IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 28 March 2023, accepted 12 May 2023, date of publication 22 May 2023, date of current version 1 June 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3278596

RESEARCH ARTICLE

Attention-Aided Generative Learning for Multi-Scale Multi-Modal Fundus Image Translation

VAN-NGUYEN PHAM[1], DUC-TAI LE[2], JUNGHYUN BUM[3], EUN JUNG LEE[4], JONG CHUL HAN[4,5,6], AND HYUNSEUNG CHOO[1,2], (Member, IEEE)

[1]Department of Electrical and Computer Engineering, Sungkyunkwan University, Suwon 16419, South Korea
[2]College of Computing and Informatics, Sungkyunkwan University, Suwon 16419, South Korea
[3]Sungkyun AI Research Institute, Sungkyunkwan University, Suwon 16419, South Korea
[4]Department of Ophthalmology, Samsung Medical Center, Seoul 06351, South Korea
[5]School of Medicine, Sungkyunkwan University, Suwon 16419, South Korea
[6]Department of Medical Device, Management and Research, SAIHST, Sungkyunkwan University, Suwon 16419, South Korea

Corresponding authors: Duc-Tai Le (ldtai@skku.edu), Jong Chul Han (heartmedic@skku.edu), and Hyunseung Choo (choo@skku.edu)

This work was supported in part by the Institute for Information and Communications Technology Planning and Evaluation (IITP) Grant funded by the Korea Government [Ministry of Science and ICT (MSIT)] under the Artificial Intelligence Graduate School under Grant 2019-0-00421, in part by the Artificial Intelligence Innovation Hub under Grant 2021-0-02068, and in part by the Information and Communications Technology (ICT) Creative Consilience Program under Grant IITP-2023-2020-0-01821.

ABSTRACT Conventional fundus images (CFIs) and ultra-widefield fundus images (UFIs) are two fundamental image modalities in ophthalmology. While CFIs provide a detailed view of the optic nerve head and the posterior pole of an eye, their clinical use is associated with high costs and patient inconvenience due to the requirement of good pupil dilation. On the other hand, UFIs capture peripheral lesions, but their image quality is sensitive to factors such as pupil size, eye position, and eyelashes, leading to greater variability between examinations compared to CFIs. The widefield retina view of UFIs offers the theoretical possibility of generating CFIs from available UFIs to reduce patient examination costs. A recent study has shown the feasibility of this approach by leveraging deep learning techniques for the UFI-to-CFI translation task. However, the technique suffers from the heterogeneous scales of the image modalities and variations in the brightness of the training data. In this paper, we address these issues with a novel framework consisting of three stages: cropping, enhancement, and translation. The first stage is an optic disc-centered cropping strategy that helps to alleviate the scale difference between the two image domains. The second stage mitigates the variation in training data brightness and unifies the mask between the two modalities. In the last stage, we introduce an attention-aided generative learning model to translate a given UFI into the CFI domain. Our experimental results demonstrate the success of the proposed method on 1, 011 UFIs, with 99.8% of the generated CFIs evaluated as good quality and usable. Expert evaluations confirm significant visual quality improvements in the generated CFIs compared to the UFIs, ranging from 10% to 80% for features such as optic nerve structure, vascular distribution, and drusen. Furthermore, using generated CFIs in an AI-based diagnosis system for age-related macular degeneration results in superior accuracy compared to UFIs and competitive performance relative to real CFIs. These results showcase the potential of our approach for automatic disease diagnosis and monitoring.

INDEX TERMS Conventional fundus images, deep learning, generative learning, ophthalmology, unpaired image-to-image translation, ultra wide-field fundus images.

I. INTRODUCTION

Fundus images are widely used in ophthalmology for the diagnosis and monitoring of various eye diseases. Conventional fundus images (CFIs) have been one of the

The associate editor coordinating the review of this manuscript and approving it for publication was K. C. Santosh.

most popular diagnostic tools used by ophthalmologists due to their ability to capture the posterior pole of the retina [1]. However, this image modality has two main disadvantages. First, a CFI examination consumes a large amount of time and is burdensome for patients. Before having an eye examination, patients are required to have dilating eye drops applied to enlarge their pupils, which takes 15 to 30 minutes. While the pupils are dilated, the eyes become sensitive to bright light and the patients' vision is blurry. These negative effects may last for several hours, which is inconvenient for the patients. Second, the number of diseases that can be detected with CFIs is limited. CFIs are taken with a small field of view, typically between 30 and 60 degrees, as shown in Figure 1(a), which covers the optic disc, macula, and nearby regions, making it mainly suitable for the diagnosis of glaucoma and macular diseases. As a result, another image modality are required to address CFI limitations.

Ultra-widefield fundus images (UFIs) do not require pupil dilation, making them faster to obtain than CFIs. With a field of view up to 200 degrees, UFIs are used to detect peripheral diseases that do not appear in CFIs, such as diabetic retinopathy, retinal vein occlusion, and retinal detachment. These advantages have made the use of UFIs increasingly popular. However, the quality of UFI images is susceptible to variations among patients due to factors such as pupil size, eye position, and eyelashes. Additionally, while UFIs include the optic disc and macula, as illustrated in Figure 1(b), most ophthalmologists are more familiar with CFIs for the diagnosis of glaucoma and macular diseases. Therefore, generating CFIs from UFIs using deep learning-based methods is useful for monitoring these diseases. This computer-based UFI-to-CFI translation approach not only helps patients avoid time-consuming and uncomfortable CFI examinations but also enables ophthalmologists to utilize their expertise in analyzing familiar CFI images for accurate diagnosis and treatment of glaucoma and macular diseases.

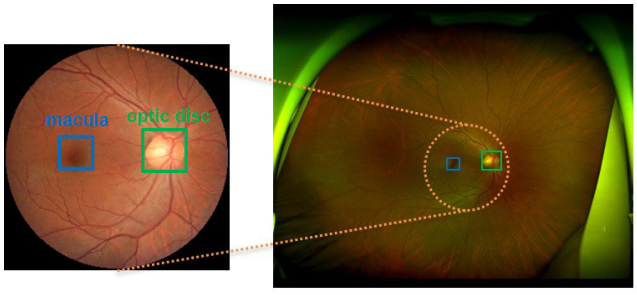

FIGURE 1. Two modalities of fundus photography. Area covered by CFI is also included in UFI.

Recent advances in deep learning, particularly generative adversarial networks (GANs), have enabled the generation of CFIs from UFIs. The first attempt to generate CFI from UFI was made in [1], where an intensity-based registration algorithm was employed to crop a central area in the UFI, which was then masked before being translated into the CFI domain using cycleGAN [2]. However, this approach has limitations in both the cropping step and the quality of the generated images. Since the registration algorithm is based on images from different eyes, the cropping may fail or the cropped images may be distorted. In addition, the variation in brightness across images poses a challenge in the translation step, leading to poor performance in generated CFIs.

In this paper, we propose a deep learning-based three-stage framework to address the limitations of the existing UFI-to-CFI translation. Our first contribution is an optic disc-centered cropping strategy that extracts a region of the UFI covering a similar area to that captured by CFI. This strategy uses the positions of the optic disc and macula to crop a portion of the UFI that includes these structures and nearby regions, ensuring that the cropped portion is useful for generating the CFI. Second, we introduce a dual illumination correction step and a mask unification method to preprocess the UFI and CFI, making them more consistent and suitable for translation. Finally, we use an attention-aided generative learning model to translate the preprocessed UFI into CFI. Our model leverages a convolutional block attention module (CBAM) [3] to improve the realism of the generated CFI. Experiments on a dataset of 2011 UFIs and 681 CFIs demonstrate that our approach outperforms existing methods in terms of visual quality and performance. The high-quality CFI images generated by our method have the potential to be utilized in a range of applications, including automatic disease diagnosis, monitoring, and research.

The rest of this paper is organized as follows: Section II reviews the related work, while section III provides a detailed description of our proposed method, including each step of the three-stage framework. Section IV presents the experimental settings, evaluation metrics, and results of our experiments. Finally, in section V, we draw our conclusions based on the results obtained from our proposed method.

II. RELATED WORK

A. DEEP LEARNING IN FUNDUS IMAGES

With superior performance compared to traditional methods, deep learning has been applied to CFI for various tasks such as multi-disease diagnosis [4], vessel segmentation [5], and optic disc and fovea localization [6]. For almost all tasks, deep learning brings impressive results which are sometimes better than human [7]. In [8], the authors designed a framework to detect 39 fundus diseases and achieved an area-under-the-roc-curve (AUC) score of 0.9984. This result is comparable to retinal specialists with more than 10 years of experience. For the vessel segmentation, the authors of [5] introduced a model called W-Net which brings state-of-the-art performance on multiple datasets. A model named FundusPosNet was designed in [6] to localize optic disc and fovea. This model is trained based on the regression of heatmap labels and it outperforms existing methods on

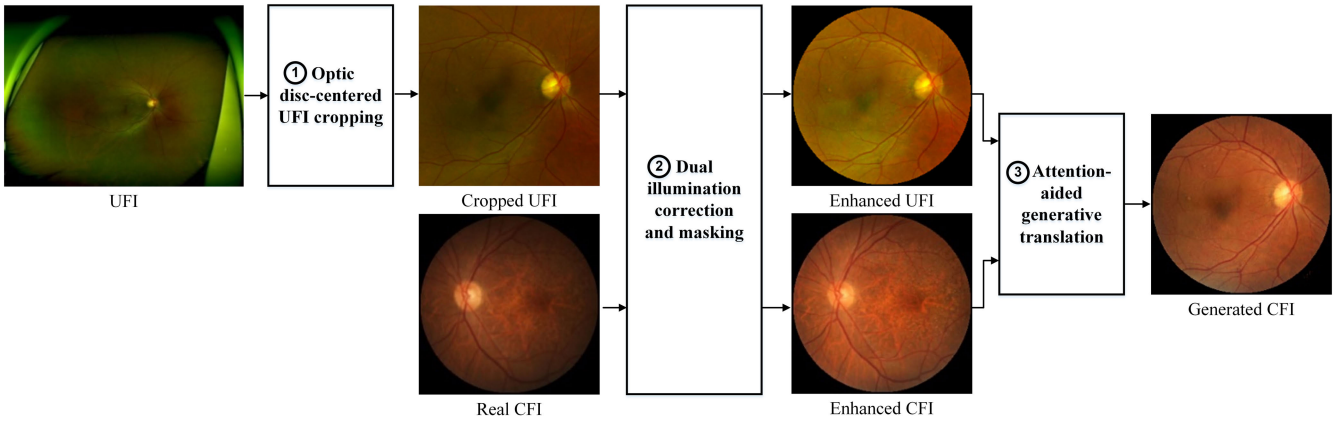

FIGURE 2. Overall process of generating CFI from UFI. There are three stages in our proposed method. The first stage is to crop a portion of UFI covering a similar area to that captured by CFI. In the second stage, the brightness of all images is improved and their masks are unified. The final stage transfers the enhanced UFI into the CFI domain.

three datasets: IDRiD [9], Messidor [10], and G1020 [11]. Recently, researchers have paid more attention to the UFI because it covers peripheral lesions which cannot be observed in the CFI. Various tasks on the CFI are also able to conduct on the UFI. In [12], the authors proposed using 6 contrast enhancement methods and the model ensemble technique to boost the performance of a multi-disease detection system. As a result, the system achieves 97.45% accuracy. For the task of vessel segmentation in UFI, the authors of [13] utilized UFIs and their corresponding fluorescein angiography images to iteratively train a multi-modal registration model and a weakly-supervised segmentation model. Once trained, the segmentation model can detect vessels without the fluorescein angiography images. To detect optic disc and fovea in the UFI, the authors of [14] proposed distance-based and direction-based losses to improve Faster RCNN detector [15]. Their method obtains an average intersection over union (IoU) score of 0.82.

B. IMAGE-TO-IMAGE TRANSLATION

The task of transforming images from one domain to another domain is referred to as image-to-image translation, most methods for this task are based on GAN [16]. In [17], a conditional GAN model is proposed for image translation using paired images which are usually difficult to obtain. To tackle this problem, the authors of [2] introduced a cycleGAN framework including two generators and two discriminators to translate images between two domains using unpaired images. The key point of cycleGAN is cycle consistency: when an image in domain A is transferred to domain B, if it is translated back to domain A, the result should be the same as the original image. For the translation between more than two domains, many works have been proposed with impressive results such as [18]. Recently, contrastive learning approach has been widely used and achieves state-of-the-art performance [19], [20], [21]. The main idea of contrastive learning is to create multiple versions of an image, then, their feature representations extracted by a deep network should be similar. For different images, the representations should be as different as possible.

C. FUNDUS IMAGE TRANSLATION

The task UFI-to-CFI translation was first tackled in [1], where the authors combined an intensity-based registration method with cycleGAN to obtain CFIs from UFIs. The goal of this translation is to make use of the additional information that CFIs provide for diagnosis. In the registration step, 20 manually cropped UFIs from normal images are used as templates to register with UFIs to obtain portions covering similar areas to CFIs. Each template results in one registered portion, and the final image is calculated based on the normalized cross-correlation between the registered portions and their corresponding templates. The registration is the affine transformation and is performed by the *imregister* function of Matlab. The registered images are then masked before being translated into the CFI domain by CycleGAN. In [22], the authors proposed a modified version of CycleGAN to transfer CFIs to UFIs, which is used as additional information for disease diagnosis. One modification made to CycleGAN is consistency regularization: different translated versions of the same image should have the same disease label. For each CFI, multiple versions created by data augmentation are passed through a generator to obtain UFIs. Labels of these UFIs are then generated by an inference model, and the consistency between the labels is represented by a consistency regularization loss.

III. PROPOSED METHOD

Our proposed three-stage framework for generating CFI from UFI is depicted in Figure 2, with the inputs being a set of UFIs and a set of real CFIs. The first stage involves an optic disc-centered cropping strategy, which extracts a portion of each UFI covering an area similar to that of the CFI. In the second stage, all images are processed to enhance their brightness and unify their masks across both domains. Finally, the

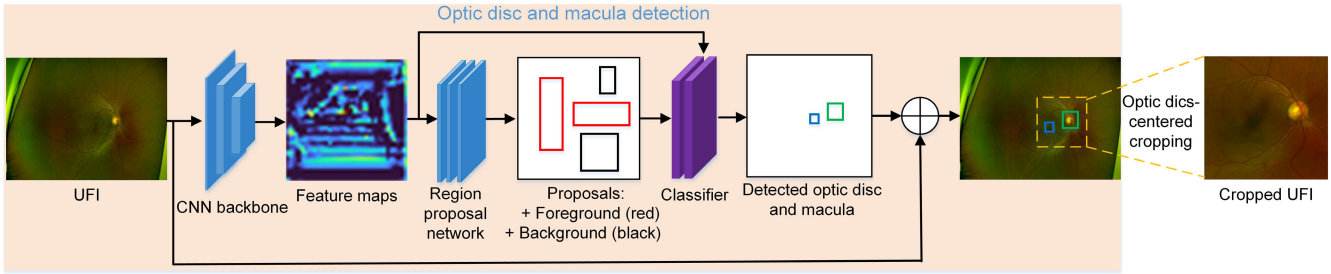

FIGURE 3. UFI cropping stage. The optic disc and macula are localized by Faster RCNN detector, then, their positions are used for cropping a desired area.

enhanced UFIs are translated into the CFI domain using an attention-aided generative model.

A. OPTIC DISC-CENTERED UFI CROPPING

At the first stage of our framework, a portion of the UFI covering a similar area to the CFI is obtained, as illustrated in Figure 3. To achieve this, we use a Faster RCNN detector [15] to localize the optic disc and macula in the UFI. These two biomarkers are then used to crop a desired portion. Specifically, the UFI is passed through a convolutional neural network (CNN) backbone to extract its feature maps. These feature maps are then used to generate proposals (anchor boxes) with corresponding objectness scores using a region proposal network. The objections scores indicate if the proposals are foreground or background. To reduce computation complexity, any proposals crossing the image boundaries are removed and non-maximum suppression is performed to eliminate boxes that overlap others with an IoU score of more than 0.7. Finally, the proposals go through a classifier for accurate prediction of categories and bounding boxes. Each proposal is classified as either optic disc, macula, or background, and the bounding box coordinates for each object are provided. If multiple boxes are predicted as the optic disc or macula, the one with the highest probability is selected as the final result, because there is only one of each object in a UFI.

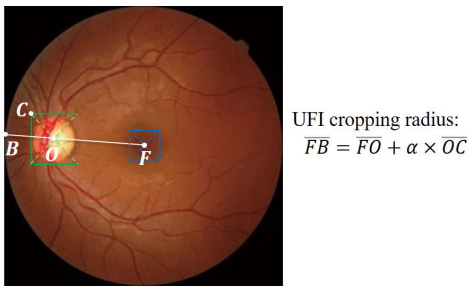

FIGURE 4. Radius of CFI, it is used for UFI cropping.

After localizing the optic disc and macula, the next step is to obtain a UFI area that is similar in appearance to the CFI. To achieve this, we propose an optic disc-centered cropping strategy based on the relationship between the positions of the optic disc, macula, and the area covered by the CFI, as illustrated in Figure 4. The retinal area covered by a CFI has a shape of a circle, is centered at the fovea (center of the macula), and includes the optic disc as well as surrounding regions [23]. Let F denote fovea, O denote the optic disc center, and $\overline{OC}$ is half the length of the diagonal of the optic disc box, then the radius of the CFI can be represented as follow:

$$\overline{FB} = \overline{FO} + \alpha\overline{OC} \tag{1}$$

where $\alpha > 1$ is a scaling factor to assure the optic disc is included in the CFI. This relationship is utilized for UFI cropping in which the center of the detected optic disc is represented by O, and F denotes the center of the detected macula. In practice, CFI is usually not centered at fovea, so, we randomly shift the center of the cropped image towards O with a maximum distance of $\frac{\overline{FO}}{2}$. We empirically find that this randomness can bring better results. The value of α is also a factor affecting the quality of generated CFIs, we will show our selection in section IV.

B. DUAL ILLUMINATION CORRECTION AND MASKING

In the second stage, we address two problems with the cropped UFIs and real CFIs: brightness and mask. Specifically, while the cropped UFIs are often underexposed, the real CFIs are either underexposed or overexposed. Additionally, the real CFIs contain a mask that is absent in the cropped UFIs. We empirically find that these issues cannot be effectively resolved by the image translation model at the final stage. Therefore, we adopt a dual illumination corrector in [24] to handle both underexposed and overexposed images, and a masking operation is performed to address the missing mask issue. The steps involved in the second stage are illustrated in Figure 5.

The illumination corrector contains three steps: dual illumination estimation, exposure correction, and fusion. For an input image, the forward and reverse illuminations are estimated, and then, based on the estimations, the underexposure and overexposure corrected images are obtained. Finally, they are fused with the input to get the corrected image. Let I and L_f denote the input image and the estimation of the forward illumination, respectively, then, the underexposure corrected

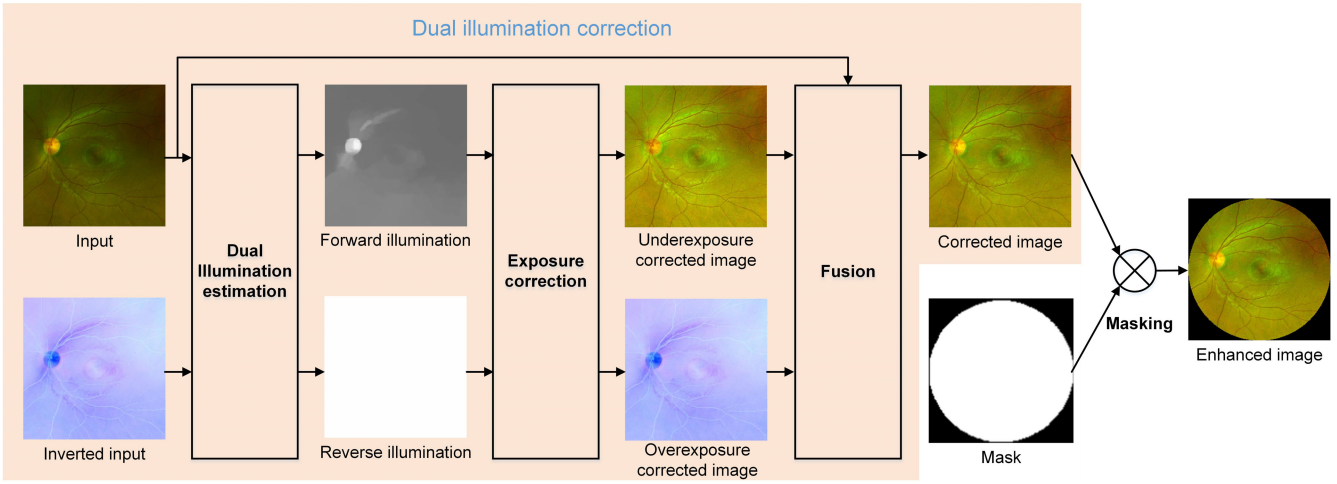

FIGURE 5. Dual illumination correction and masking stage. It includes two sub-tasks: dual illumination correction and masking. While the illumination correction is based on a dual estimation of illumination maps, masking is an element-wise multiplication between the mask and the corrected image.

image is:

$$I'_f = I * (L_f^{\gamma})^{-1},$$

where $*$ denote element-wise multiplication and γ is a Gamma adjustment to the forward illumination. A similar process is applied to the inverted input image $I_{inv} = (1 - I)$ and the reverse illumination L_r^{γ} to obtain the overexposure corrected image:

$$I'_r = I_{inv} * (L_r^{\gamma})^{-1},$$

Details about the estimations of the forward and reverse illuminations can be found in [24]. The final step of the dual illumination correction process is the fusion of the original image with the corrected under- and over-exposed images using Laplacian pyramid [25].

At the end of the second stage, the mask of images in the UFI and CFI domains is unified by a masking process. A circular binary mask, in which, the white area is a circle centered at pixel $\lfloor \frac{n}{2} \rfloor$, $\lfloor \frac{n}{2} \rfloor$ with a radius of $\frac{n}{2}$ is created, where n is the size of the mask. Pixel values at i^{th} row and j^{th} column of the mask are the same for three channels R, G, and B, it is:

$$p_{i,j} = \begin{cases} 1 & \text{if } d_{i,j} \leq \lfloor \frac{n}{2} \rfloor \\ 0 & \text{otherwise,} \end{cases}$$

where $d_{i,j}$ is the distance from the pixel $p_{i,j}$ to the center pixel $p_{\lfloor \frac{n}{2} \rfloor, \lfloor \frac{n}{2} \rfloor}$. The masking process is then performed by pixel-wise multiplication between the mask and the corrected image. The pixels inside the white area of the mask have a value of 1, which means that the contents inside the circle remain the same as those in the corrected image. On the other hand, the pixels outside the circle have a value of 0 and appear black in the enhanced image.

C. ATTENTION-AIDED GENERATIVE TRANSLATION

In the final stage of the proposed method, an attention-aided generative framework is employed to translate enhanced UFIs to the CFI domain. It uses a convolutional block attention module (CBAM) [3] to boost the performance of cycle-GAN [2]. Figure 6 depicts the framework, which comprises two generators (G_C, G_U) and two discriminators (D_C, D_U) for translating images between the UFI and CFI domains. The inputs of the model are enhanced UFIs and enhanced CFIs which can be either images of the same eyes or different eyes. While the generator G_C generates images in the CFI domain, the discriminator D_C classifies images into enhanced and

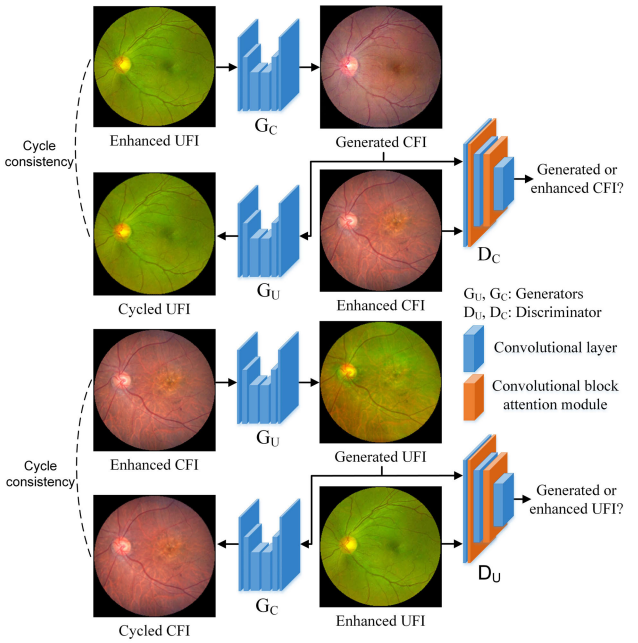

FIGURE 6. Attention aided CycleGAN framework. It includes two generators and two discriminators to translate images between UFI and CFI domains. Cycle consistency means that when a cycle of translation (*UFI → CFI → UFI* or *CFI → UFI → CFI*) is performed, the image should be unchanged. The convolutional block attention module is used to boost the quality of generated images.

generated CFI. The same applies to the generator G_U and the discriminator D_U which are used for generating and classifying images in the UFI domain. The training procedure of this framework is the same as that of CycleGAN. As a GAN-based model, the training process of the attention-aided CycleGAN includes a competition between the generators and discriminators, which is reflected by the GAN loss function:

$$L_{GAN} = [\log(1 - D_C(G_C(u))) + \log D_C(c)] + [\log(1 - D_U(G_U(c))) + \log D_U(u)],$$

where u and c denote an image in UFI and CFI domains, respectively. While the generators want to minimize the loss by generating images close to real ones, the discriminators try to classify generated images from the real ones to maximize the loss. In order to translate images between two domains, the training process contains two cycles: $UFI \rightarrow CFI \rightarrow UFI$ and $CFI \rightarrow UFI \rightarrow CFI$. The key idea is that when an image in the UFI domain is transferred into the CFI domain if it is converted back to the UFI domain, the result should be the same as the original UFI. This is called cycle consistency and is characterized by cycle-consistency loss:

$$L_{cycle} = ||u - G_U(G_C(u))||_1 + ||c - G_C(G_U(c))||_1,$$

where $||.||_1$ denotes *norm*-1. This loss is a pixel-wise comparison that not only assures the image can be recovered after being translated into another domain, but also has the role of maintaining the image structure during the translation. However, the cycle-consistency loss is not enough to transfer images into the target domain. As long as the image can be recovered to the original domain, the style of the target domain is free. For this reason, identity loss is included. The intuition behind this loss is that because the generator G_U is used to generate images in the UFI domain, if the input is already a UFI, the generator should keep the image the same, similarly for the generator G_C. Identity loss is defined as:

$$L_{identity} = ||u - G_U(u)||_1 + ||c - G_C(c)||_1.$$

Gather all components together, the loss for training is:

$$L = L_{GAN} + \lambda_1 L_{cycle} + \lambda_2 L_{identity},$$

in which λ_1 and λ_2 are parameters to balance three components.

Our network architectures for the generators and discriminators are improved from those of [1]. We maintain the generators and upgrade the discriminators by the CBAM which has been widely used to boost the performance of classification tasks. The design of the discriminators is illustrated in Figure 7. Each CBAM contains a channel attention module (CAM) followed by a spatial attention module (SAM). CBAM is inserted after convolutional layers to make the network focus on useful information. The output of a convolutional layer is a set of channels, each of which has a different contribution to the final result. The CAM makes the network pay more attention to important channels and reduces the impacts of others. Each channel contains many spatial areas

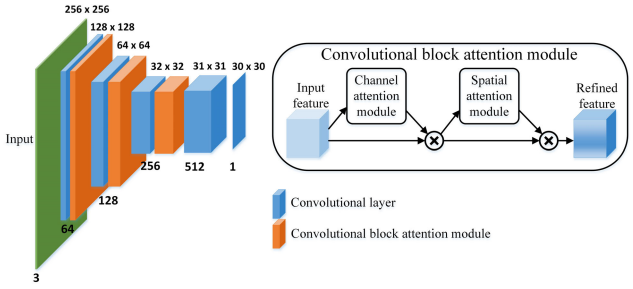

FIGURE 7. Discriminator architecture. Each convolutional layer halving spatial size of the image is followed by a convolutional block attention module. The numbers at the bottom and on top of each layer denote the number of channels and spatial size, respectively.

which have different roles in the final decision of the network. The SAM identifies the areas having high influences on the output and makes the network focus on these areas. CBAMs are injected after the convolutional layers which halve the spatial size of images. We empirically find that this way of injection brings better performance than inserting CBAM into every convolutional layer.

IV. EXPERIMENTS

A. DATASET, IMPLEMENTATION DETAILS, AND EVALUATION METRICS

1) DATASET AND IMPLEMENTATION DETAILS

For experiments, we collect a private dataset of 2, 011 UFIs from Samsung Medical Center and manually select 681 high-quality real CFIs from Eyepacs dataset [26]. This study was approved by the Institutional Review Board (IRB) at Samsung Medical Center (IRB No. 2022-06-032). While the UFIs are used in stage 1&3 of the framework, the real CFIs are used only at stage 3. For the detection of the optic disc and macula detection in the first stage, we label the optic disc and macula in 1, 204 UFIs to train the Faster RCNN detector. UFIs are resized to the resolution of 640 × 640, and the implementation of Faster RCNN is from Tensorflow object detection API [27]. The cropped UFIs are then resized to 256 × 256 before going to stage 2. The code for dual illumination correction is from: *https://github.com/pvnieo/Low-light-Image-Enhancement*. In the third stage, 1, 000 UFIs and 681 CFIs are used for training, and the rest of the UFIs is used for validation. All experiment settings of this stage are the same as those in [1]. Our codes for the ratio-based cropping, masking, and attention-aided cycleGAN are available at our Github: *https://github.com/VanNguyenPham/FundusImagesTranslation*, they are implemented on Tensorflow framework with Tesla V100 GPU.

2) EVALUATION METRICS

Since we use a GAN-based method to generate CFI, we will use two categories of metrics to evaluate the generated CFIs. The first category is to validate the quality of images generated by a GAN-based model, while the second one is to evaluate the quality of CFIs. For the first category, we select

Frechet inception distance (FID) [28] which has been commonly used in recent years because it agrees with human perception. A lower value of FID indicates that the generated images are more realistic. For the second category, we adopt Q_v [29] and an automatic tool in [30]. Q_v measures the quality of CFI based on the vessel structure without using reference images, and has a high agreement with full-reference metrics such as peak signal-to-noise ratio and structural similarity index measure [31], a higher value of Q_v indicates that the image has better quality. The automatic tool in [30] classifies CFI quality into three classes: good, usable, and reject. The quality classification is based on the ophthalmologists' perspective and determines whether the image's information is clear enough for diagnosis. While good quality states that the image has clear diagnostic information, reject class means the quality of the image is too low for diagnosis. The usable class includes images in several poor conditions but the main structure and lesions are clear enough to be detected by ophthalmologists. Since the metrics in the second category do not make sense if the generated CFIs are unrealistic, we use FID as the main metric for performance evaluation.

B. EXPERIMENT RESULTS

1) COMPARISON WITH EXISTING WORKS

In this section, we first compare our proposed method with the existing work for UFI-to-CFI translation. We then show the benefits of our attention-aided generative translation model in comparison with state-of-the-art techniques for image-to-image translation.

a: COMPARISON OF UFI-TO-CFI TRANSLATION METHODS

We compare our approach with the work in [1] where they cropped UFIs by an intensity-based registration method, masked the cropped UFIs, and then, translated the cropped images into the CFI domain by CycleGAN [2]. The cropping is considered as successful if the cropped image includes both the optic disc and macula. However, in [1], the registration between images of different eyes results in a successful cropping ratio of less than 10%, leading to too few successfully cropped images to train the translation model. In contrast, our framework utilizes the accurate localization of the optic disc and macula to implement an optic disc-centered cropping strategy with a 100% successful ratio. In addition to cropping, we want to demonstrate the benefits of our illumination enhancement and attention-aided translation model in the second and third stages. Specifically, we replace the cropped UFIs of [1] with ours and make a comparison, the result is shown in Table 1. Our framework leverages a combination of the illumination enhancement and the attention-aided translation model, which outperforms the translation-only approach used by Yoo et al. [1] in all evaluation metrics. Visualization of generated images is shown in Figure 8. Perceptually, our method produces images with higher contrast and brightness than the method of Yoo et al. [1]. Moreover, while examining the CFIs generated by the approach proposed by [1], we observed the bad generation of the optic disc and the disappearance of drusen (i.e., yellow dots), which can potentially result in an erroneous diagnosis. Such limitations were effectively addressed in our proposed framework, showcasing its potential to serve as an effective tool for accurate and reliable disease diagnosis and monitoring. In summary, our proposed approach improves two weaknesses of the existing method: cropping and the brightness variation of data. Our optic disc-centered cropping strategy ensures that all cropped images contain both the optic disc and macula, leading to a more robust and accurate translation model. Additionally, our illumination enhancement and attention-aided translation model contribute to generating high-quality CFIs with better brightness, contrast, and preservation of important features.

TABLE 1. Comparison of different methods for generating CFI from UFI. ↓ denotes lower is better, ↑ denotes higher is better.

Method	FID↓	Qv↑	Image quality	
			Good&Usable(%)↑	Reject(%)↓
Yoo et al. [1]	31.66	0.1053	99.60	0.40
Ours	**26.64**	**0.1190**	**99.80**	**0.20**

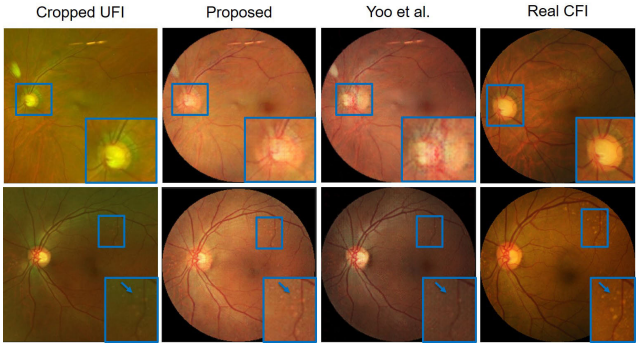

FIGURE 8. Visualization of detail preservation by two UFI-to-CFI-translation methods. While the first row shows the bad generation of the optic disc, the second row illustrates the disappearance of drusen (yellow dots) in the images generated by Yoo et al.

b: COMPARISON OF IMAGE-TO-IMAGE TRANSLATION MODELS

Next, we validate the effectiveness of our attention-aided translation model by comparing it with state-of-the-art techniques: CUT [19], ACL-GAN [20], and DCLGAN [21]. Stages 1 and 2 of our framework are fixed and the image translation model in stage 3 is replaced with the aforementioned techniques. Our results, shown in Table 2, demonstrate that our method achieves the best values for *FID* and Q_v, and

TABLE 2. Quantitative performance of different image translation models.

Method	FID↓	Qv↑	Image quality	
			Good&Usable(%)↑	Reject(%)↓
CUT [19]	53.03	0.0248	99.80	0.20
ACL-GAN [20]	41.23	0.0791	**100**	**0**
DCLGAN [21]	58.12	0.0384	99.80	0.20
Proposed	**26.64**	**0.1190**	99.80	0.20

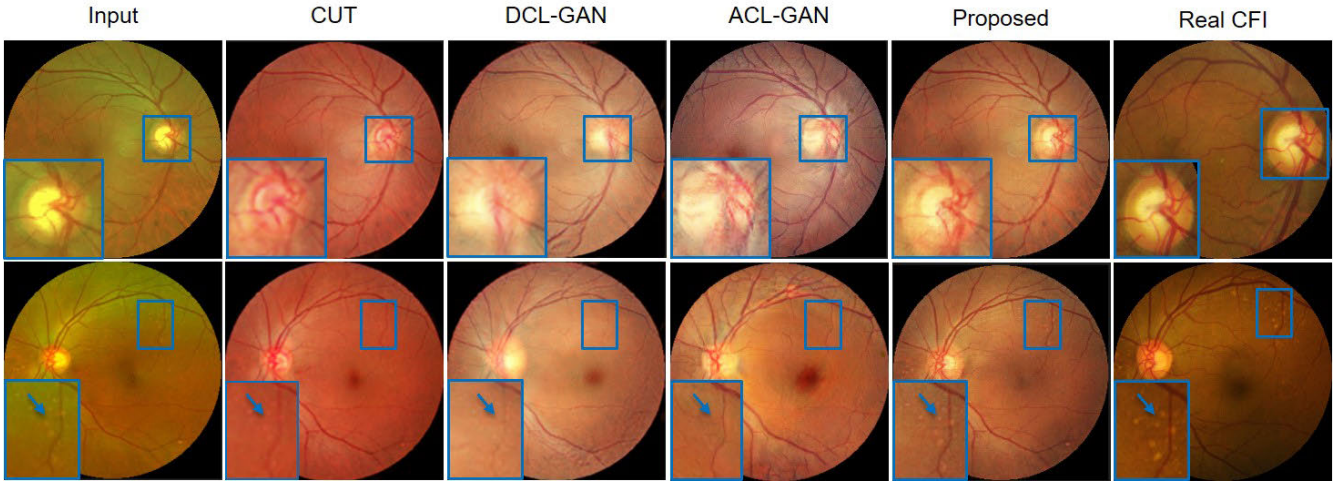

FIGURE 9. Images generated by different image-to-image translation models. The top row shows the failure of CUT, DCL-GAN, and ACL-GAN in generating the optic cup (the brightest yellow part). The bottom row illustrates the disappearance of a part of the drusen in the images generated by these models.

TABLE 3. Evaluation of UFI and generated CFI made by ophthalmologists. A higher value means a better result.

Metric		UFI	Generated CFI
Optic nerve structures	Cup-to-disc ratio	**2.20**	2.11
	Disc color	1.40	**2.58**
Vascular distribution	Overall morphology	2.16	**2.49**
	Vessel contrast	2.05	**2.97**
Drusen	Drusen pattern	2.64	**2.91**
	Drusen number	2.57	**2.86**

has competitive image quality compared to the other techniques. Especially, our model exhibits significant improvements over state-of-the-art methods in terms of FID, with performance gains of 49.76%, 35.39%, 54.16%, and 16.41% compared to CUT, ACLGAN, and DCLGAN respectively. While our model is based on the cycleGAN approach, CUT, ACL-GAN, and DCL-GAN are contrastive learning-based methods, which do not include cycle consistency loss. This loss plays a crucial role in preserving important image features such as lesions and biomarkers through a pixel-by-pixel comparison between the original image and the cycled one. In Figure 9, the first row shows that CUT, ACL-GAN, and DCL-GAN fail to maintain the boundary of the optic cup, a critical feature in glaucoma diagnosis. In the second row, a part of drusen disappears in the images generated by CUT, ACL-GAN, and DCL-GAN, which may result in a wrong diagnosis of age-related macular degeneration. In contrast, our translation model utilizes an attention module to enhance the discriminators' classification ability, thereby generating more realistic images. Our experimental results demonstrate that the attention-aided generative translation model enables the generation of high-quality CFIs from UFIs, which has the potential to improve clinical outcomes in ophthalmology.

2) EVALUATION OF OPHTHALMOLOGISTS AND AGE-RELATED MACULAR DEGENERATION CLASSIFICATION WITH GENERATED CFI

In this section, we demonstrate the clinical applicability of our generated CFI for ophthalmologists and for a computer-aided diagnosis system. First, two ophthalmologists compare how the features represented in UFI and generated CFI are close to those in real CFI (UFI and real CFI are images of the same eye, taken on the same date). Three main measurements in ophthalmology examinations are used for comparison: optic nerve structure, vascular distribution, and drusen. Each measurement contains two sub-measurements: cup-to-disc ratio and color of the disc for optic nerve structures; overall morphology and vessel contrast for vascular distribution; drusen pattern and drusen number for drusen. Each sub-measurement in UFIs and generated CFIs is evaluated independently by two ophthalmologists, who score them as either good (3), moderate (2), or poor (1) based on their similarity to the real CFIs. If there is a discrepancy in the scores, the ophthalmologists review the data again to reach a consensus. The average results of 99 random images are presented in Table 3. Except for the cup-to-disc ratio, the sub-measurements for generated CFI are superior to those of UFI, with the improvement ranging from 10.04% (drusen pattern) to 80.71% (disc color). Particularly, the scores for vessel contrast (2.97) and drusen pattern (2.91) are close to 3, meaning that these features are very similar to those in real CFI. These quantitative results demonstrate the effectiveness of our UFI-to-CFI translation method in enhancing the quality of UFIs, as evaluated by ophthalmologists.

TABLE 4. AMD diagnosis accuracy of UFI, generated CFI, and real CFI (%).

Model	UFI	Generated CFI	Real CFI
Resnet50 [32]	78.67	81.33	**86.67**
Googlenet [33]	76.67	81.00	**84.33**
Efficientnet_B3 [34]	77.00	83.67	**85.67**
MobilenetV3_Large [35]	76.67	80.67	**83.33**

Second, we evaluate the use of our generated CFIs for an automatic diagnosis system, which can support ophthalmologists in their diagnostic decisions. To this end, experiments are conducted to compare the performance of UFI,

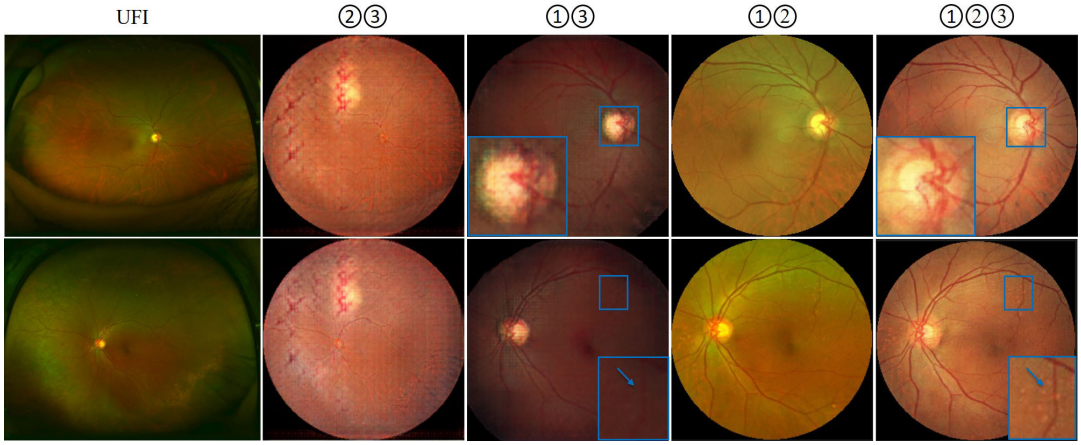

FIGURE 10. Generated images with different combinations of three stages. Without cropping (stage ①), almost all biomarkers and lesions disappear. Without illumination correction and masking (stage ②), the optic cup is not well generated and a part of drusen disappears. Without domain translation (stage ③), the images are still in the UFI domain.

generated CFI, and real CFI for the automatic diagnosis of age-related macular degeneration (AMD), one of the major causes of blindness. Four well-known classification models, namely Resnet50 [32], Googlenet [33], Efficientnet_B3 [34], and MobilenetV3 [35] are selected for comparison. For each network, all parameters were fixed, and then, UFI, generated CFI, and real CFI are used as input to the network in turn. We collect another dataset of 1, 200 pairs of UFI and CFI, with each pair containing one UFI and one CFI of the same eye, and both images having the same label for AMD. The dataset is divided into training and validation sets, both with an equal ratio of AMD to non-AMD images; the training set consists of 900 pairs. Our results demonstrate that the diagnosis accuracy of generated CFI is consistently better than that of UFI across all models tested, with a maximum improvement of 6.67% using Efficientnet_B3. This improvement indicates the potential of our approach for enhancing the accuracy of AMD diagnosis. As the lesions of AMD mostly appear in the surrounding area of the macula, the peripheral area in UFI contains redundant information. Our UFI-to-CFI translation method removes this redundancy through the optic disc-centered cropping stage. Additionally, the color variation of UFI may be misinterpreted as signs of diseases by the classifier, leading to incorrect predictions. Thanks to the generative learning translation in stage 3, the background color is made consistent, allowing the classifier to focus on the lesions and produce more accurate predictions. However, the performance of generated CFI is still inferior to that of real CFI, which is consistent with the evaluation by ophthalmologists. In conclusion, the results from both ophthalmologists and the automatic diagnosis system show that the quality of generated CFI is better than that of UFI but not as good as that of real CFI. This performance gap between generated CFI and real CFI suggests opportunities for future research.

3) ABLATION STUDY

In this section, we conduct experiments to find out a value of α in equation 1 that brings good results. Besides, there are 3 stages in our framework, we will show the contributions of each stage to generated CFI.

TABLE 5. Quantitative result when varying the cropping radius.

α	FID↓	Q_v ↑	Image quality	
			Good&Usable(%)↑	Reject(%)↓
1.5	34.28	0.0662	99.70	0.30
2.0	**26.64**	**0.1090**	99.80	0.20
2.5	27.08	0.1027	**100**	**0**
3.0	29.07	0.0934	99.90	0.10

TABLE 6. Effects of three stages to generated CFIs.

Stages	FID
② ③	106.59
① ③	76.69
① ②	94.63
① ② ③	**26.64**

a: EFFECT OF CROPPING RADIUS TO GENERATED CFI

As mentioned above, the radius for UFI cropping is $\overline{FB} = \overline{FO} + \alpha\overline{OC}$. Different values of α result in different areas covered by generated CFI. We want to figure out the value of α which brings the best result. For this purpose, we perform a grid search with multiple values of α: {1.5, 2, 2.5, 3}, the result is shown in Table 5. It can be seen that with $\alpha = 2$, generated CFI is the most realistic (smallest value of FID) and has the best value of Q_v. Based on this result, we set $\alpha = 2$ for the rest of our experiments.

b: CONTRIBUTIONS OF EACH STAGE TO GENERATED CFI

Our framework contains 3 stages, ablation study is conducted to figure out the contributions of each stage to the generated CFIs. For this purpose, we take turns removing each stage from the framework, the result is reported in Table 6, and the visualization is shown in Figure 10. It can be seen that all stages have huge effects on the output images. Without cropping (stage 1), the huge scale difference between the two domains causes difficulty for the translation model. As a result, almost all biomarkers and lesions disappear in the generated image. If the cropped UFI is not improved

brightness and masked (stage 2) before the translation stage, generated CFI contains faked details near the mask (marked by yellow rectangles) which may hide useful information such as lesions. Stage 3 is the key stage to transfer UFI into the CFI domain, without the translation, the color of UFI remains and is very different from that of CFI. In this comparison, we do not use the metrics Q_v and Image Quality because they are only used for CFI. In Table 6, high values of FID indicate the images are very far from the CFI domain, so Q_v and Image Quality do not make sense for these cases.

V. CONCLUSION AND FUTURE WORK

In this work, we have presented a novel framework for multi-scale multi-modal fundus image translation, addressing the challenges of scale difference and brightness variation between UFI and CFI. Our extensive experiments demonstrate that our proposed method outperforms state-of-the-art approaches, with the majority of the generated CFIs evaluated as high quality. These results indicate the promising potential of our approach for clinical applications, such as automatic disease diagnosis and monitoring, which can reduce patient examination costs and improve clinical outcomes. Furthermore, expert evaluations confirm significant visual quality improvements in the generated CFIs compared to UFIs. However, we acknowledge the limitations of our approach, such as the unsatisfactory optic disc and the remaining artifacts that cause a performance gap between the generated CFI and real CFI. Future work will focus on addressing these remaining issues and exploring the application of generated CFIs for tasks such as vessel segmentation and explainable diagnosis. Moreover, the principles employed in UFI-to-CFI translation, such as addressing scale differences, brightness variations, and image quality enhancement, have relevance in various domains where image transformation and enhancement are critical. While our current investigation focuses on the specific context of fundus images, we recognize the transferability of our methodology to other applications and potential avenues for future research in image translation and enhancement.

REFERENCES

[1] T. K. Yoo, I. H. Ryu, J. K. Kim, I. S. Lee, J. S. Kim, H. K. Kim, and J. Y. Choi, "Deep learning can generate traditional retinal fundus photographs using ultra-widefield images via generative adversarial networks," *Comput. Methods Programs Biomed.*, vol. 197, Dec. 2020, Art. no. 105761.

[2] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2223–2232.

[3] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.

[4] J. Wang, L. Yang, Z. Huo, W. He, and J. Luo, "Multi-label classification of fundus images with EfficientNet," *IEEE Access*, vol. 8, pp. 212499–212508, 2020.

[5] A. Galdran, A. Anjos, J. Dolz, H. Chakor, H. Lombaert, and I. B. Ayed, "State-of-the-art retinal vessel segmentation with minimalistic models," *Sci. Rep.*, vol. 12, no. 1, p. 6174, Apr. 2022.

[6] B. J. Bhatkalkar, S. V. Nayak, S. V. Shenoy, and R. V. Arjunan, "FundusPosNet: A deep learning driven heatmap regression model for the joint localization of optic disc and fovea centers in color fundus images," *IEEE Access*, vol. 9, pp. 159071–159080, 2021.

[7] S. Matsuba, H. Tabuchi, H. Ohsugi, H. Enno, N. Ishitobi, H. Masumoto, and Y. Kiuchi, "Accuracy of ultra-wide-field fundus ophthalmoscopy-assisted deep learning, a machine-learning technology, for detecting age-related macular degeneration," *Int. Ophthalmol.*, vol. 39, no. 6, pp. 1269–1275, Jun. 2019.

[8] L.-P. Cen et al., "Automatic detection of 39 fundus diseases and conditions in retinal photographs using deep neural networks," *Nature Commun.*, vol. 12, no. 1, p. 4828, Aug. 2021.

[9] P. Porwal, S. Pachade, R. Kamble, M. Kokare, G. Deshmukh, V. Sahasrabuddhe, and F. Meriaudeau, "Indian diabetic retinopathy image dataset (IDRiD): A database for diabetic retinopathy screening research," *Data*, vol. 3, no. 3, p. 25, Jul. 2018.

[10] E. Decencière, X. Zhang, G. Cazuguel, B. Lay, B. Cochener, C. Trone, P. Gain, R. Ordonez, P. Massin, and A. Erginay, B. Charton, and J.-C. Klein, "Feedback on a publicly distributed image database: The messidor database," *Image Anal. Stereol.*, vol. 33, no. 3, pp. 231–234, 2014.

[11] M. N. Bajwa, G. A. P. Singh, W. Neumeier, M. I. Malik, A. Dengel, and S. Ahmed, "G1020: A benchmark retinal fundus image dataset for computer-aided glaucoma detection," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2020, pp. 1–7.

[12] W. Zhang, X. Zhao, Y. Chen, J. Zhong, and Z. Yi, "DeepUWF: An automated ultra-wide-field fundus screening system via deep learning," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 8, pp. 2988–2996, Aug. 2021.

[13] L. Ding, A. E. Kuriyan, R. S. Ramchandran, C. C. Wykoff, and G. Sharma, "Weakly-supervised vessel detection in ultra-widefield fundus photography via iterative multi-modal registration and learning," *IEEE Trans. Med. Imag.*, vol. 40, no. 10, pp. 2748–2758, Oct. 2021.

[14] Z. Yang, X. Li, X. He, D. Ding, Y. Wang, F. Dai, and X. Jin, "Joint localization of optic disc and fovea in ultra-widefield fundus images," in *Proc. 10th Int. Workshop Mach. Learn. Med. Imag. (MLMI)*, Shenzhen, China. Cham, Switzerland: Springer, Oct. 2019, pp. 453–460.

[15] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 91–99.

[16] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," 2014, *arXiv:1406.2661*.

[17] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1125–1134.

[18] Y. Wang, Z. Zhang, W. Hao, and C. Song, "Multi-domain image-to-image translation via a unified circular framework," *IEEE Trans. Image Process.*, vol. 30, pp. 670–684, 2021.

[19] T. Park, A. A. Efros, R. Zhang, and J.-Y. Zhu, "Contrastive learning for unpaired image-to-image translation," in *Proc. 16th Eur. Conf. Comput. Vis. (ECCV)*, Glasgow, U.K. Cham, Switzerland: Springer, 2020, Aug. 2020, pp. 319–345.

[20] Y. Zhao, R. Wu, and H. Dong, "Unpaired image-to-image translation using adversarial consistency loss," in *Proc. 16th Eur. Conf. Comput. Vis. (ECCV)*, Glasgow, U.K. Cham, Switzerland: Springer, Aug. 2020, pp. 800–815.

[21] J. Han, M. Shoeiby, L. Petersson, and M. A. Armin, "Dual contrastive learning for unsupervised image-to-image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 746–755.

[22] L. Ju, X. Wang, X. Zhao, P. Bonnington, T. Drummond, and Z. Ge, "Leveraging regular fundus images for training UWF fundus diagnosis models via adversarial learning and pseudo-labeling," *IEEE Trans. Med. Imag.*, vol. 40, no. 10, pp. 2911–2925, Oct. 2021.

[23] M. Monjur, I. T. Hoque, T. Hashem, M. A. Rakib, J. E. Kim, and S. I. Ahamed, "Smartphone based fundus camera for the diagnosis of retinal diseases," *Smart Health*, vol. 19, Mar. 2021, Art. no. 100177.

[24] Q. Zhang, Y. Nie, and W.-S. Zheng, "Dual illumination estimation for robust exposure correction," *Comput. Graph. Forum*, vol. 38, no. 7, pp. 243–252, 2019.

[25] P. J. Burt and E. H. Adelson, "The Laplacian pyramid as a compact image code," in *Readings in Computer Vision*. Amsterdam, The Netherlands: Elsevier, 1987, pp. 671–679.

[26] *Eyepacs Dataset*. Accessed: Jan. 12, 2023. [Online]. Available: https://www.kaggle.com/datasets/mariaherrerot/eyepacspreprocess

[27] *TensorFlow Object Detection API*. Accessed: Jan. 15, 2023. [Online]. Available: https://github.com/tensorflow/models/tree/master/research/object_detection

[28] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 6626–6637.
[29] T. Kohler, A. Budai, M. F. Kraus, J. Odstrcilik, G. Michelson, and J. Hornegger, "Automatic no-reference quality assessment for retinal fundus images using vessel segmentation," in *Proc. 26th IEEE Int. Symp. Comput.-Based Med. Syst.*, Jun. 2013, pp. 95–100.
[30] H. Fu, B. Wang, J. Shen, S. Cui, Y. Xu, J. Liu, and L. Shao, "Evaluation of retinal image quality assessment networks in different color-spaces," in *Proc. 22nd Int. Conf. Med. Image Comput. Comput. Assist. Intervent. (MICCAI)*, Shenzhen, China. Cham, Switzerland: Springer, Oct. 2019, pp. 48–56.
[31] A. Hore and D. Ziou, "Image quality metrics: PSNR vs. SSIM," in *Proc. 20th Int. Conf. Pattern Recognit.*, Aug. 2010, pp. 2366–2369.
[32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[33] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.
[34] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.
[35] A. Howard, M. Sandler, B. Chen, W. Wang, L. Chen, M. Tan, G. Chu, V. Vasudevan, Y. Zhu, R. Pang, H. Adam, and Q. Le, "Searching for MobileNetV3," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1314–1324.

VAN-NGUYEN PHAM received the B.S. degree in electronics and telecommunications from the Hanoi University of Science and Technology, Vietnam, in August 2020. He is currently pursuing the integrated M.S./Ph.D. degree with the Department of Electrical and Computer Engineering, Sungkyunkwan University, South Korea. His current research interest includes AI-based applications for medical images.

DUC-TAI LE received the M.S. degree in computer science from the University of Science, Vietnam National University Ho Chi Minh City, Vietnam, in 2010, and the Ph.D. degree in computer engineering from Sungkyunkwan University, South Korea, in 2016. He was a Postdoctoral Researcher with the Convergence Research Institute, Sungkyunkwan University, from 2016 to 2019. In 2019, he joined the College of Computing in Informatics, Sungkyunkwan University, as a Research Professor. His research interests include wireless ad hoc and sensor networks, intelligent networking, and medical image processing.

JUNGHYUN BUM received the M.S. degree in computer science from Chonnam National University, in 1997, and the Ph.D. degree in computer engineering from Sungkyunkwan University, in 2021. She was with KT for 17 years, where she was involved in IT service and research and development. Since 2016, she has been focusing on big data and artificial intelligence. She is currently a Research Professor with the Sungkyun AI Research Institute. Her research interests include lifelog analysis and visualization as well as applying AI techniques to the medical field.

EUN JUNG LEE received the M.D. degree. She is currently a fellow with the Department of Ophthalmology, Samsung Medical Center, School of Medicine, Sungkyunkwan University, Seoul, South Korea. Her research interests include myopic glaucoma pathogenesis and imaging of lamina cribrosa.

JONG CHUL HAN received the M.D. and Ph.D. degrees from the School of Medicine, Sungkyunkwan University, Seoul, South Korea, in 2000 and 2019, respectively. He is currently an Assistant Professor with the Department of Ophthalmology, Samsung Medical Center, School of Medicine, Sungkyunkwan University. His research interests include myopic glaucoma pathogenesis and imaging of lamina cribrosa.

HYUNSEUNG CHOO (Member, IEEE) received the B.S. degree from Sungkyunkwan University (SKKU), South Korea, in 1988, the M.S. degree from The University of Texas at Dallas, USA, in 1990, and the Ph.D. degree from The University of Texas at Arlington, USA, in 1996. He is currently a Professor with the College of Computing and Informatics, SKKU, and the Director of the ICT Creative Consilience Program supported by the Ministry of Science and ICT (MIST), South Korea. Previously, he was the Director of the Intelligent HCI Convergence Research Center supported by the Ministry of Knowledge Economy, South Korea. He has also served as a Technical Adviser of the Samsung Electronics DMC Research and Development Center (next-generation interaction). He specializes in network softwarization, intelligent mobile computing, multi-access edge computing, and medical image processing, with over 475 publications in international journals and refereed conferences, and 29 international (USA) and 231 domestic patents (South Korea) in the field of mobile and sensor networks with intelligence and autonomicity. He is a member of ACM and IEICE. For his outstanding research, he has received two excellence awards and one commendation award over the years from MIST. He was the Editor-in-Chief of the *Journal of Korean Society for Internet Information* for three years and a Journal Editor of *ACM Transactions on Internet Technology*, *Journal of Communications and Networks*, *Journal of Supercomputing*. He has been the founding Editor of *Transactions on Internet and Information Systems*, since 2010.