal models also benefit from the synthetic instances. As Table 3 shows, CFP-CNN-da has higher sensitivity for dryAMD (0.817) and PCV (0.667) than CFP-CNN (0.783 and 0.533). This, however, comes at the cost of

---

[4]MM-MHSA is implemented as follows. Given $\bar{F}_f$ and $\bar{F}_o$ produced by CFP-CNN and OCT-CNN, each feature first goes through a specific 512×512 FC layer and is stacked afterwards to form a 2×512 feature array. The feature array is then fed into a 4-head self-attention block followed by an Add&Norm layer, a feedforward network and another Add&Norm. The resultant 2×512 features are flattened and fed to a 1024×4 FC layer for classification.

TABLE III

COMPARING SINGLE-MODAL AND MULTI-MODAL MODELS. WE USE OCT-CNN, THE BEST SINGLE-MODAL BASELINE, AS A REFERENCE. RELATIVE IMPROVEMENTS OVER OCT-CNN ARE SHOWN IN PARENTHESES. Sen. INDICATES SENSITIVITY AND Spe. AS SPECIFICITY. FOR EACH MODEL WE RUN TRAINING AND EVALUATION THREE TIMES AND REPORT AVERAGE SCORES PER METRIC. OUR PROPOSED MM-CNN-da PERFORMS THE BEST.

Model	normal			dryAMD			PCV			wetAMD			Overall	
	Sen.	Spe.	F1.	Sen.	Spe.	F1.	Sen.	Spe.	F1.	Sen.	Spe.	F1.	F1.	Accuracy
CFP-CNN	**1.000**	**1.000**	**1.000**	0.783	0.867	0.798	0.533	0.928	0.636	0.550	0.828	0.661	0.774 (↓-12.6%)	0.717 (↓-12.3%)
CFP-CNN-da	**1.000**	0.989	0.994	0.817	0.867	0.831	0.667	0.906	0.764	0.417	0.872	0.555	0.786 (↓-11.3%)	0.725 (↓-11.4%)
OCT-CNN	**1.000**	**1.000**	**1.000**	0.877	1.000	0.934	0.745	0.889	0.810	0.754	0.854	0.801	0.886	0.818
OCT-CNN-da	**1.000**	0.992	0.996	0.868	0.994	0.927	0.766	0.910	0.831	0.763	0.860	0.809	0.891 (↑+0.6%)	0.825 (↑+0.9%)
EarlyFusion	**1.000**	0.992	0.996	0.825	0.997	0.902	0.759	0.847	0.796	0.640	0.851	0.730	0.856 (↓-3.4%)	0.779 (↓-4.8%)
LateFusion	**1.000**	**1.000**	**1.000**	**1.000**	0.943	**0.971**	0.553	**0.948**	0.698	0.790	0.827	0.808	0.869 (↓-1.9%)	0.792 (↓-3.2%)
MM-MHSA	**1.000**	**1.000**	**1.000**	0.763	0.997	0.863	0.667	0.924	0.753	0.816	0.771	0.785	0.850 (↓-4.1%)	0.779 (↓-4.8%)
MM-CNN	**1.000**	**1.000**	**1.000**	0.737	0.997	0.847	0.787	0.910	0.841	0.790	0.819	0.801	0.872 (↓-1.6%)	0.804 (↓-1.7%)
MM-CNN-da	**1.000**	**1.000**	**1.000**	0.868	**1.000**	0.929	**0.794**	**0.948**	**0.864**	**0.868**	**0.860**	**0.864**	**0.914** (↑+3.2%)	**0.863** (↑+5.5%)

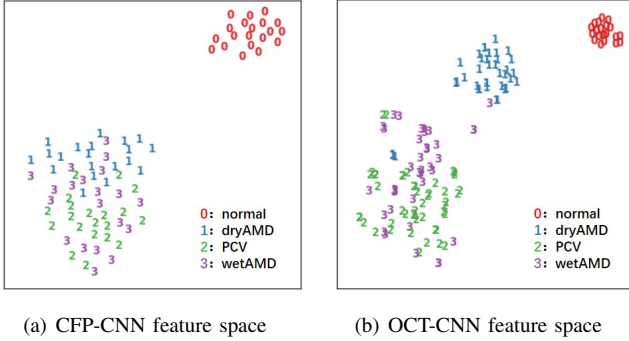(a) CFP-CNN feature space   (b) OCT-CNN feature space

Fig. 7. **Visualization of test image distribution in specific CNN feature spaces by t-SNE [39].**

losing sensitivity in wetAMD (from 0.550 to 0.417). More examples of wetAMD are misclassified. Recall that for both human and CNNs, the CFP modality alone is inadequate for recognizing wetAMD. We tried to train CFP-CNN on the synthetic data without finetuning on real data. That model recognizes wetAMD at sensitivity of 0.417, specificity of 0.928 and F1 of 0.547. The performance, though lower than that of CFP-CNN trained on real data, remains meaningful. We therefore attribute the performance degeneration of wetAMD to the unreliability of the CFP modality other than the CAM-conditioned image synthesis. The data augmentation is in general beneficial for CFP-CNN and OCT-CNN.

Compared to the single-modal CNNs, the amount of trainable parameters in MM-CNN is doubled, from 11.2M to 22.4M. This makes MM-CNN more data-hungry and more likely to overfit when learning from the small amount of real multi-modal samples, thus worse than OCT-CNN. For this reason, the proposed data augmentation brings more improvement to MM-CNN than the single-modal baselines. MM-MHSA, despite its excellent performance in various CV tasks [21], is not on par with MM-CNN for the current task, where multi-modal training samples are in short supply. Some qualitative results are shown in Fig. 4.

C. Experiment 2. Ablation Study on Data Augmentation

In the previous experiment we report the performance obtained by the joint use of the two data augmentation methods.

We now present an ablation study to reveal the effect of the individual method and the training strategy. We also investigate the advantage of the proposed CAM-based image synthesis against the state-of-the-art alternative [26].

Effect of the individual data augmentation methods. Table IV shows the effect of different data augmentation methods and their combinations on the performance of multi-modal CNNs. Note that common data augmentation is applied to all methods. For instance, the performance of Loose Pairing in the third row is actually obtained by a joint use of common data augmentation and Loose Pairing, while "Loose Pairing + CAM-conditioned image synthesis" is a combination of common data augmentation, Loose Pairing and CAM-conditioned image synthesis. As shown in Table IV, using either Loose Pairing or CAM-conditioned image synthesis is beneficial, bringing in relative improvements of 2.9% and 2.3% in F1 and 4.1% and 5.2% in Accuracy, respectively, when compared to MM-CNN that uses only common data augmentation. The best combination is to use CAM-conditioned image synthesis to expand the single-modal samples, Loose Paring to expand the multi-modal samples, followed by the two-stage training strategy to properly balance the synthetic and real samples. Such a strategy gives relative improvements of 4.8% in F1 and 7.3% in Accuracy.

Training strategy. To justify the necessity of the two-stage training strategy described in Section III-C, we instead train the two-stream CNN on the real and synthetic data simultaneously, with a ratio of 1:3. The performance of this one-stage training strategy is provided in the second last row of Table IV. While also improving over MM-CNN, it is not on a par with the two-stage training strategy.

Alternative for image synthesis. As we have noted in Section II-B, Burlina et al. [26] use PGGAN to generate CFP images for binary AMD categorization. In order to use PGGAN to generate CFP / OCT images for each of the three AMD subclasses in this study, we construct the input of PGGAN by concatenating a 3-dimensional one-hot label vector and a random noise vector. Following [40], we set the size of the input vector to be 100. We term this baseline label-conditioned image synthesis. As Table IV shows, the baseline is less effective than CAM-conditioned image synthesis. Note that our proposed method is capable of explicitly leveraging the spatial information of abnormal regions in a source image
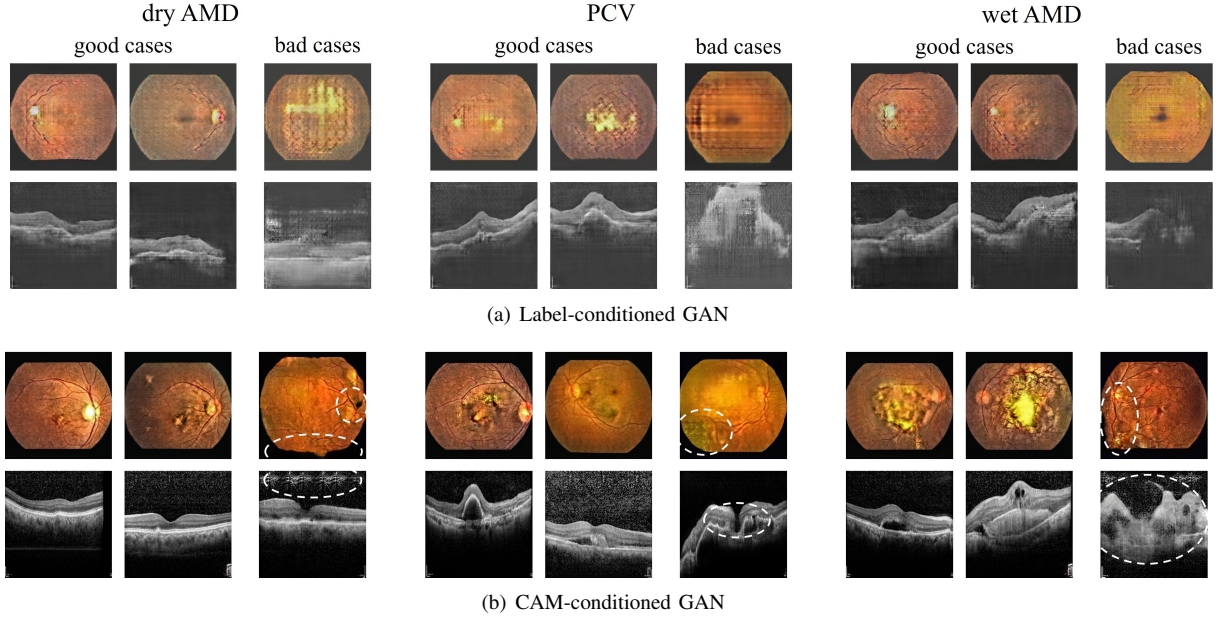
(a) Label-conditioned GAN

(b) CAM-conditioned GAN

Fig. 8. **Images generated by label-conditioned / CAM-conditioned GANs**. The latter generates images that are visually more realistic and show better details around the macula. For our method, a synthetic image can be paired with its source image (from which CAMs are extracted), whilst a synthetic image by label-conditioned GAN can only be paired with a class label and random noise. So a controlled comparison given the same source image is unfeasible. Failed cases are shown in the last column in each group, with unrealistic regions marked out by dotted ellipses.

TABLE IV

PERFORMANCE OF MM-CNN TRAINED WITH VARIOUS DATA AUGMENTATION METHODS, SORTED IN ASCENDING ORDER IN TERMS OF F1. WE USE MM-CNN, EXCLUSIVELY TRAINED WITH COMMON DATA AUGMENTATION, AS A REFERENCE. RELATIVE IMPROVEMENTS OVER MM-CNN ARE SHOWN IN PARENTHESES. WHILE ALL DATA AUGMENTATION METHODS ARE HELPFUL, THE JOINT USE OF LOOSE PAIRING AND CAM-CONDITIONED IMAGE SYNTHESIS IS THE BEST.

Data augmentation	F1	Accuracy
∘ Common data augmentation (MM-CNN in Table III)	0.872	0.804
∘ Common data augmentation ∘ Label-conditioned image synthesis	0.887 (↑+1.7%)	0.825 (↑+2.6%)
∘ Common data augmentation ∘ CAM-conditioned image synthesis	0.892 (↑+2.3%)	0.846 (↑+5.2%)
∘ Common data augmentation ∘ Loose Pairing (Our conference version [24])	0.897 (↑+2.9%)	0.837 (↑+4.1%)
∘ Common data augmentation ∘ Loose Pairing + CAM-conditioned image synthesis, *one-stage training*	0.904 (↑+3.7%)	0.849 (↑+5.6%)
∘ Common data augmentation ∘ Loose Pairing + CAM-conditioned image synthesis, *two-stage training*	**0.914 (↑+4.8%)**	**0.863 (↑+7.3%)**

in the form of CAMs (Fig. 5). By contrast, the input of the baseline is limited to the class label and the random noise. Some synthetic images are presented in Fig. 8. Compared to images generated by the label-conditioned GAN, images generated by our CAM-conditioned GAN are visually more realistic with better details, in particular around the macular. We also observe failed cases that contain unrealistic regions, see the last column of each group in Fig. 8. Nonetheless, such base cases are not common. As the confusion matrices in Fig. 6 show, the synthetic training examples contribute to AMD

categorization by reducing misclassification between PCV and wetAMD. To assess the overall quality of the synthetic data, we train a four-class MM-CNN on the synthetic data plus real images of the normal class. With F1 of 0.826 and Accuracy of 0.741, the model is somewhere between CFP-CNN and OCT-CNN, suggesting that the synthetic data is largely meaningful.

D. Experiment 3. Comparison to the State-of-the-Art

Baselines. As we have noted in Section I, the only prior art on multi-modal AMD categorization is by Yoo *et al.* [14]. As their experimental data is non-public, we replicate their method, *i.e.,* feature extraction by a pretrained CNN and classification by a random forest classifier. To make the comparison fair, we replace VGG19 used in [14] by ResNet-18. In addition, we are interested in whether the proposed multi-modal data augmentation method is also useful for this baseline. We term this enhanced baseline Yoo *et al.* *-da*.

Results. As Table V shows, the proposed MM-CNN-*da* clearly outperforms Yoo *et al.* , 0.914 *vs* 0.792 in F1 and 0.863 *vs* 0.690 in Accuracy. the proposed data augmentation methods also work for Yoo *et al.* , brining in relative improvements of 5.7% in F1 and 7.8% in Accuracy, respectively. Still, Yoo *et al.* *-da* is less effective than MM-CNN-*da*. The result justifies the superiority of our end-to-end learning method.

TABLE V

COMPARISON WITH SOTA FOR MULTI-MODAL AMD CATEGORIZATION.

Model	F1-score					Accuracy
	normal	*dryAMD*	*PCV*	*wetAMD*	*Overall*	
Yoo *et al.* [14]	**1.000**	0.783	0.648	0.736	0.792	0.690
Yoo *et al.* *-da*	**1.000**	0.880	0.740	0.726	0.837	0.744
MM-CNN-*da*	**1.000**	**0.929**	**0.864**	**0.864**	**0.914**	**0.863**

E. Discussion

There are several limitations in the current study. Due to the tremendous cost of collecting multi-modal data with ground-truth, the dataset used in our experiments is relatively small. While this does not affect our major conclusions (see the Appendix), the performance of our model is likely to drop when directly applying it on external data collected from different devices. Although we believe that our multi-modal approach can go beyond AMD categorization, its applicability for recognizing more retinal diseases remains to be justified. We have release our source code[1] so readers of interest may train and evaluate based on their own data.

Despite its effectiveness for improving classification, CAM-based image synthesis is not guaranteed to generate clinically and physiologically valid images. For instance, the fundus image generated with manipulated CAM on the first row of Fig. 5(a) appears to have two macular structures. A hotfix to deal with the issue is probably to detect primary physiological structures such as macula and optic disc [41] on the generated images and heuristically exclude the invalid.

Our multi-modal framework accepts only one single OCT image per input. So there is a gap between our setup and the real clinical settings, where an OCT scan typically renders a 3D volume, and very often the operator will choose more than one B-scan from the volume. To handle the OCT volume as input, a simple way is to select the central B-scan, with the other B-scans ignored. Alternatively, we may consider pairing each of the B-scans with the CFP image, using the current framework to make predictions per pair, and finally adopting certain score aggregation strategy, say max pooling, to make the volume-level prediction. We leave this for future work.

V. CONCLUSIONS

This paper tackles multi-modal AMD categorization by end-to-end deep learning. Extensive experiments on a clinical dataset allow us to answer the questions asked in the beginning and consequently draw conclusions as follows. Comparing the two modalities, OCT is more effective than CFP. When end-to-end trained, OCT-CNN obtains F1 of 0.886 and Accuracy of 0.818 on our dataset, making it a nontrivial baseline to beat. Compared to OCT-CNN, our multi-modal CNN recognizes PCV and wetAMD with higher sensitivity and specificity, scoring F1 of 0.914 and Accuracy of 0.863. However, using the two-stream network architecture alone is insufficient. In order to surpass the best single-modal baseline, the multi-modal CNN needs to be properly trained on data augmented by CAM-conditioned image synthesis and loose pairing.

APPENDIX

The effect of CLAHE. Table VI shows the performance of models with and without the CLAHE enhancement on color fundus images. Concerning the overall performance, CLAHE is helpful for CFP-CNN and MM-CNN-*da*, while no positive nor negative effect is observed on MM-CNN.

The effect of distinct data splits. So far all the experiments are performed based on the data split provided in Table II. To check whether our major conclusions depend on this specific

Model	CLAHE	F1-score					Accuracy
		normal	dryAMD	PCV	wetAMD	Overall	
CFP-CNN	−	1.000	0.816	0.663	0.570	0.762	0.700
	+	1.000	0.798	0.636	0.661	0.774 (↑+1.6%)	0.717 (↑+2.4%)
MM-CNN	−	1.000	0.859	0.818	0.809	0.872	0.804
	+	1.000	0.847	0.841	0.801	0.872 (0.0%)	0.804 (0.0%)
MM-CNN-*da*	−	0.989	0.882	0.858	0.860	0.839	0.897
	+	1.000	0.929	0.864	0.864	0.863 (↑+2.9%)	0.914 (↑+2.5%)

data split, we re-run our experiments on four new data splits, each obtained by shuffling the entire dataset followed by training / validation / test partition as described in Sec. IV-A. Viewing the data split as a random factor, averaging the performance of a model obtained on distinct data splits allows us to cancel out this random factor. As shown in Fig. 9, our major conclusions, *i.e.*, OCT-CNN is better than CFP-CNN, MM-CNN alone is insufficient, and MM-CNN-*da* is the best, remain valid. In addition, we observe that the relatively small amount of dryAMD for training mainly affects CFP-CNN, which has a relatively large standard deviation (std) of 0.120 in F1 for dryAMD. In contrast, OCT-CNN has a much smaller std of 0.040, largely because dryAMD-related visual patterns such as drusen are more visible in OCT images.
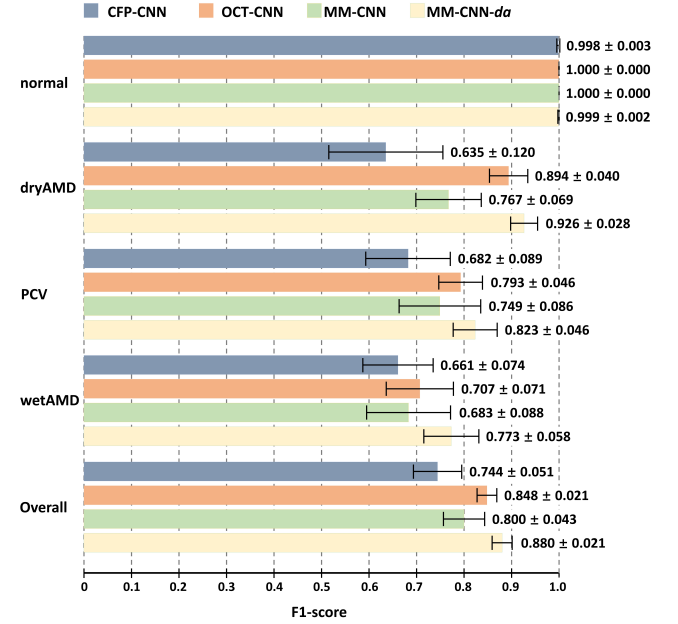

Fig. 9. **Model performance given five distinct data splits**.

We have also tried with subject-level split, adjusting the previous data split in Table II to ensure zero subject overlap between the training / validation / test sets, see Table VII. The performance of the primary models on the new data split is presented in Table VIII. Our major conclusions, *i.e.*, the two-stream network architecture alone is insufficient to beat the best single-modal baseline and the proposed multi-modal data augmentation is effective, remain valid.

REFERENCES

[1] W. L. Wong *et al.*, "Global prevalence of age-related macular degeneration and disease burden projection for 2020 and 2040: A systematic

TABLE VII

SUBJECT-LEVEL DATA SPLIT. IN PARENTHESES ARE NUMBER OF SUBJECTS PER CLASS.

Class	Training set		Validation set		Test set	
	CFP	OCT	CFP	OCT	CFP	OCT
normal	155 (155)	156 (155)	20 (20)	20 (20)	20 (20)	20 (20)
dryAMD	67 (57)	42 (23)	20 (19)	36 (19)	20 (18)	28 (17)
PCV	259 (228)	294 (137)	20 (20)	43 (20)	20 (19)	43 (19)
wetAMD	453 (340)	532 (245)	20 (20)	35 (19)	20 (19)	40 (19)

TABLE VIII

PERFORMANCE OF SINGLE-MODAL / MULTI-MODAL MODELS BASED ON THE DATA SPLIT IN TABLE VII.

Model	F1-score					Accuracy
	normal	dryAMD	PCV	wetAMD	Overall	
CFP-CNN	0.997	0.732	0.824	0.643	0.799	0.733
OCT-CNN	**1.000**	0.849	0.869	0.789	0.877	0.819
LateFusion	**1.000**	0.813	0.919	0.74	0.868	0.807
MM-CNN	**1.000**	0.800	0.852	0.680	0.833	0.766
MM-CNN-da	**1.000**	**0.904**	**0.913**	**0.823**	**0.910**	**0.863**

review and meta-analysis," *The Lancet Global Health*, vol. 2, no. 2, pp. e106–e116, 2014.

[2] A. DeWan *et al.*, "HTRA1 promoter polymorphism in wet age-related macular degeneration," *Science*, vol. 314, no. 5801, pp. 989–992, 2006.

[3] J. E. Lee *et al.*, "Efficacy of fixed-dosing aflibercept for treating polypoidal choroidal vasculopathy: 1-year results of the vault study," *Graefe's Archive for Clinical and Experimental Ophthalmology*, vol. 255, no. 3, pp. 493–502, 2017.

[4] A. Laude *et al.*, "Polypoidal choroidal vasculopathy and neovascular age-related macular degeneration: Same or different disease?" *Progress in Retinal and Eye Research*, vol. 29, no. 1, pp. 19–29, 2010.

[5] C. W. Wong *et al.*, "Age-related macular degeneration and polypoidal choroidal vasculopathy in Asians," *Progress in Retinal and Eye Research*, vol. 53, pp. 107–139, 2016.

[6] G. T. Kokame, T. E. deCarlo, K. N. Kaneko, J. N. Omizo, and R. Lian, "Anti–vascular endothelial growth factor resistance in exudative macular degeneration and polypoidal choroidal vasculopathy," *Ophthalmology Retina*, vol. 3, no. 9, pp. 744–752, 2019.

[7] J. De Fauw *et al.*, "Clinically applicable deep learning for diagnosis and referral in retinal disease," *Nature medicine*, vol. 24, no. 9, pp. 1342–1350, 2018.

[8] P. Burlina, D. E. Freund, N. Joshi, Y. Wolfson, and N. M. Bressler, "Detection of age-related macular degeneration via deep learning," in *ISBI*, 2016.

[9] P. M. Burlina, N. Joshi, M. Pekala, K. D. Pacheco, D. E. Freund, and N. M. Bressler, "Automated grading of age-related macular degeneration from color fundus images using deep convolutional neural networks," *JAMA Ophthalmology*, vol. 135, no. 11, pp. 1170–1176, 2017.

[10] F. Grassmann *et al.*, "A deep learning algorithm for prediction of age-related eye disease study severity scale for age-related macular degeneration from color fundus photography," *Ophthalmology*, vol. 125, no. 9, pp. 1410–1420, 2018.

[11] C. S. Lee, D. M. Baughman, and A. Y. Lee, "Deep learning is effective for classifying normal versus age-related macular degeneration OCT images," *Ophthalmology Retina*, vol. 1, no. 4, pp. 322–327, 2017.

[12] S. P. K. Karri, D. Chakraborty, and J. Chatterjee, "Transfer learning based classification of optical coherence tomography images with diabetic macular edema and dry age-related macular degeneration," *Biomedical Optics Express*, vol. 8, no. 2, pp. 579–592, 2017.

[13] M. Treder, J. L. Lauermann, and N. Eter, "Automated detection of exudative age-related macular degeneration in spectral domain optical coherence tomography using deep learning," *Graefe's Archive for Clinical and Experimental Ophthalmology*, vol. 256, no. 2, pp. 259–265, 2018.

[14] T. K. Yoo, J. Y. Choi, J. G. Seo, B. Ramasubramanian, S. Selvaperumal, and D. W. Kim, "The possibility of the combination of OCT and fundus images for improving the diagnostic accuracy of deep learning for age-related macular degeneration: a preliminary experiment," *Medical & Biological Engineering & Computing*, vol. 57, no. 3, pp. 677–687, 2019.

[15] P. Burlina, K. D. Pacheco, N. Joshi, D. E. Freund, and N. M. Bressler, "Comparing humans and deep learning performance for grading AMD: A study in using universal deep features and transfer learning for automated AMD analysis," *Computers in Biology & Medicine*, vol. 82, pp. 80–86, 2017.

[16] D. B. Russakoff, A. Lamin, J. D. Oakley, A. M. Dubis, and S. Sivaprasad, "Deep learning for prediction of AMD progression: A pilot study," *Investigative Ophthalmology & Visual Science*, vol. 60, no. 2, pp. 712–722, 2019.

[17] D. S. Kermany *et al.*, "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Cell*, vol. 172, no. 5, p. 1122–1131.e9, 2018.

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *NIPS*, 2012.

[19] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015.

[20] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *CVPR*, 2016.

[21] J. Yu, J. Li, Z. Yu, and Q. Huang, "Multimodal transformer with multi-view visual representation for image captioning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 12, pp. 4467–4480, 2019.

[22] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *CVPR*, 2016.

[23] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional GANs," in *CVPR*, 2018.

[24] W. Wang *et al.*, "Two-stream CNN with loose pair training for multimodal AMD categorization," in *MICCAI*, 2019.

[25] Y. Kanagasingam, A. Bhuiyan, M. D. Abramoff, R. T. Smith, L. Goldschmidt, and T. Y. Wong, "Progress on retinal image analysis for age related macular degeneration," *Progress in Retinal and Eye Research*, vol. 38, pp. 20–42, 2014.

[26] P. M. Burlina, N. Joshi, K. D. Pacheco, T. A. Liu, and N. M. Bressler, "Assessment of deep generative models for high-resolution synthetic retinal image generation of age-related macular degeneration," *JAMA Ophthalmology*, vol. 137, no. 3, pp. 258–264, 2019.

[27] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," in *ICLR*, 2018.

[28] R. Zheng *et al.*, "Detection of exudates in fundus photographs with imbalanced learning using conditional generative adversarial network," *Biomedical Optics Express*, vol. 9, no. 10, pp. 4863–4878, 2018.

[29] H.-C. Shin *et al.*, "Medical image synthesis for data augmentation and anonymization using generative adversarial networks," in *International Workshop on Simulation and Synthesis in Medical Imaging*, 2018.

[30] Y. Xing *et al.*, "Adversarial pulmonary pathology translation for pairwise chest x-ray data augmentation," in *MICCAI*, 2019.

[31] Y. Zhou, X. He, S. Cui, F. Zhu, L. Liu, and L. Shao, "High-resolution diabetic retinopathy image synthesis manipulated by grading and lesions," in *MICCAI*, 2019.

[32] X. Yi, E. Walia, and P. Babyn, "Generative adversarial network in medical imaging: A review," *Medical Image Analysis*, vol. 58, p. 101552, 2019.

[33] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *CVPR*, 2017.

[34] Q. Wei, X. Li, H. Wang, D. Ding, W. Yu, and Y. Chen, "Laser scar detection in fundus images using convolutional neural networks," in *ACCV*, 2018.

[35] X. Lai, X. Li, R. Qian, D. Ding, J. Wu, and J. Xu, "Four models for automatic recognition of left and right eye in fundus images," in *MMM*, 2019.

[36] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *CVPR*, 2009.

[37] T. Jintasuttisak and S. Intajag, "Color retinal image enhancement by Rayleigh contrast-limited adaptive histogram equalization," in *ICCAS*, 2014.

[38] A. Paszke *et al.*, "PyTorch: An imperative style, high-performance deep learning library," in *NeurIPS*, 2019.

[39] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, no. 11, pp. 2579–2605, 2008.

[40] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in *ICLR*, 2016.

[41] Z. Yang *et al.*, "Joint localization of optic disc and fovea in ultra-widefield fundus images," in *MLMI@MICCAI*, 2019.