# Bridging Synthetic and Real Images: A Transferable and Multiple Consistency Aided Fundus Image Enhancement Framework

Erjian Guo, Huazhu Fu, *Senior Member, IEEE*, Luping Zhou, *Senior Member, IEEE*, and Dong Xu, *Fellow, IEEE*

***Abstract*—Deep learning based image enhancement models have largely improved the readability of fundus images in order to decrease the uncertainty of clinical observations and the risk of misdiagnosis. However, due to the difficulty of acquiring paired real fundus images at different qualities, most existing methods have to adopt synthetic image pairs as training data. The domain shift between the synthetic and the real images inevitably hinders the generalization of such models on clinical data. In this work, we propose an end-to-end optimized teacher-student framework to simultaneously conduct image enhancement and domain adaptation. The student network uses synthetic pairs for supervised enhancement, and regularizes the enhancement model to reduce domain-shift by enforcing teacher-student prediction consistency on the real fundus images without relying on enhanced ground-truth. Moreover, we also propose a novel multi-stage multi-attention guided enhancement network (MAGE-Net) as the backbones of our teacher and student network. Our MAGE-Net utilizes multi-stage enhancement module and retinal structure preservation module to progressively integrate the multi-scale features and simultaneously preserve the retinal structures for better fundus image quality enhancement. Comprehensive experiments on both real and synthetic datasets demonstrate that our framework outperforms the baseline approaches. Moreover, our method also benefits the downstream clinical tasks.**

***Index Terms*—Fundus image, teacher-student model, image enhancement.**

## I. Introduction

RETINAL images are widely used by ophthalmologists or automated image analyzing systems as a non-invasive way to detect and monitor various eye and body diseases [1], such as glaucoma, diabetic retinopathy, and hypertension. Unfortunately, a study of 5,575 patients found that about 12% of fundus images are not of adequate quality to be readable by ophthalmologists [2]. The quality of fundus images varies due to equipment limitations, ophthalmologists' experience, and patient eye movement, which could negatively affect clinical decision making. Image enhancement methods are therefore proposed as a remedy. Traditional fundus image enhancement methods [3], [4], [5], [6] were mainly based on hand-crafted priors, and they could not satisfactorily handle the complexity of varied low-quality cases. To solve this issue, the deep learning methods were proposed to learn more general priors from large amounts of paired low-quality and high-quality images [7], [8], [9], [10], [11], [12], [13], [14]. Therefore, the existing methods resort to either i) synthetic image pairs, such as synthesizing low-quality fundus images by degrading real high-quality ones [7], or ii) unpaired supervision models, such as CycleGAN-like ones [11], [15], for enhancement. However, both approaches have limitations. On one hand, due to the domain shift between the synthetic and the real fundus images, the models trained on synthetic image pairs have limited capability to generalize well to real clinical fundus images. On the other hand, the models trained with unpaired supervision mainly translate image styles and could not well preserve the local details of structures.

To bridge this gap, in this work, we propose a new end-to-end optimized method that simultaneously conducts image enhancement and domain adaptation in one-shot based on the well-known mean teacher framework [16]. By imitating self-supervised learning, mean teacher framework was proposed to be used for unsupervised domain adaptation task in [17]. The domain gap is naturally reduced by the consistency regularization in the mean teacher framework, which enforces the predictions of the teacher network and the student network to be consistent around each unlabeled (target domain) image. Mean teacher aims to learn a smoother domain-invariant function from unlabeled (target domain) images than the model purely trained on labeled (source domain) images. In this paper, we adapt the mean teacher framework to our cross-domain enhancement network through both multi-stage enhancement consistency and multi-level segmentation consistency. Specifically, our method consists of a student network and a teacher network with identical architecture, while the latter is an exponential moving average of the former. The

Manuscript received 7 November 2022; revised 2 February 2023; accepted 12 February 2023. Date of publication 22 February 2023; date of current version 1 August 2023. This work was supported in part by the Huazhu Fu's Agency for Science, Technology and Research (A*STAR) Central Research Fund, in part by Advanced Manufacturing and Engineering (AME) Programmatic Fund under Grant A20H4b0141, in part by the Hong Kong Jockey Club Charities Trust under Grant 2022-0174, and in part by The University of Hong Kong. The work of Luping Zhou was supported by the Australian Research Council (ARC) under Grant DP200103223. *(Corresponding author: Erjian Guo.)*

Erjian Guo and Luping Zhou are with the School of Electrical and Information Engineering, University of Sydney, Camperdown, NSW 2006, Australia (e-mail: eguo9622@uni.sydney.edu.au; luping.zhou@sydney.edu.au).

Huazhu Fu is with the Institute of High Performance Computing (IHPC), Agency for Science, Technology and Research (A*STAR), Singapore 138632 (e-mail: hzfu@ieee.org).

Dong Xu is with the Department of Computer Science, The University of Hong Kong, Pokfulam, Hong Kong (e-mail: dongxu@hku.hk).

Digital Object Identifier 10.1109/TMI.2023.3247783

student network is trained for two tasks. On one hand, it uses synthetic image pairs for supervised enhancement. On the other hand, it uses the unlabeled real images (without the enhanced ground-truth) to regularize the enhancement model trained on the synthetic input, in order to reduce the domain shift. This is achieved by feeding a real image and its augment simultaneously into the student and the teacher networks, respectively, and enforcing consistent predictions between the two networks. Moreover, we also propose a powerful multi-stage Multi-Attention Guided Enhancement Network called MAGE-Net, which also serves as the backbones of the student and the teacher network. Our MAGE-Net is comprised of a multi-stage enhancement (MSE) module and a retina structure preservation (RSP) module. The MSE module consists of a UNet-Shaped stage (Stage-1) to encode broad semantic information and an original-scale stage (Stage-2) to provide spatial details. Multi-type attentions are further employed to guide the enhancement, including our newly proposed fundus attention. Compared with the commonly used skipped connections that directly link encoder-decoder levels, our multi-stage multi-attention architecture provides a more delicate way to effectively integrate multi-scale features. The RSP module is proposed to maintain the vital structures information in fundus images, e.g., the vessels, the optic disc, and the cup, for clinical observation. It sequentially produces essential structure features to guide the enhancement process. Building upon MAGE-Net and the supervised enhancement loss, we further propose multiple consistency losses to bridge the student and teacher networks, including the multi-stage enhancement consistency and the multi-level segmentation consistency of the RSP module.

The contributions of this work are summarized as follows:

1) We propose a new teacher-student based framework with specifically designed consistency losses to reduce the domain shift between the synthetic and the real low-quality fundus images, which is conducted simultaneously with the image enhancement task.
2) We propose a new multi-stage multi-attention guided fundus image enhancement network, which corrects low-quality fundus images while catering for contextual accuracy, spatial accuracy, and anatomical structure accuracy.
3) The experimental results show that our fundus enhancement method also improves the performance of multiple downstream tasks, such as vessel segmentation, optic disc, and cup detection, and disease recognition.

## II. Related Work

In this section, we briefly discuss the fundus image enhancement approaches, domain adaptation methods, and the mean teacher framework, which are related to our work.

### A. Fundus Image Enhancement

Fundus image enhancement methods have two main categories: prior-based methods and learning-based methods.

*1) Prior-Based Methods:* The traditional prior-based methods could not successfully address multiple low-quality cases including noise, blurring, missed focus, illumination, and contrast. Histogram equalization (HE) [18], [19], [20] is a popular method in this category to improve image contrast of retinal images, but the decreasing of gray levels results in the loss of image details. Therefore, negative observations were found for it in many retina image cases, especially in color retinal images. Alternatively, contrast limited adaptive histogram equalization (CLAHE) is also widely adopted to enhance medical images, e.g., Setiawan et al. [5] applied a specifically designed CLAHE to retinal fundus images. However, the CLAHE method may produce artificial boundaries at the region containing an abrupt change in the gray levels. Moreover, although these methods perform efficiently due to their simplicity, their heavy dependence on global image statistics leads to severe image degradation in practice. These hand-crafted priors from human observations do not always work in diverse real-world low-quality retinal images. They tend to suffer from undesirable color and structure distortions.

*2) Learning-Based Methods:* Recently, due to the advantage in image representation, deep learning methods have dominated the computer vision field. There are different types of methods for promoting image quality, such as image enhancement [21], dehazing [22], denoizing [23], and so on. Unfortunately, due to the differences between medical and natural images, the above image correction methods are not suitable to fundus image enhancement which needs specific design to cater for the special characteristic of retinal images. Retinal image enhancement should use pixel-wise translation to preserve retinal structures, which is critical in retinal image analysis. The deep learning networks applied to retinal image enhancement are composed of two categories: synthetic image-pairs-based methods and unpaired-supervision-based methods. The first ones like [24] require high-low quality retinal image pairs to learn a mapping from one representation to another. The widely used fundus degradation method [7] simulates real images of low quality to build retinal image pairs from real high-quality images. However, the image pair-based methods ignore the domain gap between the synthetic low-quality images and the real low-quality images, thus generalizing unsatisfactorily to clinic use. The unpaired supervision-based methods [7], [15] are usually based on CycleGAN-like frameworks to restore fundus images directly from real unpaired images of high or low quality. However, these methods mainly translate image styles to simulate clean results without well preserving the important details of fundus structures. To this end, vessel segmentation was employed as a useful way to enhance retinal structures. For example, CofeNet [7], a method using synthetic image pairs, was designed to preserve the retinal structures in fundus enhancement process through benefiting from vessel segmentation outputs. Recently, Transformer-based methods have achieved great success in high-level vision tasks [25], such as image classification [26], semantic segmentation [27], object detection [28], etc. Due to the advantage of capturing long-range dependencies and good performance in many high-level vision tasks, Transformer has also been introduced into low-level vision tasks, such as image restoration [29], [30]. The transformer-based method is rarely applied in the fundus image enhancement task. The transformer-based method: RFormer [11] relies on an in-house Real Fundus (RF) dataset including 120 paired high- and low-quality real fundus images to learn to synthesize high-quality images from low-quality ones. Unlike other fundus image enhancement methods using synthetic low-quality images for training, RFormer is directly trained based on paired real fundus images of different qualities, unfortunately, are costly to be collected in practice. Moreover, the RF dataset has not been publicly released.

In this paper, to alleviate the issues from both the paired- and the unpaired-supervision-based methods and integrate their advantages, we develop a fundus image enhancement framework that learns feature presentations from synthetic image pairs and leverages real low-quality images to improve enhancement performance. This further calls for domain adaptation to alleviate the discrepancy between the synthetic and the real image domains involved in the fundus image enhancement task.

### B. Domain Adaptation

Due to the gap between the source and the target domains, a model trained on the source domain may suffer significant performance drops on the target domain in practice. Domain adaptation is therefore proposed to bridge the domain gap so that the model learned from the source domain is able to perform decently on the target domain. There is a large corporation of literature tackling the problem of domain adaptation. We focus on deep learning based methods as these are most relevant to our work. Unsupervised domain adaptation could be addressed from different perspectives. Discrepancy-based methods guide the feature learning by minimizing the domain gap with Maximum Mean Discrepancy [31], while the works in [32] and [33] estimate the domain confusion by learning a domain discriminator. Differently, self-ensembling [17] extended the mean teacher framework [16] to reduce domain gap and established several cross-domain benchmarks for recognition task. Recently, Mean Teacher has been extensively used as a transfer learning method for various tasks, e.g., image dehazing [34], object detection [35] and semantic segmentation [36]. For example, to reduce the domain gap in image dehazing, Liu et al. [34] developed a disentangle-consistency mean-teacher network (DMT-Net) collaborating with unlabeled real-world hazy images to address the domain shift problem. Similar to [34], our method aims to leverage additional real low-quality fundus images without ground-truth to alleviate the domain discrepancy between the synthetic and the real images. Along this line, we explore the Mean Teacher framework to bridge the domain gap by imposing consistency regularization in fundus image enhancement, which has not been previously explored in this field. Moreover, the enhancement loss and the segmentation loss between synthetic to real fundus image is elegantly integrated into the Mean Teacher paradigm to boost cross-domain enhancement results.

### C. Mean Teacher Framework

Mean Teacher framework [16] is widely used in semi-supervised learning. The main idea of mean teacher is to enforce the predictions of the teacher and the student networks consistent under small perturbations of the input or the network parameters. Mean teacher consists of two networks with the same architecture: let $S(\cdot)$ and $T(\cdot)$ represent the embedding functions of the student network with weight $w_s$ and the teacher network with weight $w_t$, respectively. Let us denote a labeled data as $\mathbf{I}_l$, an unlabeled data as $\mathbf{I}_u$ and its augment as $\tilde{\mathbf{I}}_u$. The consistency loss penalizes the difference between the student's prediction $S(\mathbf{I}_u)$ and the teacher's prediction $T(\tilde{\mathbf{I}}_u)$, which is typically computed as the Mean Squared Error:

$$\mathcal{L}_{cons}(\mathbf{I}_u) = \|S(\mathbf{I}_u; w_s) - T(\tilde{\mathbf{I}}_u; w_t)\|_2^2. \tag{1}$$

The student network is trained by using gradient descent, and the weights $w_t$ of the teacher network at the $n$-th iteration are the exponential moving average of the student weights $w_s$:

$$w_t^n = \alpha \cdot w_t^{n-1} + (1-\alpha) \cdot w_s^{n-1}, \tag{2}$$

where $\alpha$ is a smoothing coefficient parameter that controls the updating of the teacher's weights. The total loss of the mean teacher framework is a combination of the supervised losses on labeled data and the consistency losses on unlabeled data, balanced with the trade-off parameter $\mu$:

$$\mathcal{L} = \sum_{i=1}^{M} \mathcal{L}_{super}(\mathbf{I}_l^i) + \mu \sum_{j=1}^{N} \mathcal{L}_{cons}(\mathbf{I}_u^j), \tag{3}$$

where $M$ and $N$ denote the total number of the labeled and unlabeled images, respectively.

## III. Methodology

The overview of our proposed teacher-student framework is shown in Fig. 1. It consists of a student network and a teacher network, both built on our proposed MAGE-Net. In order to integrate synthetic and real images, we also design specific consistency losses to reduce the domain shift. The detailed structure of MAGE-Net is shown in Fig. 2a. It is composed of two modules: multi-stage enhancement module (Stage-1 and Stage-2) and retinal structure preservation module (RSP). These two modules are effectively fused through a newly designed fundus attention block (See Fig. 2c).

### A. Multi-Attention Guided Enhancement Network (MAGE-Net)

*1) Multi-Stage Enhancement (MSE) Module:* The MSE module (Fig. 2a) consists of two stages to restore clean fundus images. Stage-1 employs encoder-decoder structure with large receptive fields to extract the contextualized features in the fundus images. However, the downsampling operations in Stage-1 loss spatial details and thus yield over-smoothed results. Therefore, Stage-2 is proposed for three reasons: preserving local image details by operating on the original image resolution, maintaining the anatomical structure by adding the feature from the RSP module, and fusing features through the fundus attention block effectively. Given a low-quality input image, we feed it into each enhancement stage. To enrich the features from Stage-2, we also adopt a multi-patch hierarchy strategy on the input image from Stage-1. We split each input

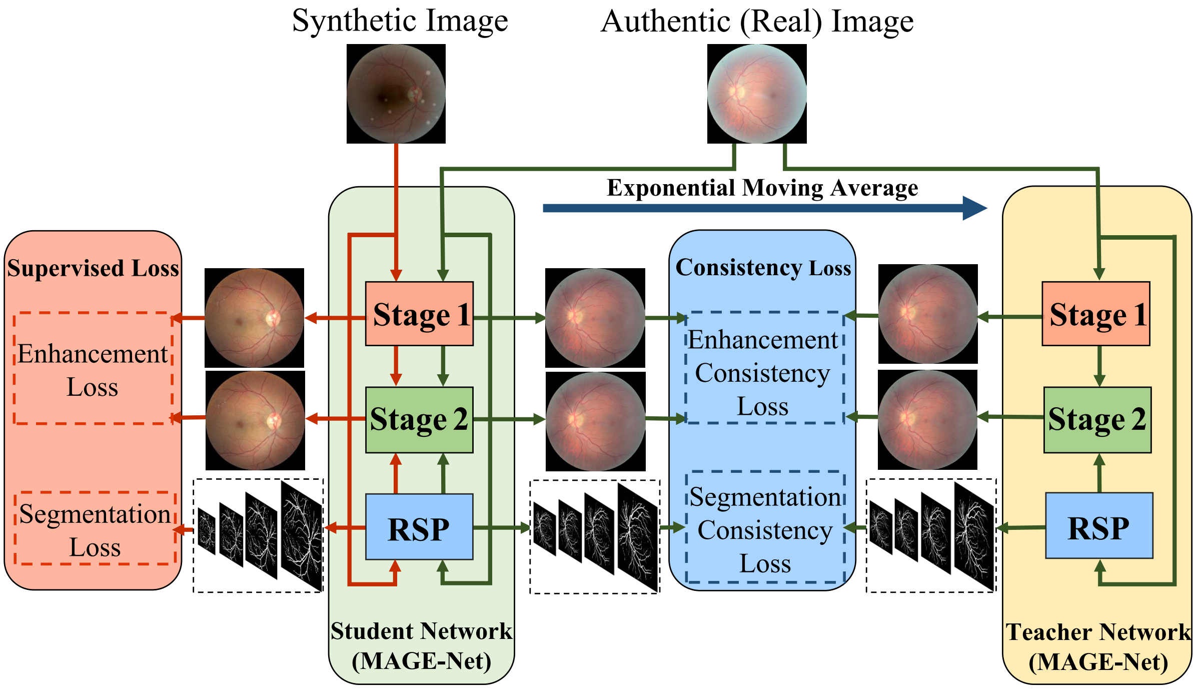

Synthetic Image
Authentic (Real) Image
Exponential Moving Average
Supervised Loss
Enhancement Loss
Segmentation Loss
Stage 1
Stage 2
RSP
Student Network (MAGE-Net)
Consistency Loss
Enhancement Consistency Loss
Segmentation Consistency Loss
Stage 1
Stage 2
RSP
Teacher Network (MAGE-Net)

Fig. 1. **Overview of Transferred MAGE-Net with Multi-Stage Consistency (T-MAGE-Net).** We use MAGE-Net as a teacher network and a student network, respectively. Each paired synthetic image is fed into student model to conduct the supervised learning of image enhancement and segmentation. We calculate supervised multi-stage enhancement losses and supervised segmentation losses in the student network. Each unlabeled real image is firstly transformed into two perturbed samples by adding Gaussian noise and then we inject the two perturbed samples into student and teacher models separately. Two types consistency regularization are devised to facilitate reducing the domain gap in mean teacher paradigm: 1) Multi-Stage Enhancement Consistency Loss to align the clean predictions between teacher and student; 2) Multi-Stage Segmentation consistency Loss for matching the retinal structures between teacher and student. The whole T-MAGE-Net is trained by minimizing the supervised losses on paired synthetic image data plus the two consistency losses on the unlabeled real image in an end-to-end manner. Note that the student network is optimized with Adam and the weights of teacher network are the exponential moving average of student model weights.

image into two non-overlapping subimages with 50% of the original resolution. Then, we feed them into Stage-2.

In Stage-1, we first utilize a convolution layer and a channel attention block (CAB) [37] (Fig. 2b) to extract the features from the input. Specifically, the CAB generates different attention maps for each channel-wise feature, making the network focus on more informative features. Then a UNet-shaped [38] architecture is adopted as our sub-network to restore low-quality fundus images. Each encoder layer employs CABs to extract high-level semantic features, and each decoder layer uses a bilinear upsampling operation followed by a convolution layer instead of using transposed convolution due to the checkerboard artifacts introduced by transposed convolution [39]. Both the encoder and decoder features are resized and fused with the intermediate feature maps of Stage-2. Moreover, the output of the decoder is sent through the supervised attention (SAM) module (Fig. 2d) [40] to provide the attention maps to Stage-2 with the aid of supervision information from the ground-truth high-quality fundus images.

In Stage-2, we employ a residual network. The features of the original image is firstly extracted by a convolution layer and a CAB like in Stage-1, and then sent to a series of our newly proposed blocks: fundus attention block (FAB) (Fig. 2c). To preserve local image details, the FAB keeps the feature maps at the original image resolution and does not employ any downsampling operation. In each FAB, we first extract high-resolution features by using several CABs. Then, we add the resized contextual features from Stage-1 with those from Stage-2 to refine the feature maps of whole images. To boost the performance of the downstream clinical analysis tasks, the feature maps from the RSP module are resized and concatenated with the feature maps from Stage-2. Finally, the fused feature maps are passed through a learnable non-linear transformation filter to maintain stable and efficient model performance. We sequentially employ three FABs for feature extraction and fusion to generate the residual and then add it with the low-quality input image to produce the final enhancement output.

*2) Retinal Structure Preservation (RSP) Module:* Nature image enhancement methods focus on producing visually satisfying results for humans, without necessarily preserving valuable clinical information in the reconstructed images. However, as important clinical diagnosis evidence, the enhanced fundus images should preserve the retinal structures without incorrectly synthesizing the content. Therefore, we propose the RSP module to further guide image enhancement in the MSE module. As illustrated in Fig. 2a, the RSP module is based on the pre-trained AG-Net [41], which is a UNet-shaped segmentation network with its layers respectively supervised by multi-scale segmentation masks. The main retinal structures are encoded as an attention feature map by the decoder at each scale layer. These multi-scale feature maps from different decoder layers are then fused with the MSE module after being scaled up to the original image size. In this

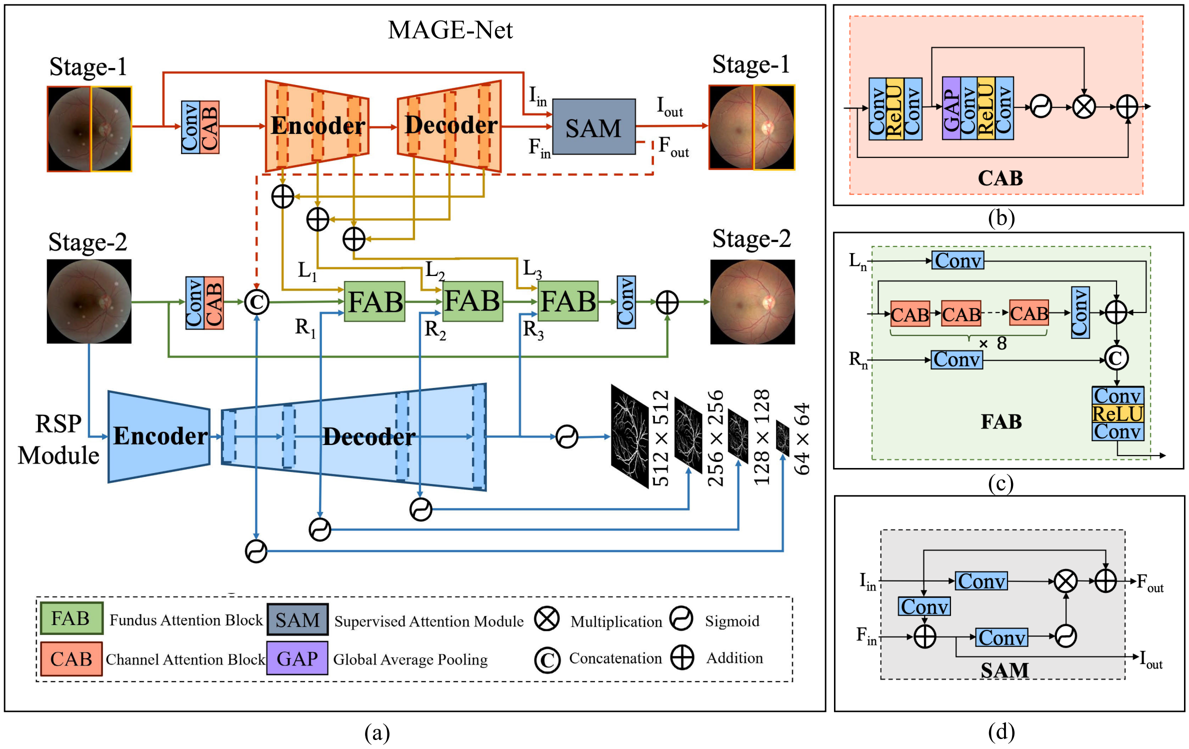


Fig. 2. **Illustration of our proposed Multi-Attention Guided Network (MAGE-Net).** MAGE-Net consists of two parts: multi-stage enhancement module (MSE) and retinal structure preservation module (RSP). For both teacher and student networks, given an input image, we simultaneously feed it into the RSP module and the two stages of the MSE module. The RSP module sends the features maps of important retinal structures to correct stage-2 of the MSE module at each FAB. The stage-1 of the MSE module extracts the contextualized features by a UNet-shaped network. Then, both the encoder-decoder and SAM features are fused into stage-2 from stage-1. (a) Multi-Attention Guided Network (MAGE-Net). (b) Channel Attention Block. (c) Fundus Attention Block. (d) Supervised Attention Module.

way, the important components of fundus images are injected into the image correction network to preserve the clinically useful contents.

*3) MAGE-Net Loss:* Our MAGE-Net is supervised by an enhancement loss and a segmentation loss. Given a labeled synthetic low-quality image $\mathbf{I}_l$, denoting its enhancement output at the $s$-th stage as $\mathbf{I}_e^s$ and the ground-truth high-quality image as $\mathbf{I}_h$, the enhancement loss at the $s$-th stage is the sum of the Charbonnier loss [42]:

$$\mathcal{L}_{char}^s(\mathbf{I}_l) = \sqrt{\|\mathbf{I}_e^s - \mathbf{I}_h\|^2 + \varepsilon^2} \tag{4}$$

and the Edge loss [40]:

$$\mathcal{L}_{edge}^s(\mathbf{I}_l) = \sqrt{\|\Delta(\mathbf{I}_e^s) - \Delta(\mathbf{I}_h)\|^2 + \varepsilon^2}, \tag{5}$$

where $\varepsilon$ is set as 0.001 in both loss functions, and $\Delta(\cdot)$ is the gradient function. Denoting the segmentation result of $\mathbf{I}_l$ at the $v$-th scale in the RSP module as $\mathbf{I}_{seg}^v$ and the ground-truth mask as $\mathbf{G}_{seg}^v$, the segmentation loss at the $v$-th scale is calculated as:

$$\mathcal{L}_{seg}^v(\mathbf{I}_l) = \|\mathbf{I}_{seg}^v - \mathbf{G}_{seg}^v\|_2 \tag{6}$$

for each of the four scales in the RSP module. The overall supervised loss of our MAGE-Net is provided below, where the trade-off coefficient $\lambda$ is set as 0.5:

$$\mathcal{L}_{mage}(\mathbf{I}_l) = \sum_{s=1}^{2}(\mathcal{L}_{char}^s(\mathbf{I}_l) + \mathcal{L}_{edge}^s(\mathbf{I}_l)) + \lambda\sum_{v=1}^{4}\mathcal{L}_{seg}^v(\mathbf{I}_l), \tag{7}$$

where $s$ denotes the stage index of MSE module, and $v$ denotes the scale index of the RSP module.

## B. Transferable MAGE-Net With Multiple Consistency Losses

Fig. 1 depicts the overall architecture of our proposed Transferable MAGE-Net (T-MAGE-Net) with multiple consistency losses, which corrects the low-quality fundus images for clinical observation and leverages real images for fundus images enhancement. We use the MAGE-Net for both the teacher and student networks. During the training process, we feed the synthetic image pairs into the student network and compute the supervised loss of our MAGE-Net. Meanwhile, for each unlabeled real image without the corresponding enhanced ground-truth, we create an auxiliary image from it by adding Gaussian noise, separately feeding them into the student and teacher networks, and enforce the consistent prediction results from the two networks. The consistency loss is computed for the unlabeled real images at each enhancement stage and each scale of the RSP decoder outputs between the teacher and student networks. The whole architecture is then optimized with two consistency regularizations: 1) each stage of the enhancement results consistency to align the clean outputs predictions between teacher and student, 2) each scale of the segmentation consistency for matching and preserving the retinal structure between teacher and student networks. The weights of the teacher model are updated by the exponential moving average weights of the student model [16], which will not significantly increase burden to our MAGE-Net becauses of shared weights between teacher and student models. The employment of a teacher-student framework does not introduce additional parameters to learn as the teacher is simply the exponential moving average of the student [43].

*1) Consistency Loss:* Denote an unlabeled real image as $\mathbf{I}_u$ and its augment as $\tilde{\mathbf{I}}_u$, and let $S(\cdot)$ and $T(\cdot)$ represent the embedding functions of the student and the teacher networks, respectively. We enforce the two networks to output consistent enhancement results ($S_e(\mathbf{I}_u)$ and $T_e(\tilde{\mathbf{I}}_u)$) of each stages and segmentation results($S_{seg}(\mathbf{I}_u)$ and $T_{seg}(\tilde{\mathbf{I}}_u)$) of each segmentation scale. The overall consistency loss $\mathcal{L}_{cons}(\mathbf{I}_u)$ sums over the multi-stage enhancement consistency loss and the multi-level segmentation consistency losses:

$$\mathcal{L}_{cons}(\mathbf{I}_u) = \sum_{s=1}^{2} \|S_e^s(\mathbf{I}_u) - T_e^s(\tilde{\mathbf{I}}_u)\|_1 + \sum_{v=1}^{4} \|S_{seg}^v(\mathbf{I}_u) - T_{seg}^v(\tilde{\mathbf{I}}_u)\|_1. \quad (8)$$

where $s$ denotes the stage index of MSE module, and $v$ denotes the scale index of the RSP module.

*2) Total Loss:* The total loss of our method is the sum of the supervised loss from our MAGE-Net and the unsupervised multi-stage multi-level consistency loss:

$$\mathcal{L} = \sum_{i=1}^{M} \mathcal{L}_{mage}(\mathbf{I}_l^i) + \mu \sum_{j=1}^{N} \mathcal{L}_{cons}(\mathbf{I}_u^j), \quad (9)$$

where $M$ and $N$ denote the total number of the labeled and unlabeled images, respectively. The weight $\mu$ is computed by a time-dependent Gaussian warming up function [16]. The parameters of teacher network are updated by the exponential moving average (EMA) strategy in each training iteration.

# IV. Experiments

## A. Implementation Details

Our method is implemented by PyTorch, and trained on a single NVIDIA RTX V6000 GPU. The Adam optimizer is adopted. The initial learning rate is $2 \times 10^{-5}$, which is decreased to $1 \times 10^{-7}$ by the cosine annealing strategy [44]. All of the labeled and unlabeled images are re-scaled to the size of 512 × 512 . The mini-batch size is 24, including 16 labeled synthetic images and 8 unlabeled real images. We use a two-stage training strategy. In order to accelerate our training process, we pretrain the RSP module, and then train the whole enhancement framework in an end-to-end fashion.

## B. Datasets

Our training set is formed from the EyeQ [45] dataset, whose images are captured by various cameras from different hospitals. From this perspective, our model is generally trained to enhance images from different centers or equipments [46]. The EyeQ [45] dataset is a subset of the Kaggle [47] dataset for fundus image quality assessment, which has 28,792 retinal images with three quality grades (“Good”, “Usable”, and “Reject”). Specifically, we select 10,000 high-quality fundus images (labeled as “Good”) as the clean images and produce segmentation masks by using pre-trained AG-Net [41]. We randomly choose degradation factors (e.g., light transmission disturbance, image blurring, and retinal artifacts) to synthesize degraded images by using the method [7]. Moreover, we randomly select 5,000 low-quality images (labeled as “Uable”) as the unlabeled data. In the test stage, to evaluate the quality of image enhancement, we also utilize the degradation model [7] to randomly generate degraded images on the DRIVE [48] test set and REFUGE [49] test set. Moreover, For the Subtest-EyeQ dataset, we chose another 500 images which are labeled as “Good” in EyeQ but not present in Subtrain-EyeQ dataset to evaluate the image enhancement quality.

## C. Degradation Model Settings

The degradation method is based on the ophthalmoscope imaging systems, which is also verified by Cofe-Net [7]. Clinical image collection in a complex environment using an ophthalmoscope often encounters several types of interference, as introduced in the optical feed-forward system. Light transmission disturbance is often caused by exposure issues. Due to the interspace between the eye and camera, stray light may enter into the ophthalmoscope, mix with the lighting source and result in uneven exposure. This also affects the tuning setting of the programmed exposure, leading to global over-/under-exposure. In addition, image blurring caused by human factors (such as eyeball movement, fluttering, and defocus) results in low-quality images. Besides, the capturing of undesired objects (e.g., dust) during imaging is also a crucial factor that reduces image quality and impedes subsequent diagnosis. Therefore, Cofe-Net [7] proposes a reformulated representation of the interference that occurs during the collection of fundus images. The degradation model could be used to not only support current fundus propagation models, but also synthesize a high-quality pairwise fundus dataset for subsequent research. Cofe-Net [7] summarized the interference in terms of three factors, including light transmission disturbance, image blurring, and retinal artifacts. Thanks to Cofe-Net [7], the degradation method was directly adopted in subsequent works, such as I-SECRET [24]. We put a high-quality fundus image $\mathbf{x}$ into the degradation model to get the paired degraded image $x'$.

*1) Light Transmission Disturbance:* The light transmission disturbance contains two types of degraded factors: global factors and local factors. The global factors include contrast factor, brightness, and saturation, which are caused by unstable stray light, subjective situation, and manual mydriasis. The local factors produce additional non-uniform illumination due to the initiative light leak phenomenon, diverse lens apertures, and embedded optical compensation mechanism. Therefore, light transmission disturbance is simulated by using:

$$\mathbf{x}' = clip(\alpha(\mathbf{J} \cdot G_L(r_L, \sigma_L) + \mathbf{x}) + \beta; s), \quad (10)$$

where $\alpha$, $\beta$, and $s$ refer to the contrast factor, brightness, and saturation, respectively. To simulate global factors, we randomly set them between $-0.5$ to $0.5$. $G_L$ is a Gaussian filter with the radius $r_L$ and the variance $\sigma_L$. For local factors, an illumination bias $\mathbf{J}$ is defined as:

$$\mathbf{J}_{ij} = n_l \mid_{(i-a)^2+(j-b)^2<r_L^2}, \quad (11)$$

where $c = (a, b)$ is the center with the radius of $r_L$. We randomly set $c \in [0.375r_L, 0.625r_L]$. We define $r_L \in [0.75w, w]$; $\sigma_L \in [0.66cr_L, 0.66(w - c)r_L]$ and $r_L \in [0.3w, 0.5w]$; $\sigma_L \in [0.55r_L, 0.75r_L]$ for light leak phenomenon and uneven exposure problem, respectively, where $w$ denotes the image size.

*2) Image Blurring:* The image blurring is caused by undesired object distance in funduscopy. It is simulated by using:

$$\mathbf{x}' = \mathbf{x} \cdot G_B(r_B, \sigma_B) + n, \tag{12}$$

where $G_B$ is a Gaussian filter with a radius $r_B$ and the spatial constant $\sigma_B$, and $n$ denotes the additive random Gaussian noise. Here we set $\sigma_B = 0.03w$, and $r_B \in [0.01w, 0.015w]$.

*3) Retinal Artifact:* The retinal artifacts are caused by dust and grains attaching on the lens of the imaging plane. It is simulated by using:

$$\mathbf{x}' = \mathbf{x} + \sum_{k}^{K} G_R(r_k/4, \sigma_k) \cdot \mathbf{o_k}, \tag{13}$$

where $K$ is the undesired object number. To simulate the interference in real clinical scenarios, we randomly increase the number of undesired objects from 10 to 30. For each undesired object $k$, $r_k$ and $\sigma_k$ are defined as the radius and the variance of a Gaussian filter $G_R$. We randomly set the radius $r_k \in [0.025w, 0.05w]$, the variance $\sigma_k = 5 + 0.8r_k$, and the illumination bias $\mathbf{o_k} = 1 - e^{-(0.5+0.04r_k)\times(0.012r_k)}$ for each object $k$.

## V. Evaluation

In this section, we evaluate the enhancement performance of different methods in terms of the image quality and three downstream tasks including vessel segmentation task, optic disc/cup detection task and real clinical image analysis task.

### A. Image Quality Enhancement

For quantitative evaluation, we use both PSNR and SSIM as the evaluation metrics in Table I. Our method is the best performer in terms of both PSNR and SSIM [51]. Specifically, among all comparing methods, the baseline method proposed by Setiawan et al. [5] and the DCP [50] method correct each fundus image based on the global image statistics and functions, so their results contain undesired distortion; The relatively poor work results from both methods Setiawan et al. [5] and DCP [50] are expected since they are non-deep learning methods. The method from Setiawan et al. [5] exploits the image contrast normalization and contrast limited adaptive histogram equalization (CLAHE) techniques to restore the color of retinal images. Instead of simply considering the color and texture information like Setiawan et al. [5], the DCP [50] method decomposes the reflection and illumination, which achieves image enhancement and correction by estimating the solution through an alternative minimization scheme. While these algorithms based on the bottom-up frameworks are effective, their optimal solutions rely heavily on global image statistics and mapping functions, namely, these methods ignore discriminative features, which may introduce undesired artifacts and distortion. StillGAN [15], a CycleGAN-like method, is trained with unpaired supervision that could not well preserve the local structure details; Cofe-Net [7] model trained on synthetic image pairs ignores the gap between the synthetic and the real low-quality images for the practical diagnosis, so it is not well generalized to the authentic clinical fundus images. Although utilizing unlabelled real fundus images, I-SECRET [24] method still loses to ours as it only uses a single CNN-based network to enhance the fundus images without well preserving the retinal and lesion details. Moreover, by cross-referencing Table VI, it is found that our method still wins all the baseline methods even without using the RSP module (i.e., without the additional guidance from the extra segmentation masks), since our method without RSP (i.e., "Ours w/o RSP") achieves PSNR/SSIM of 24.47/0.9273, better than the best baseline I-SECRET (24.09/0.906) shown in Table I. This may suggest that our mean teacher based domain adaptation more effectively utilizes the unlabelled real low-quality images, compared with the contrastive learning employed by I-SECRET.

TABLE I

Image Enhancement Quality Comparison of Different Methods on DRIVE [48], REFUGE [49], and Subtest-EyeQ

| | DRIVE [48] | | REFUGE [49] | | Subtest-EyeQ | |
|---|---|---|---|---|---|---|
| **Methods** | **PSNR** | **SSIM** | **PSNR** | **SSIM** | **PSNR** | **SSIM** |
| Low-quality | 16.28 | 0.804 | 16.79 | 0.823 | 18.49 | 0.811 |
| Setiawan [5] | 16.68 | 0.680 | 16.65 | 0.668 | 18.59 | 0.770 |
| DCP [50] | 17.55 | 0.792 | 14.66 | 0.702 | 17.98 | 0.786 |
| Cofe-Net [7] | 20.31 | 0.881 | 24.45 | 0.897 | 21.88 | 0.880 |
| StillGAN [15] | 22.48 | 0.890 | 22.90 | 0.871 | 22.77 | 0.850 |
| I-SECRET [24] | 24.09 | 0.906 | 25.04 | 0.914 | 23.53 | 0.889 |
| **T-MAGE-Net** | **24.72** | **0.928** | **25.66** | **0.929** | **24.14** | **0.896** |

In addition, visual comparisons of different enhancement results are given in Fig. 3 for synthetic low-quality images and in Fig. 4 and Fig. 5 for real low-quality images without ground-truth. From Fig. 3, we can see that neither StillGAN nor Cofe-Net could eliminate the undesired light spot (indicated by the blue arrows) presented in the synthetic low-quality image. This problem is alleviated in the images enhanced by I-SECRET, as it considers both labeled and unlabeled retina images to learn robust features. However, compared with our method, I-SECRET generates blurred vessel boundaries (indicated by the green arrows), confirming the importance of introducing RSP module. Our advantage over I-SECRET could also be observed by the enhancement result from a real low-quality image in Fig. 4. Compared with I-SECRET, our method produces a sharper image with much less undesired light spots. Visual comparison with more methods on real low-quality images is given in Fig. 5. Since there is no ground-truth for the enhancement, we additionally show the vessel segmentation from the enhanced images. As shown, our method could recover finer vessels than other methods. We compared the number of training parameters between the proposed T-MAGE-Net and other baseline approaches, as shown in Table II. It could be found that the numbers of parameters of Cofe-Net [7] and StillGAN [15] are much more than that of our method. Ours is just a bit more than that of I-SECRET [24], but ours achieves better performance.

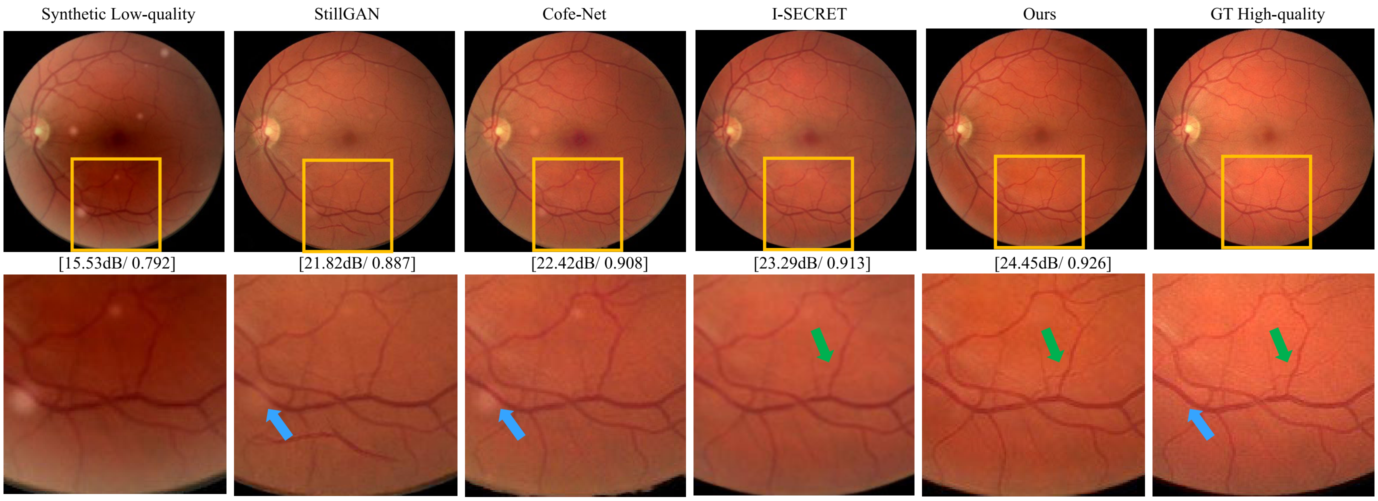


Fig. 3. **Visual comparison of enhancement results (Columns 2-5) on a synthetic low-quality fundus image (Column 1**). The ground-truth (GT) high-quality image is given in Column 6. The symbol [. / .] denotes [PSNR / SSIM] scores. The bottom row contains the zoom-in views of the images in the top row. Arrows point to the visual differences for attention.

TABLE II
COMPARISON OF DIFFERENT METHODS IN TERMS OF THE NUMBER OF PARAMETERS

| Methods | Number of Parameters (M) |
|---|---|
| Cofe-Net [7] | 41.218 |
| StillGAN [15] | 78.644 |
| I-SECRET [24] | 11.756 |
| T-MAGE-Net(ours) | 26.379 |

## B. Vessel Segmentation

The enhancement quality is also validated through the downstream task of vessel segmentation. We evaluate different methods on the degraded DRIVE [48] test set. Quantitative results are shown in Table I. Our method outperforms the competing methods in terms of AUC, Accuracy (Acc.), and IoU. We use CE-Net [52] trained on DRIVE training set as the segmentation method, which achieves AUC/Acc/IoU of 0.953/0.982/0.830 on high-quality DRIVE [48] test set. The vessel segmentation results of real low-quality fundus images from different deep learning methods are shown in Fig. 5. Obviously, the vessel structure is better preserved by using our enhancement method. In contrast, the baseline method by Setiawan et al. [5] and DCP [50] produce unsatisfactory results, because their solutions highly rely on global image contrast and illumination. They cannot correct local light spots, holes, and halos, which influences the vessel observation. Other baseline methods, like StillGAN [15] and I-SECRET [24], use different supervised losses to preserve the structure information of the whole images, but the useful retinal structures for clinic diagnosis are not emphasized during the process, so that the vessel details are missed in Fig. 5. Differently, Cofe-Net [7] designed a retinal structure activation module to emphasize the anatomical retinal structures. Comparing with it, our RSP module together with a multi-resolution method can provide more effective and robust structural features. Moreover, using the restored images from our MAGE-Net without RSP for vessel segmentation, we achieve the results of AUC/Acc/IoU as 0.910/0.970/0.726, better than I-SECRET's 0.877/0.963/0.662 as reported in Table III. The results indicate that our improvements come from both the RSP module to preserve retinal structure and the MSE module to achieve clear images. Therefore, our enhanced images with better quality also lead to the smallest error of vessel segmentation over those produced by other baseline methods.

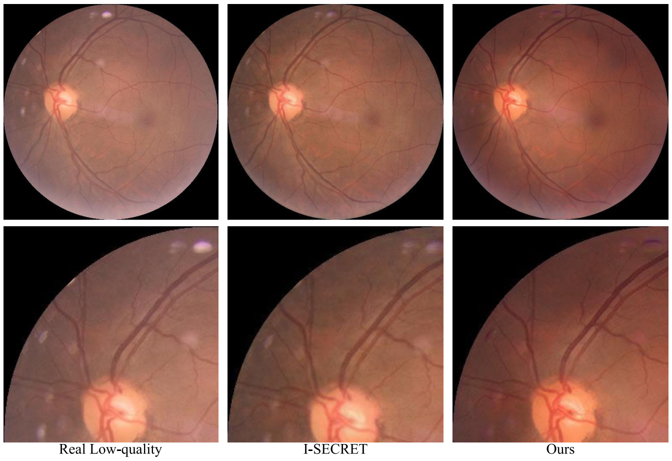


Fig. 4. Visual comparison between I-SECRET [24] and our method by enhancement from a real low-quality fundus image. The top row shows the real low-quality fundus image and the enhancement results. The bottom row shows the zoom-in views.

TABLE III
VESSEL SEGMENTATION PERFORMANCE COMPARISON OF DIFFERENT METHODS ON DRIVE [48] DATASET

| Methods | AUC | Acc | IoU |
|---|---|---|---|
| Low-quality | 0.781 | 0.938 | 0.479 |
| high-quality | 0.953 | 0.982 | 0.830 |
| Setiawan [5] | 0.809 | 0.938 | 0.504 |
| DCP [50] | 0.813 | 0.948 | 0.547 |
| Cofe-Net [7] | 0.875 | 0.961 | 0.654 |
| StillGAN [15] | 0.867 | 0.959 | 0.639 |
| I-SECRET [24] | 0.877 | 0.963 | 0.662 |
| MAGE-Net(ours) | 0.910 | 0.970 | 0.726 |
| **T-MAGE-Net(ours)** | **0.932** | **0.974** | **0.764** |

## C. Optic Disc/Cup Detection

Optic disc and cup detection is important for diagnosing glaucoma. For the downstream task of optic disc/cup detection,

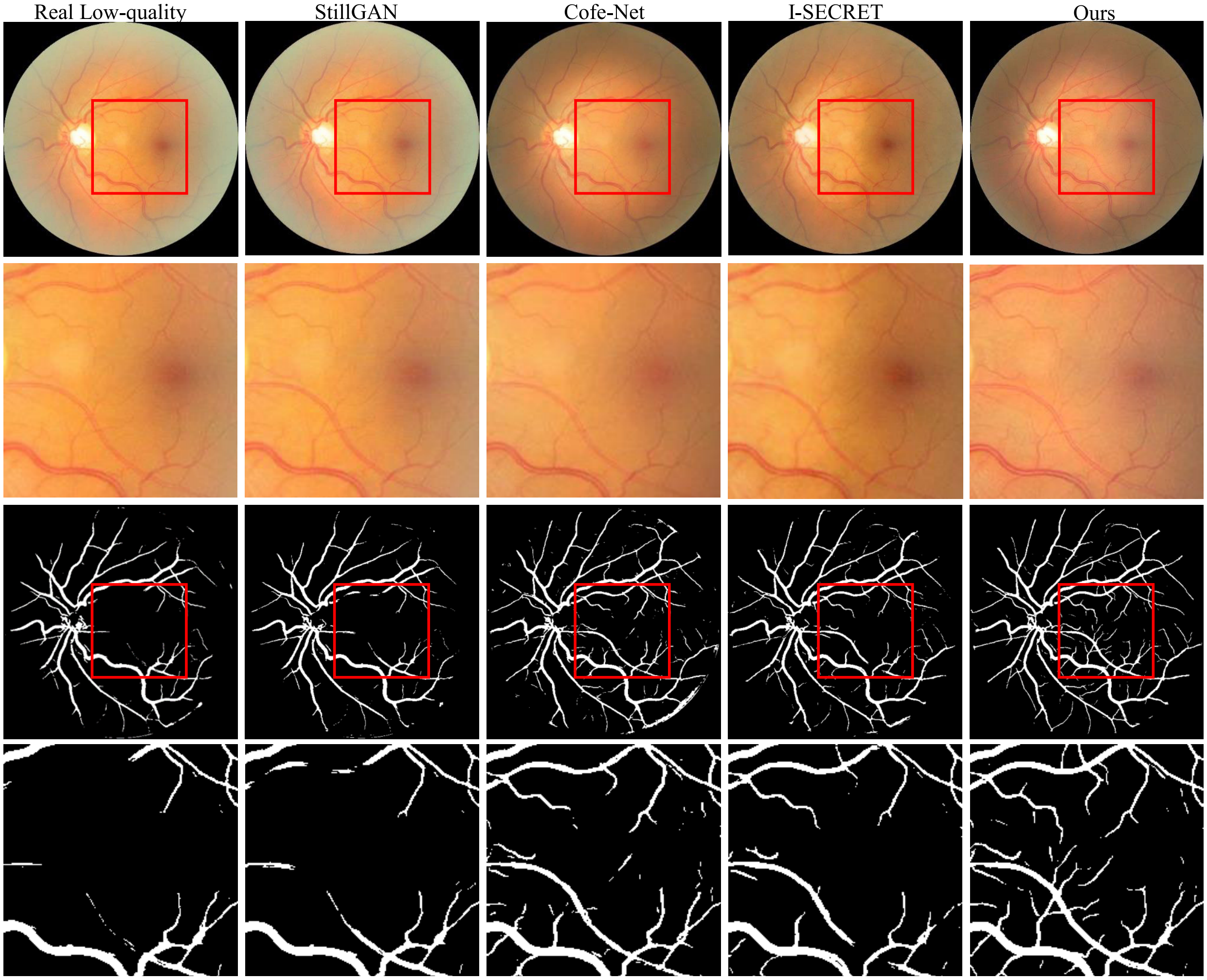


Fig. 5. **Visual comparison of enhancement results on real low-quality fundus images.** From top to bottom the images are the enhanced images and their zoom-in, and the vessel segmentation results and their zoom-in. The real low-quality images are given in the 1st column. There are no ground-truth high-quality images.

TABLE IV
Optic Disc/Cup Segmentation Performance Comparison of Different Methods on REFUGE [49] Dataset

| Methods | mIoU | Dice |
|---|---|---|
| Low-quality | 0.709 | 0.823 |
| high-quality | 0.789 | 0.882 |
| Setiawan [5] | 0.727 | 0.841 |
| DCP [50] | 0.720 | 0.831 |
| Cofe-Net [7] | 0.758 | 0.858 |
| StillGAN [15] | 0.725 | 0.834 |
| I-SECRET [24] | 0.750 | 0.852 |
| **T-MAGE-Net(ours)** | **0.762** | **0.862** |

we evaluate different methods on the degraded REFUGE [49] test set. We report the Dice and mIoU in Table IV. Specifically, we use Pra-Net [53] trained on the REFUGE training set(obtains mIoU/Dice of 0.789/0.882) for the detection on the enhanced images. The results consistently show that the enhanced images by our method could benefit the optic disc and cup detection for clinical observation. By performing paired t-tests based on the optic disc/cup detection results from our method and the best baseline (I-SECRET [24]), we achieve the p-values of 0.00157 (mIoU) and 0.00063 (Dice), indicating our improvements are statistically significant, as both p-values are lower than 0.05.

### D. Real Clinical Image Analysis

The enhancement model is expected to provide clean images with lesion preservation to assist diagnosis. This is validated by using the ODIR-5K dataset [54] collected from different hospitals and medical centers with different image qualities, which contains eight different labels including “normal”, “diabetes”, “glaucoma”, “cataract”, “age-related macular degeneration (AMD)”, “hypertension”, “myopia”, and “other diseases”. We adopt the Jordi et al. [55] model to classify ocular diseases. The results are shown in the Table V. Our method boosts the disease recognition performance evidently. Due to the gap between the synthetic and the real low-quality images, Cofe-Net [7] increases the risk of changing lesion areas, such as color distortion in Fig. 3. The enhanced images by I-SECRET [24] and StillGAN [15] are over-smoothed and fail to restore the retinal structure details that are important diagnose clue for eye diseases such as age-related macular degeneration, hypertension, and myopia [56]. In contrast, we preserve the retina structure by our RSP module, and

TABLE V

THE OCULAR DISEASE RECOGNITION RESULTS OF EACH DISEASE ON THE ODIR [54] DATASET

| | normal | | | diabetes | | | glaucoma | | | cataract | | | AMD | | | hypertension | | | myopia | | | other diseases | | | average | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Methods** | Kappa | ACC | AUC | Kappa | ACC | AUC | Kappa | ACC | AUC | Kappa | ACC | AUC | Kappa | **ACC** | AUC | Kappa | ACC | AUC | Kappa | ACC | AUC | Kappa | ACC | AUC | Kappa | ACC | AUC |
| Real fundus image | 0.2989 | 0.7875 | 0.7925 | 0.5100 | 0.8974 | 0.7926 | 0.4986 | 0.8875 | 0.8901 | 0.8107 | 0.9625 | 0.9776 | 0.7151 | 0.9425 | 0.9596 | 0.2142 | 0.8898 | 0.7895 | 0.4049 | 0.9100 | 0.9117 | 0.1694 | 0.7825 | 0.6595 | 0.4438 | 0.8825 | 0.8358 |
| Setiawan et al. [5] | 0.1501 | 0.7835 | 0.7835 | 0.2258 | 0.8800 | 0.7177 | 0.5862 | **0.9249** | 0.8761 | 0.5985 | 0.9325 | 0.9013 | 0.3181 | 0.7000 | 0.9199 | 0.0049 | 0.8725 | 0.5980 | 0.3043 | 0.9000 | **0.9716** | 0.0040 | **0.8475** | 0.4363 | 0.2887 | 0.8587 | 0.7541 |
| DCP [40] | 0.2899 | 0.7475 | 0.8100 | 0.3671 | 0.8675 | 0.7722 | 0.6026 | 0.9247 | 0.8975 | 0.7350 | 0.9500 | 0.9427 | 0.6270 | 0.9275 | 0.9259 | 0.3225 | **0.8950** | 0.7877 | 0.7615 | **0.9550** | 0.9437 | 0.1572 | 0.8425 | 0.6562 | 0.4662 | 0.8887 | 0.8488 |
| Cofe-Net [7] | 0.2495 | 0.7275 | 0.7868 | 0.3766 | 0.8800 | 0.7771 | 0.5405 | 0.9150 | 0.8780 | 0.7894 | 0.9575 | 0.9676 | 0.6832 | 0.9325 | 0.9590 | 0.2781 | 0.8800 | 0.8076 | 0.4049 | 0.9100 | 0.8626 | **0.2445** | **0.8475** | 0.6348 | 0.4328 | 0.8812 | 0.8336 |
| StillGAN [11] | 0.2500 | 0.7840 | 0.7515 | 0.3591 | 0.8550 | 0.7932 | 0.5580 | 0.9025 | 0.8870 | 0.6794 | 0.9425 | 0.9817 | **0.7977** | **0.9550** | 0.9628 | **0.3533** | 0.8948 | **0.8250** | 0.4741 | 0.9175 | 0.9514 | 0.2363 | 0.7900 | **0.6945** | 0.4500 | 0.8828 | 0.8528 |
| I-SECRET [21] | 0.2527 | 0.7875 | 0.7977 | 0.4110 | 0.8800 | 0.8074 | 0.5934 | 0.9225 | 0.8910 | 0.7961 | 0.9600 | 0.9747 | 0.7346 | 0.9350 | 0.9713 | 0.3529 | 0.8900 | 0.8197 | 0.5169 | 0.9220 | 0.8963 | 0.1566 | 0.8250 | 0.6658 | 0.4707 | 0.8903 | 0.8523 |
| MAGE-Net(ours) | 0.2835 | 0.7600 | 0.7900 | 0.4388 | 0.8825 | **0.8090** | **0.6160** | 0.9225 | 0.8935 | **0.8527** | **0.9700** | **0.9840** | 0.7942 | 0.9547 | **0.9790** | 0.2440 | 0.8800 | 0.8125 | 0.4741 | 0.9175 | 0.9266 | 0.1656 | 0.8300 | 0.6672 | 0.4772 | 0.8897 | 0.8526 |
| **T-MAGE-Net(ours)** | **0.3438** | **0.7900** | **0.8107** | **0.5147** | **0.9075** | 0.8076 | 0.5660 | 0.9075 | **0.8989** | 0.8390 | 0.9675 | 0.9766 | 0.7537 | 0.9475 | 0.9692 | 0.2203 | 0.8850 | 0.7758 | **0.5175** | 0.9225 | 0.9121 | 0.2267 | 0.7975 | 0.6727 | **0.4879** | **0.8906** | **0.8545** |

TABLE VI

THE ABLATION STUDY RESULTS ON THE DRIVE DATASET [48], WHERE"S1" DENOTES THE STAGE-1, "S2" DENOTES THE STAGE-2, "TS" DENOTES THE TEACHER-STUDENT FRAMEWORK, "LE" DENOTES THE SUPERVISED ENHANCEMENT LOSS, "LS" DENOTES THE SUPERVISED SEGMENTATION LOSS, "LCE" DENOTES THE CONSISTENT ENHANCEMENT LOSS, AND "LCS" DENOTES THE CONSISTENT SEGMENTATION LOSS

| Loss Combination | S1 | S2 | RSP | TS | **PSNR** | **SSIM** |
|---|---|---|---|---|---|---|
| Le | ✓ | | | | 22.93 | 0.9109 |
| Le | ✓ | ✓ | | | 23.35 | 0.9168 |
| Le+Ls | ✓ | ✓ | ✓ | | 23.72 | 0.9228 |
| Le+Lce | ✓ | ✓ | | ✓ | 24.47 | 0.9273 |
| Le+Ls+Lce+Lcs(w/o CAB) | ✓ | ✓ | ✓ | ✓ | 23.33 | 0.9212 |
| Le+Ls+Lce+Lcs | ✓ | ✓ | ✓ | ✓ | **24.72** | **0.9281** |

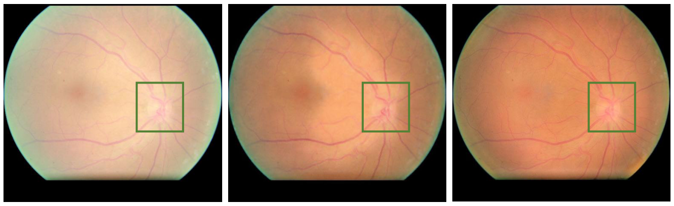
enhance the feature learning of our MAGE-Net based on both synthetic low-quality images (via the supervised enhancement process from the student model) and real low-quality images (via the unsupervised consistent enhancement process from both teacher and student models). In this way, the learned features can cater for both domains, which helps reduce domain gap to some extent. Without the teacher student framework, the results of Kappa/ACC/AUC become 0.4772/0.8897/0.8526, worse than 0.4879/0.8906/0.8545 achieved by our complete model under the teacher student framework, as shown in Table V.

### E. Ablation Study

To investigate the contribution of each component in our method, the ablation study is reported in the Table VI. Initially, stage-1 is a UNet-shaped network to correct fundus images, so some fine details are lost due to the sequential downsampling. To solve this problem, stage-2 is introduced to maintain important information. In order to emphasize retinal structures, we utilize the RSP module for clinical purposes. Our MAGE-Net combines the two stages and the RSP module. We observe that either the multi-stage strategy or the RSP module improves the PSNR and SSIM results. The employment of the teacher-student framework reduces the domain shift between the authentic image pairs and the real clinic images, which further improves both PSNR and SSIM results.

### F. Limitation

Although our proposed method outperforms other strong competitors, our method may not well reconstruct the images with too much noise. For example, if the image is extremely over/under-exposured, the proposed method will not work. Also, the vessel and disc/cup can be preserved only with moderate level of noise. A visual example about this limitation is shown in Fig. 6, which includes one extremely blurred real low-quality fundus image and the enhancement results from I-SECRET [24] and our method. Neither I-SECRET [24] nor our method produces clear disc/cup and vessel details. Fundus image enhancement under extreme noise condition is still a challenging problem, and will be investigated in our future work.



Extremely blurred real image I-SECRET Ours

Fig. 6. Visualization of failure cases. From the left to the right are an extremely blurred real low-quality fundus image, an enhanced image by I-SECRET [24], and an enhanced image by our method. Neither I-SECRET [24] nor our method produces clear disc/cup and vessel details.

## VI. CONCLUSION

In this paper, we propose a new transferred MAGE-Net method by integrating synthetic and real-world low-quality fundus images for multi-stage fundus image enhancement guided by multi-attentions. Furthermore, we design an RSP module to preserve the anatomical retinal structures and integrate it with our mean teacher based multi-stage enhancement framework seamlessly. Comprehensive experimental results demonstrate that our proposed method can simultaneously perform fundus image enhancement and reduce the domain gap between the synthetic and the authentic images. In addition, our method can boost the downstream tasks and assist clinical diagnosis for ophthalmologists and automated image analysis systems.

## REFERENCES

[1] M. D. Abràmoff, M. K. Garvin, and M. Sonka, "Retinal imaging and image analysis," *IEEE Rev. Biomed. Eng.*, vol. 3, pp. 169–208, 2010.

[2] S. Philip, "The impact of the health technology board for Scotland's grading model on referrals to ophthalmology services," *Brit. J. Ophthalmol.*, vol. 89, no. 7, pp. 891–896, Jul. 2005.

[3] M. Foracchia, E. Grisan, and A. Ruggeri, "Luminosity and contrast normalization in retinal images," *Med. Image Anal.*, vol. 9, no. 3, pp. 179–190, Jun. 2005.

[4] J. Cheng, Z. Li, Z. Gu, H. Fu, D. W. K. Wong, and J. Liu, "Structure-preserving guided retinal image filtering and its application for optic disk analysis," *IEEE Trans. Med. Imag.*, vol. 37, no. 11, pp. 2536–2546, Nov. 2018.

[5] A. W. Setiawan, T. R. Mengko, O. S. Santoso, and A. B. Suksmono, "Color retinal image enhancement using CLAHE," in *Proc. Int. Conf. ICT Smart Soc.*, Jun. 2013, pp. 1–3.

[6] M. Liao, Y.-Q. Zhao, X.-H. Wang, and P.-H. Dai, "Retinal vessel enhancement based on multi-scale top-hat transformation and histogram fitting stretching," *Opt. Laser Technol.*, vol. 58, pp. 56–62, Jun. 2014.

[7] Z. Shen, H. Fu, J. Shen, and L. Shao, "Modeling and enhancing low-quality retinal fundus images," *IEEE Trans. Med. Imag.*, vol. 40, no. 3, pp. 996–1006, Mar. 2021.

[8] A. D. Pérez, O. Perdomo, H. Rios, F. Rodríguez, and F. A. A. G. Lez, "A conditional generative adversarial network-based method for eye fundus image quality enhancement," in *Proc. Int. Workshop Ophthalmic Med. Image Anal.* Cham, Switzerland: Springer, 2020, pp. 185–194.

[9] U. Sevik, C. Köse, T. Berber, and H. Erdöl, "Identification of suitable fundus images using automated quality assessment methods," *J. Biomed. Opt.*, vol. 19, no. 4, Apr. 2014, Art. no. 046006.

[10] T. Li et al., "Applications of deep learning in fundus images: A review," *Med. Image Anal.*, vol. 69, Apr. 2021, Art. no. 101971.

[11] Z. Deng et al., "RFormer: Transformer-based generative adversarial network for real fundus image restoration on a new clinical benchmark," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 9, pp. 4645–4655, Sep. 2022.

[12] H. Li et al., "An annotation-free restoration network for cataractous fundus images," *IEEE Trans. Med. Imag.*, vol. 47, no. 7, pp. 1699–1710, Jul. 2022.

[13] H. Liu et al., "Degradation-invariant enhancement of fundus images via pyramid constraint network," in *Proc. MICCAI*, 2022, pp. 507–516.

[14] H. Li et al., "Structure-consistent restoration network for cataract fundus image enhancement," in *Proc. MICCAI*, 2022, pp. 487–496.

[15] Y. Ma et al., "Structure and illumination constrained GAN for medical image enhancement," *IEEE Trans. Med. Imag.*, vol. 40, no. 12, pp. 3955–3967, Dec. 2021.

[16] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–13.

[17] G. French, M. Mackiewicz, and M. Fisher, "Self-ensembling for visual domain adaptation," 2017, *arXiv:1706.05208*.

[18] A. Mitra, S. Roy, S. Roy, and S. K. Setua, "Enhancement and restoration of non-uniform illuminated fundus image of retina obtained through thin layer of cataract," *Comput. Methods Programs Biomed.*, vol. 156, pp. 169–178, Mar. 2018.

[19] W.-Y. Hsu and C.-Y. Chou, "Medical image enhancement using modified color histogram equalization," *J. Med. Biol. Eng.*, vol. 35, no. 5, pp. 580–584, Oct. 2015.

[20] G. D. Joshi and J. Sivaswamy, "Colour retinal image enhancement based on domain knowledge," in *Proc. 6th Indian Conf. Comput. Vis., Graph. Image Process.*, Dec. 2008, pp. 591–598.

[21] W. Ren et al., "Low-light image enhancement via a deep hybrid network," *IEEE Trans. Image Process.*, vol. 28, no. 9, pp. 4364–4375, Sep. 2019.

[22] B. L. Cai, X. M. Xu, K. Jia, C. M. Qing, and D. C. Tao, "DehazeNet: An end-to-end system for single image haze removal," *IEEE Trans. Image Process.*, vol. 25, no. 11, pp. 5187–5198, Aug. 2016.

[23] T. Huang, S. Li, X. Jia, H. Lu, and J. Liu, "Neighbor2Neighbor: Self-supervised denoising from single noisy images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14781–14790.

[24] P. Cheng, L. Lin, Y. Huang, J. Lyu, and X. Tang, "I-SECRET: Importance-guided fundus image enhancement via semi-supervised contrastive constraining," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2021, pp. 87–96.

[25] F. Shamshad et al., "Transformers in medical imaging: A survey," 2022, *arXiv:2201.09873*.

[26] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lucic, and C. Schmid, "ViViT: A video vision transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 6836–6846.

[27] H. Cao et al., "Swin-Unet: Unet-like pure transformer for medical image segmentation," 2021, *arXiv:2105.05537*.

[28] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. 16th Eur. Conf. Comput. Vis.* Glasgow, U.K.: Springer, Aug. 2020, pp. 213–229.

[29] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "SwinIR: Image restoration using Swin transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 1833–1844.

[30] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Restormer: Efficient transformer for high-resolution image restoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5728–5739.

[31] M. Long, H. Zhu, J. Wang, and M. I. Jordan, "Deep transfer learning with joint adaptation networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 2208–2217.

[32] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1180–1189.

[33] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7167–7176.

[34] Y. Liu et al., "From synthetic to real: Image dehazing collaborating with unlabeled real data," in *Proc. 29th ACM Int. Conf. Multimedia*, 2021, pp. 50–58.

[35] A. Raj, V. P. Namboodiri, and T. Tuytelaars, "Subspace alignment based domain adaptation for RCNN detector," 2015, *arXiv:1507.05578*.

[36] Y. Zhang, Z. Qiu, T. Yao, D. Liu, and T. Mei, "Fully convolutional adaptation networks for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6810–6818.

[37] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 286–301.

[38] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241.

[39] A. Odena, V. Dumoulin, and C. Olah, "Deconvolution and checkerboard artifacts," *Distill*, vol. 1, no. 10, p. e3, Oct. 2016.

[40] S. W. Zamir et al., "Multi-stage progressive image restoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 14821–14831.

[41] S. Zhang et al., "Attention guided network for retinal image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2019, pp. 797–805.

[42] P. Charbonnier, L. Blanc-Féraud, G. Aubert, and M. Barlaud, "Two deterministic half-quadratic regularization algorithms for computed imaging," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, vol. 2, Sep. 1994, pp. 168–172.

[43] G. Yang, Q. Ye, and J. Xia, "Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond," *Inf. Fusion*, vol. 77, pp. 29–52, Jan. 2022.

[44] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," 2016, *arXiv:1608.03983*.

[45] H. Fu et al., "Evaluation of retinal image quality assessment networks in different color-spaces," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2019, pp. 48–56.

[46] Y. Nan et al., "Data harmonisation for information fusion in digital healthcare: A state-of-the-art systematic review, meta-analysis and future research directions," *Inf. Fusion*, vol. 82, pp. 99–122, Jan. 2022.

[47] *Kaggle Diabetic Retinopathy Detection*. Accessed: 2015. [Online]. Available: https://www.kaggle.com/c/diabetic-retinopathy-detection/data/

[48] J. Staal, M. D. Abramoff, M. Niemeijer, M. A. Viergever, and B. van Ginneken, "Ridge-based vessel segmentation in color images of the retina," *IEEE Trans. Med. Imag.*, vol. 23, no. 4, pp. 501–509, Apr. 2004.

[49] J. I. Orlando et al., "REFUGE challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs," *Med. Image Anal.*, vol. 59, Jan. 2020, Art. no. 101570.

[50] K. He, J. Sun, and X. Tang, "Single image haze removal using dark channel prior," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 12, pp. 2341–2353, Dec. 2011.

[51] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[52] Z. Gu et al., "CE-Net: Context encoder network for 2D medical image segmentation," *IEEE Trans. Med. Imag.*, vol. 38, no. 10, pp. 2281–2292, Oct. 2019.

[53] D.-P. Fan et al., "PraNet: Parallel reverse attention network for polyp segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2020, pp. 263–273.

[54] *Peking University International Competition on Ocular Disease Intelligent Recognition (ODIR-2019)*. Accessed: 2019. [Online]. Available: https://odir2019.grand-challenge.org/dataset/

[55] C. Jordi, N. J. Manuel, and V. Carles, "Ocular disease intelligent recognition through deep learning architectures," IEEE, Universitat Oberta de Catalunya, Barcelona, Spain, Tech. Rep. 1-114, 2019.

[56] T. A. Soomro et al., "Deep learning models for retinal blood vessels segmentation: A review," *IEEE Access*, vol. 7, pp. 71696–71717, 2019.