scientific reports

Check for updates

OPEN

A two-stage multi-scale attention-based network for weakly supervised cataract fundus image enhancement

Xiaoyong Fang[1], Yue Wang[2], Xiangyu Li[2], Wanshu Fan[2]✉ & Dongsheng Zhou[2,3]✉

Cataract is a major cause of vision loss and hinders further diagnosis. However, enhancing cataract fundus images remains challenging due to limited paired cataract retinal images and the difficulty of recovering fine details in the retinal images. To mitigate these challenges, we in this paper propose a two-stage multi-scale attention-based network (TSMSA-Net) for weakly supervised cataract fundus image enhancement. In Stage 1, we introduce a real-like cataract fundus image synthesis module, which utilizes domain transformation via CycleGAN to generate realistic paired cataract images from unpaired clear and cataract fundus images, thus alleviating the scarcity of paired training data. In Stage 2, we employ a multi-scale attention-based enhancement module, which incorporates hierarchical attention mechanisms to extract rich, fine-grained features from the degraded images under weak supervision, effectively restoring image details and reducing artifacts. Experiments conducted on the Kaggle and ODIR-5K datasets show that TSMSA-Net outperforms existing state-of-the-art methods for cataract fundus image enhancement, even without paired images, and demonstrates strong generalization ability. Moreover, the enhanced images contribute to improved performance in downstream tasks such as vessel segmentation and disease classification.

Keywords Cataract fundus enhancement, Multi-scale attention, Weakly supervised learning

With the development of deep learning techniques, researchers have proposed numerous retinal disease detection and segmentation algorithms to aid in clinical diagnosis[1]. However, these algorithms require high-quality retinal image inputs, while cataract retinal images are typically characterized by capture device and patient variability, making it difficult to ensure image quality. The quality of cataract retinal images is often shown as blurriness and poor image readability, Rendering the diagnosis of diseases by ophthalmologists challenging. Meanwhile, these poor-quality retinal images also may lead to suboptimal outcomes in automatic image processing, such as disease detection and segmentation, consequently impacting further disease diagnosis. Therefore, the restoration of cataract fundus images has clinical value[2]. Figure 1 shows the fundus images of cataract image and the enhancement image restored by our TSMSA-Net. It can be observed that compared to Fig. 1c, d, Fig. 1a, b are relatively blurry, with lower visibility of the fundus structures. It is difficult to accurately extract the fundus information of these blurred images, indicating the effectiveness of our TSMSA-Net.

In addressing the blurriness of the cataract retinal images, researchers have extensively explored retinal image enhancement methods[3–7], utilizing classical methods to improve image quality. However, these manually designed algorithms fail to adequately preserve image details and suffer from the issue of amplifying image noise, leading to erroneous guidance in image restoration. Subsequently, with the advancement of deep learning, researchers began to utilize deep learning to enhance retinal images, achieving promising results[8–11].

Although the artificial synthesis methods[12–14] can effectively obtain a large number of paired cataract retinal images, the synthetic function cannot fully cover the degradation conditions of cataract images, resulting in poor generalization ability of well-trained networks in real cataract retinal image application scenarios. In addition, unsupervised methods[9,15] are prone to losing detailed information of retinal images during the enhancement process due to the lack of supervised constraints.

[1]Department, School of Safety and Management Engineering, Hunan Institute of Technology, Hengyang 421002, China. [2]National and Local Joint Engineering Laboratory of Computer Aided Design, School of Software Engineering, Dalian 116622, China. [3]School of Computer Science and Technology, Dalian University of Technology, Dalian 116024, China. ✉email: fanwanshu@dlu.edu.cn; zhouds@dlu.edu.cn

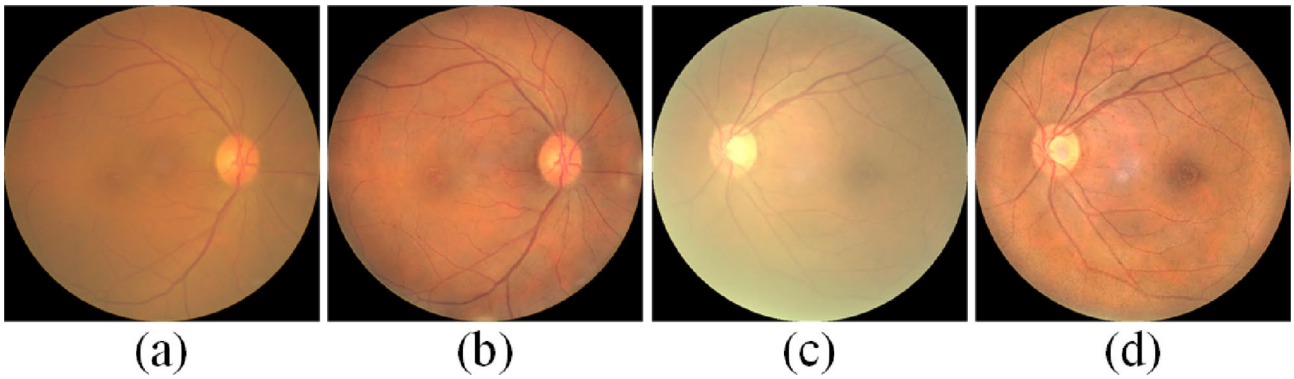

Fig. 1. Fundus images. (**a**,**b**) Cataract images. (**c**,**d**) The image enhanced by our TSMSA-Net.

To address the above challenges, we propose a two-stage multi-scale attention-based weakly supervised cataract retinal image enhancement network (TSMSA-Net). Our TSMSA-Net mainly leverages a real-like cataract image synthesis stage to simulate more realistic paired degradation images. Specifically, we firstly apply a synthetic function to narrow the domain gap between synthetic and real cataract retinal images. Subsequently, we use the synthetic cataract images as the source domain and real cataract images as the target domain for domain transformation, resulting in real-like paired synthetic cataract retinal images. To further enhance the detail rendition, we introduce a multi-scale attention-based cataract retinal image enhancement stage. Unlike the widely used U-Net[16], we utilize multi-scale attention modules to extract more abundant image detail features, avoiding the loss of details during the down-sampling process. TSMSA-Net combines the conversion capabilities of CycleGAN, the structural information retention strategy of ArcNet, and the unsupervised learning idea of EnlightenGAN, introduces a multi-scale attention mechanism and a high-frequency extraction module, and achieves better image enhancement effects.

We summarise the main contributions of this paper as follows:

- We propose a real-like cataract image synthesis stage to obtain more paired realistic synthesized cataract images, addressing the problem of difficult acquisition of paired images and the inability of the synthesis function to effectively cover the degradation of cataract images.
- We propose a multi-scale attention-based cataract image enhancement stage to better fuse multi-scale features, enabling the enhancement network to better recover image details and reduce the generation of artifacts.
- Qualitative and quantitative experiments on the Kaggle and ODIR-5K datasets demonstrate that our TSMSA-Net outperforms existing state-of-the-art cataract image enhancement methods. And the enhancement can improve the automatic image processing, such as disease classification and segmentation.

Related work

Classical retinal image enhancement methods

Classical methods for retinal image enhancement typically involve manually designing algorithms using prior information of the images, often focusing on enhancing contrast and brightness. For instance, many methods utilize contrast limited adaptive histogram equalization (CLAHE) to improve image contrast and achieve retinal image enhancement[17]. Additionally, some researchers employ filtering techniques for image enhancement[18]. In addition, Dash et al.[4] propose a joint model of fast guided filter and matching filter to enhance vascular extraction performance. Mohammed et al.[3] present a hybrid algorithm that utilizes wiener filtering and CLAHE to enhance color retinal fundus images, reducing noise generation and achieving better enhancement effects.

Although the aforementioned methods can achieve excellent enhancement results, they also exist some limitations. Firstly, they are unable to precisely control the enhancement level, resulting in extracted features that struggle to preserve image details and are prone to amplifying image noise. Secondly, these manually designed prior information is often simplistic and cannot fully adapt to the various retinal degradations present in the real world, thus limiting their applicability.

Deep learning-based retinal image enhancement methods

In recent years, deep learning has shown outstanding performance and has been widely applied to low-level visual tasks, such as segmentation[19], dehazing[20], and image enhancement[21]. In the field of image enhancement, there are also many methods that have achieved excellent enhancement effects[22–24]. However, most methods use the supervised learning method, leveraging a large amount of paired training data to learn the mapping from low-quality images to high-quality images[25]. However, obtaining the paired data in medical scenarios is extremely difficult and time-consuming. Therefore, researchers have proposed artificially degrading high-quality images to synthesize low-quality images for supervised training[10,12,26–28].nBy computing the OT cost in feature space, this method is able to better preserve local structures and minimize unnecessary artifacts[29]. An end-to-end optimized teacher-student framework is proposed for simultaneous image enhancement and domain

adaptation[30]. This paper proposes a model based on a teacher-student network, combining NoiseContextNet Block and iterative pruning technology to improve denoising effect and computational efficiency[31].

Since the degradation conditions that can be covered by artificially degraded methods are limited and cannot fully restore the degradation of real cataract images, and there exists a domain gap between artificially synthesized fundus degradation images and real degraded fundus images, the models trained by these methods often lack generalization ability when processing real degraded fundus images. Subsequently, some researchers propose semi-supervised methods for fundus image enhancement to reduce the dependence on the requirement of paired data. Wu et al.[8] propose a semi-supervised generative adversarial network (SSGAN-ASP) to train the network using both supervised data and unsupervised data. In addition, there are also methods proposed to use unpaired images for unsupervised learning in fundus image enhancement to reduce the need for paired data. Yang et al.[15] introduce an unpaired fundus image enhancement method based on high-frequency extraction and feature description to preserve the structural information of the image and reduce the generation of vascular-like artifacts during the enhancement process. Li et al.[9] propose an unsupervised cataract fundus image restoration network (ArcNet) that does not require annotations. However, due to the lack of supervision constraints, unsupervised learning methods mainly simulate the results of high-quality images from low-quality images through image style transformation, which easily leads to the loss of detailed information in low-quality images.

Proposed approach

Due to the inability of synthesized images generated by the composite function to fully simulate the degradation of real cataract fundus images, and the risk of losing structural and detailed information in unsupervised learning methods, we propose a Two-Stage Multi-Scale Attention-based Network (TSMSA-Net) for weakly supervised cataract fundus image enhancement. To better obtain synthesized cataract fundus images that are closer to realistic scenarios, we propose a real-like cataract fundus image synthesis stage. To better capture the detailed information of cataract fundus images, we design a multi-scale attention-based enhancement stage, which learns informative features under weak supervision to preserve fine details and improve image quality.

Overall pipeline

Figure 2 illustrates the overall architecture of our two-stage multi-scale attention-based weakly supervised cataract retinal image enhancement network (TSMSA-Net), which consists of a real-like cataract fundus image synthesis stage as stage 1 and a multi-scale attention-based cataract fundus image enhancement stage as stage 2. In stage 1, we first utilize a composite function $C(\cdot)$ to generate simulated cataract fundus images, aiming to reduce the domain gap between the synthesized and real cataract fundus images. Subsequently, we use the CycleGAN[32] network for domain translation between the synthesized cataract fundus images and the corresponding real cataract fundus images, to make the synthesis images are closer to the real ones. We utilize these synthesized

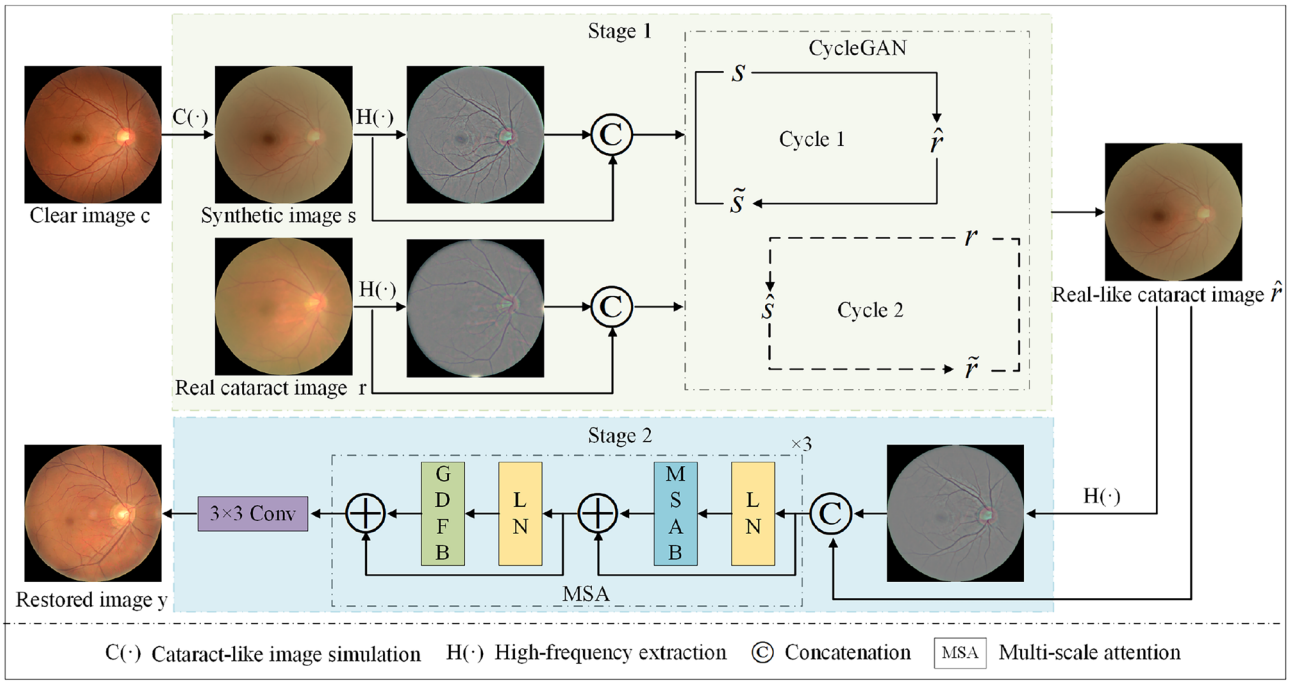

Fig. 2. The overall architecture of the proposed network. Our network consists of stage 1 and stage 2. In stage 1, given a clear image c and a real cataract image r, we first extract structural information from each image for guidance. Subsequently, we employ CycleGAN[32] to perform domain transformation on the synthetic image and the real cataract image, And in stage 2, we use a series of multi-scale attention (MSA) to extract more details from the synthetic real-like cataract images.

cataract fundus images paired with their corresponding clear fundus images as the input for the stage 2. In stage 2, we employ multi-scale attention to further extract the detailed information of the degraded images more precisely.

Real-like cataract fundus image synthesis stage

To generate synthetic cataract fundus images that better reflect real-world conditions, we propose a real-like cataract fundus image synthesis stage, as shown stage 1 in Fig. 2. S represents the source domain, $\tilde{S}$ represents the image mapped from the source domain S to the target domain T through the generative network. r represents the target domain,$\tilde{r}$ represents the image mapped from the source domain S to the target domain T through the generative network. $\hat{r}$ represents the image mapped from the target domain T to the source domain S through the generative network.resulting in the generation of real-like cataract image. Inspired by ArcNet[9], we adopt the degradation model proposed in[33] to simulate the cataract-induced degradation process and further improve it to better suit our task. The simulate formulation is represented as:

$$C(s_C) = \alpha \cdot s_C * g_B\left(r_B, \sigma_B\right) + \beta \cdot J * g_L\left(r_L, \sigma_L\right) \cdot \left(L_C - s_C\right), \tag{1}$$

where $C(s_C)$ represents the simulated cataract fundus image, s_C denotes the clear image, c stands for the image's r, g, b channel, α and β represent the weights of the clear fundus image and the noise from the cataract, $*$ denotes the convolution operation, g_B and g_L represent gaussian filters for smoothing the clear image and the cataract panel, respectively, $g\left(r, \sigma\right)$ denotes a gaussian filter with radius r and spatial constant σ, J represents the cataract panel, and L_C represents the highest intensity of s_C.

In the proposed two-stage multi-scale attention weakly supervised cataract retinal image enhancement network (TSMSA-Net), the reason for dividing F_{a} into three parts along the channel dimension is to better capture feature information at different scales. After decomposition, features can be learned on different channels, thereby understanding the image content more comprehensively. The network is also able to focus on feature representations at different levels, which helps to improve the effect of image enhancement. Specifically, F_a is divided into three parts F_{a_1}, F_{a_2}, and F_{a_3}, and each part passes through an attention block with different convolution kernel sizes, so that diverse features can be extracted. These different scale features are then merged and multiplied with F_{b} to obtain multi-scale attention features. This helps the network work effectively on different image regions and objects of different sizes.

The segmentation and multi-scale processing enhance the expressiveness of feature representations. If the original feature F is directly input into the attention block, the details and hierarchical structures that can be captured by segmentation and multi-scale processing may be missed. In addition, the segmented features are processed through different attention blocks, which can learn richer and more diverse feature representations.

However, the degradation that can be simulated by mathematical formulas is limited. Therefore, we further synthesize cataract fundus degraded images that are closer to real-world scenarios through an improved CycleGAN[32]. Through two GAN networks, we learn mappings from the sythesized cataract image domain s to the real cataract image domain r, and from real cataract image domain r to the synthesized cataract image domain s. We ensure the similarity between generated images and input image content through GAN loss functions and cycle consistency loss functions. The generative networks are adapted from Johnson et al.[34], which contains two convolutional layers with stride 2, several residual blocks and two convolutional layers with stride $\frac{1}{2}$. For the discriminator network, we use 70 $\times$ 70 PatchGANs[35,36]. In order to retain as much structural information of the fundus image as possible, inspired by ArcNet[9], based on the Retinex theory[37], we utilize a high-frequency extraction module $H\left(\cdot\right)$ to extract structural information from the fundus image as the guidance. The high-frequency extraction module $H\left(\cdot\right)$ can be represented as:

$$H\left(I\right) = I - I * g_P\left(r_P, \sigma_P\right), \tag{2}$$

where I represents the fundus image, and $g_P\left(r_P, \sigma_P\right)$ denotes a gaussian filter with radius r_P and spatial constant σ_P.

Multi-scale attention-based cataract fundus image enhancement stage

To fully leverage the features of the synthesized real-like cataract fundus images and enhance the enhancement effect, we propose a multi-scale attention-based cataract fundus image enhancement stage, as shown in the stage 2 of Fig. 2. This stage consists of a high-frequency extraction module and three multi-scale attention (MSA) modules. The high-frequency extraction module aims to preserve the structural information of the fundus image for better restoration of cataract fundus image details. Inspired by[38], the MSA is designed to exploit the features of the synthesized real-like cataract fundus images and learn richer characteristics of the fundus images. The learning process of the multi-scale attention modules can be represented as follows:

$$M_i = F_i + f_{MSAB}\left(LN\left(F_i\right)\right), \tag{3}$$

$$F_{i+1} = M_i + f_{GDFB}\left(LN\left(M_i\right)\right), \tag{4}$$

where F_i and F_{i+1} represent the input and output features of the multi-scale attention module, M_i represents the extracted multi-scale features, $f_{MSAB}\left(\cdot\right)$ denotes the MSAB module, $f_{GDFB}\left(\cdot\right)$ denotes the Global Dual-Branch Fusion Block (GDFB), and $LN\left(\cdot\right)$ represents the layer normalization operation.

As shown in Fig. 3, the MSAB module consists of three attention blocks (AB). Firstly, the features are divided into two parts along the channel dimension, resulting in features F_a and F_b. Next, F_a is further divided into

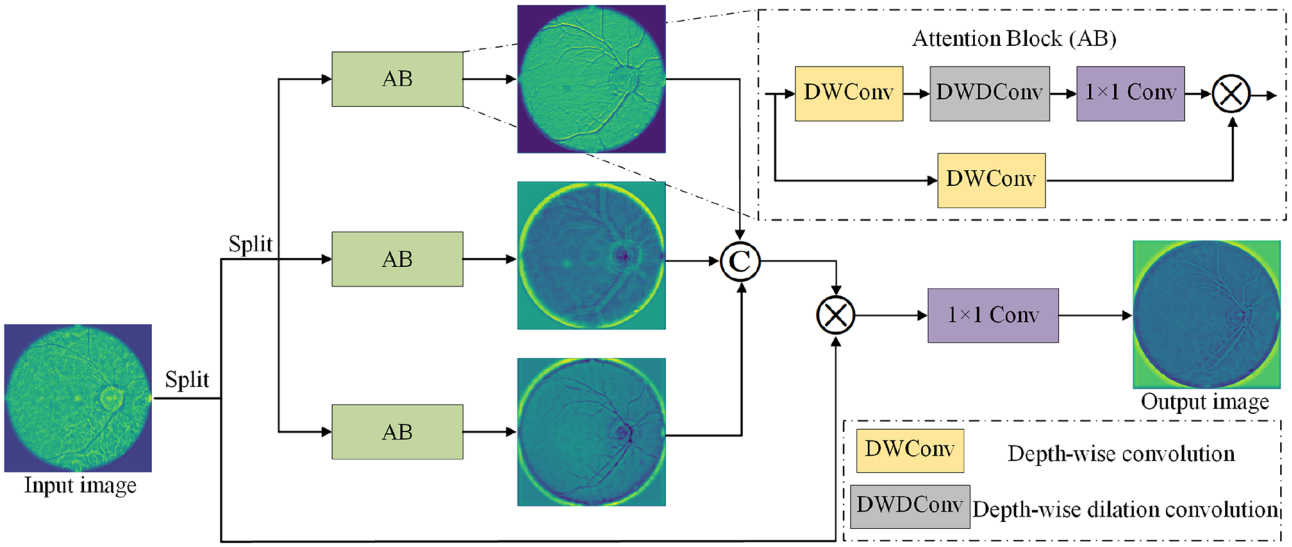

Fig. 3. The architecture of multi-scale attention block (MSAB). Our MSAB mainly consists of 3 attention blocks (AB) with different scales of convolution kernel to explore the multi-scale features for the better enhancement.

three parts along the channel dimension, yielding F_{a_1}, F_{a_2}, and F_{a_3}, which are then fed into the attention blocks (AB). Each AB module employs different convolutional kernels to extract multi-scale features and capture richer image detail information. Subsequently, the features of different scales are concatenated and multiplied element-wise with F_b to obtain the multi-scale attention features. AB is shown in Fig. 3, which consists of two deep-wise convolutions, a deep-wise dilation convolution, and a 1×1 convolution to achieve a larger receptive field for feature extraction. The GDFB firstly adopt a channel-wise splitting to split the input into two halves. We remain one branch while applying a deep-wise convolution operationt to the other branch. Afterwards, we merge the dual cranch together by multiplied to obtain the global spatial information.

Loss function

Real-like cataract fundus image synthesis stage We train the real-like cataract fundus image synthesis stage using the loss function from CycleGAN[32], which can be formulated as follows:

$$L_{CycleGAN} = \lambda L_{cyc}\left(G_S^R, G_R^S\right) + L_{GAN}\left(G_S^R, D_R, S, R\right) + L_{GAN}\left(G_R^S, D_S, S, R\right), \tag{5}$$

where S and T respectively represent the source domain and target domain, G_{XY} represents the mapping function from domain X to Y, D_S and D_T represent the discriminators for the source and target domains, L_{GAN} and L_{cyc} represent the adversarial loss and cycle consistency loss, λ represents the weight parameter. In addition, we also use the structural loss L_R in[10] to preserve the structural information of the fundus. The overall loss function can be expressed as:

$$L_{total} = L_{CycleGAN}(\cdot) + L_R\left(S_G, \hat{R_G}\right) + L_R\left(R_G, \hat{S_G}\right), \tag{6}$$

where X_G represents the structural graph of X.

Multi-scale attention-based cataract fundus image enhancement stage: To preserve the content information of fundus images, retain their structural details, color brightness, and minimize the occurrence of artifacts during the restoration process, we utilize several loss functions. These include the mean squared error loss L_{MSE} to constrain structural variations, color loss L_{Color} to maintain color brightness consistency, total variation loss L_{TV} and L_1 loss to preserve image edges and promote smoothness, and structural similarity index loss L_{ssim} to recover both brightness and structural details. L_{ssim} can improve the structural quality of the generated image, mainly helping to maintain the structural information of the image, not just the pixel-level error, but also maintaining consistency at a higher level of visual structure. L_{color} is used to maintain the consistency of the image color, making it visually closer to the true value image. The overall loss function can be expressed as:

$$L_{multi} = \lambda_{TV}\Delta(c-y) + \lambda_{MSE}||c-y||_2 + \lambda_{L1}||c-y||_2 + \lambda_{ssim}\,\tilde{ssim}((c,y)) + \lambda_{Colour}||\max_{rgb} c^{rgb} - \max_{rgb} y^{rgb}||, \tag{7}$$

where λ represents the hyperparameter, rgb denotes the r, g, b channels, and $\tilde{ssim} = 1 - ssim$.

Method	NIQE ↓	IS ↑
CycleGAN[32]	9.50	1.41
CofeNet[27]	9.12	1.39
EnlightenGAN[24]	8.95	1.49
ArcNet[9]	7.09	1.42
PCENet[12]	6.32	1.53
GFENet[10]	6.37	1.41
TSMSA-Net(Ours)	**6.22**	**1.57**

Table 1. Quantitative results on Kaggle dataset. The best results are highlighted in bold.

Method	NIQE ↓	IS ↑
CycleGAN[32]	9.85	1.37
CofeNet[27]	9.75	1.39
EnlightenGAN[24]	8.47	1.50
ArcNet[9]	6.43	1.31
PCENet[12]	6.20	1.53
GFENet[10]	6.21	1.47
TSMSA-Net(Ours)	**6.12**	**1.56**

Table 2. Quantitative results on ODIR-5K dataset. The best results are highlighted in bold.

Experiments

Dataset and evaluation metrics

Dataset. We train and test our TSMSA-Net on publicly available datasets. Specifically, we use the normal and cataract subsets of the Kaggle dataset to create an unpaired dataset for unsupervised training in stage 1. And in stage 2, we use the normal subset of the Kaggle dataset and the degraded normal subset generated from stage 1 to form a paired dataset for supervised training. All images are resized to 512 $\times$ 512 before being sent to the model. For testing, we use a subset of cataract-labeled images from the ODIR-5K dataset and the cataract subset of the Kaggle dataset.

Evaluation metrics. We assess the effectiveness of our TSMSA-Net by using the natural image quality evaluator (NIQE)[39] and initial score (IS)[40] as metrics to evaluate the image enhancement quality, where lower values of NIQE and higher values of IS indicate better performance.

Implementation details

We implement our TSMSA-Net on the PyTorch framework, optimize using the Adam optimizer[41] and train on a single V100 GPU. In stage 1, we start with an initial learning rate of 0.0002, and the model is trained for 150 epochs with a linear decay of the learning rate. The batch size is set to 8. In stage 2, during the training phase, we start with an initial learning rate of 0.0001, and the model is trained for 100 epochs. The input image size is 512×512, and we random crop size to 256×256 and fed into the network. The batch size is set to 4. For the loss function, we use the weights as follows: λ_{TV}=1, λ_{MSE}=1, λ_{Color}=0.1, λ_{ssim}=0.1, and λ_{L_1}=0.5. The MSAB employ convolutional kernels with scales of 7-9-1, 5-7-1, and 3-5-1 for different branches. During the testing phase, the input image size is set to 512×512, and the batch size is set to 1.

Comparision with state-of-the-art methods

Quantitative results

We compare our TSMSA-Net with six state-of-the-art models, including CycleGAN[32], CofeNet[27], EnlightenGAN[24], ArcNet[9], PCENet[12], and GFENet[10]. Tables 1 and 2 summarize the comparison results, which are based on the pre-trained models provided by the networks and tested on our dataset. From Tables 1 and 2, we can clearly observe that our TSMSA-Net achieves the best results in NIQE and IS on Kaggle and ODIR-5K. Concretely, our method surpasses GFENet by 0.15 and 0.09 in NIQE on Kaggle and ODIR-5K respectively, although GFENet is trained on a larger dataset. Although ArcNet[9] has seen the test set images during the training process, our method still outperforms ArcNet by 0.87 and 0.15 in NIQE and IS on the Kaggle dataset. As shown in Table 6, We also test the value of PSNR and our method achieved the best results.

Visual comparison

Meanwhile, we also provide the visual comparison between TSMSA-Net and other models on the cataract sub-dataset of Kaggle and the ODIR-5K cataract sub-dataset in Figs. 4 and 5, respectively. It is important to note that unlike other models, our model is trained only on the 300 images in the Kaggle sub-dataset, which may contribute to the stylistic differences in the restored images compared to other models. From Fig. 4, it can be observed that our model is capable of restoring clean and clear vessels in the optic disk, demonstrating the ability of the multi-scale attention module to extract image features. Additionally, as our model has not seen the ODIR-

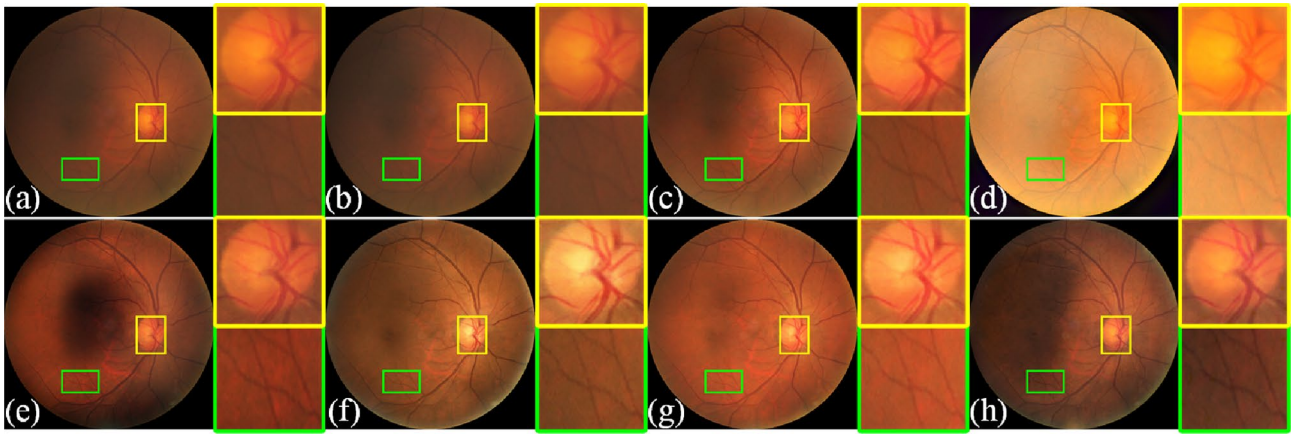

Fig. 4. Visual comparsion with state-of-the-art methods on Kaggle, and areas of contrast are marked with green and yellow boxes on the original image. (**a**) Cataract image. (**b**) CycleGAN[32]. (**c**) CofeNet[27]. (**d**) EnlightenGAN[24]. (**e**) ArcNet[9]. (**f**) PCENet[12]. (**g**) GFENet[10]. (**h**) TSMSA-Net(ours).

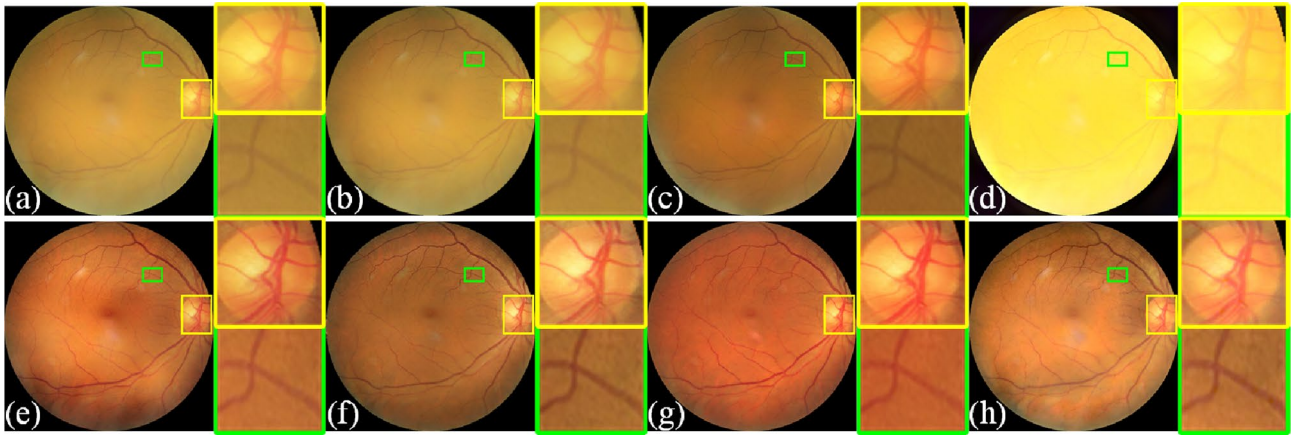

Fig. 5. Visual comparsion with state-of-the-art methods on ODIR-5K, and areas of contrast are marked with green and yellow boxes on the original image. (**a**) Cataract image. (**b**) CycleGAN[32]. (**c**) CofeNet[27]. (**d**) EnlightenGAN[24]. (**e**) ArcNet[9]. (**f**) PCENet[12]. (**g**) GFENet[10]. (**h**) TSMSA-Net(Ours).

w/o Stage1	w/o MSAB	NIQE ↓	IS ↑
✓		6.28	1.53
	✓	6.64	1.54
✓	✓	6.41	1.53
		6.22	**1.57**

Table 3. Ablation study on Kaggle dataset. The best results are highlighted in bold.

5K dataset during the entire training process, Fig. 5 also reflects the generalization capability of our model. It is evident that our model can effectively restore detailed features such as blood vessels even when trained on a relatively small dataset.

Ablation study

In this section, we conduct ablation experiments to investigate the effect of the proposed different components. We test the NIQE[39] and IS metrics[40] on the Kaggle dataset. The experimental results are summarized in Table 3. "w/o stage1" indicates that paired images are not obtained through training in stage 1, and synthetic images generated using the composition function are directly paired with corresponding clear images. "w/o MSAB" means that the multi-scale module is not used in stage 2. We only use a single AB module with a convolution kernel of scale 3-5-1 instead of the three AB modules of different scales.

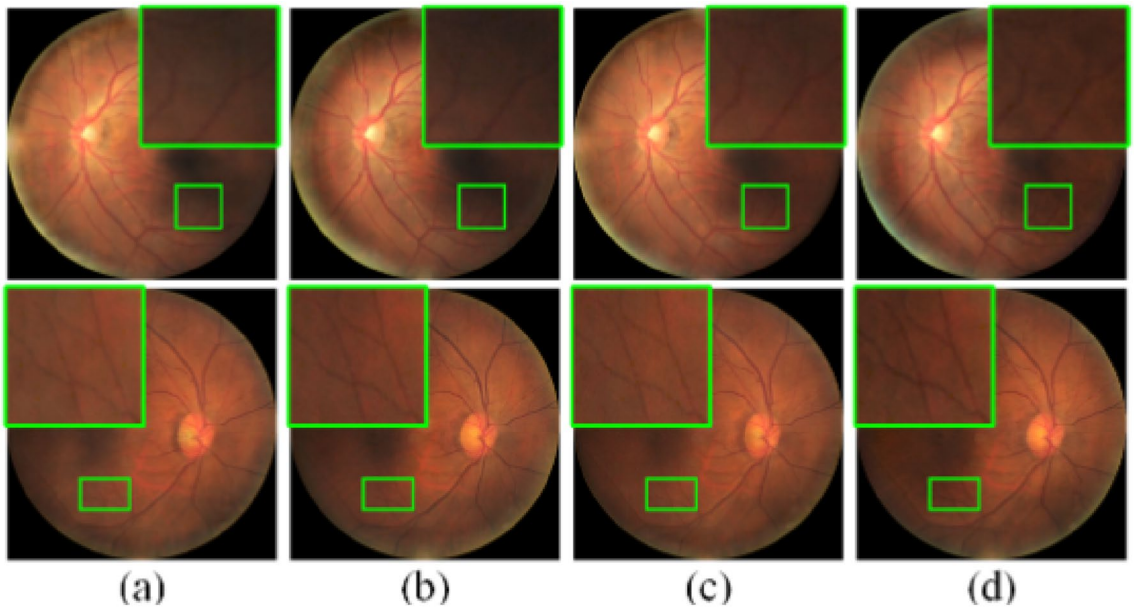

Fig. 6. Visual comparison on ablation study. (**a**) Represents using synthetic cataract images by simulate formulation directly without the first stage and without multi-scale feature extraction. (**b**) Represents using synthetic cataract images by simulate formulation directly without the first stage. (**c**) Represents without multi-scale feature extraction. (**d**) Represents our method, utilizing both the first stage and multi-scale feature extraction. The magnified areas are indicated by green boxes in the retinal image.

Method	Param (M)	Inference time (s)
CycleGAN[32]	57.1	11.64
CofeNet[27]	41.2	79.72
EnlightenGAN[24]	8.6	34.84
ArcNet[9]	54.4	59.29
PCENet[12]	26.6	12.63
GFENet[10]	89.3	25.4
TSMSA-Net(ours)	**0.17**	**9.43**

Table 4. The comparison of model complexity. The best results are highlighted in bold.

From the first and fourth rows of Table 3, it can be observed that without the learning from stage 1 network, the NIQE and IS indicators decrease by 0.06 and 0.04, respectively. This indicates that the synthesized cataract images from stage 1 are closer to real cataract degradation compared to the synthesized function. As a result, better enhancement effects can be achieved for real cataract images. Moreover, from the second and fourth rows of Table 3, it can be seen that employing multi-scale attention significantly boosts performance by 0.42 and 0.03 in NIQE and IS metrics respectively, demonstrating that the extraction of multi-scale information can effectively enhance the ability of image enhancement. Notablely, from the second and third rows of Table 3, it can be observed that the effect of using only the image enhanced by the stage 1 network is slightly inferior to directly using the synthesized image. This may be because there is some degree of feature loss during the domain adaptation process. As shown in Table 3, using both stage 1 and the multi-scale attention module can achieve the best results on the Kaggle dataset, which can further demonstrate the proposed components's effectiveness. We also add the ablation study to evaluate the effect of concatenating the real-like cataract image with the high-frequency image. As shown in Table 5, when concatenating the real-like cataract image with the high-frequency image, the NIQE and IS achieve the better results.

Furthermore, we also provide visual comparison, as shown in Fig. 6. It can be observed that compared with Fig. 6d, without the stage 1 training, Fig. 6a,b have more obvious black shadows in the enhanced images. Moreover, comparing the magnified vessel images in Fig. 6c,d, it shows that the introduction of multi-scale feature extraction can extract richer vessel details, resulting in clearer recovery of vessels.

Model complexity comparisons

An analysis of the model's complexity is essential for a comprehensive evaluation. As shown in Table 4, a comparative study of parameters and inference time among various models is presented. The inference time is obtained by inferring 100 images on a V100 GPU. In addition, since our model only synthesizes pseudo cataract

Real-like cataract image	High-frequency image	NIQE ↓	IS ↑
✓		10.12	1.46
✓	✓	**6.22**	**1.57**

Table 5. Ablation study on Kaggle dataset. The best results are highlighted in bold.

Method	SSIM
TLLR[42]	0.50
LLIE[43]	0.55
HIEA[44]	0.65
RPCA[45]	0.78
Ours	**0.82**

Table 6. Quantitative results on widely-used benchmarks. The best results are marked in bold. Higher SSIM values reflect improved performance.

images in the stage 1, the cataract image enhancement process only occurs in the stage 2, which is equivalent to the image pre-processing process in stage 1. Therefore, when comparing model complexity, only the complexity of the stage 2 of the model is calculated. It can be seen that our method has achieved optimal results in terms of both parameter quantity and inference time. Specifically, the parameter quantity of our model is only 1.9% of that of GFE-Net, and the inference time is only 37% of that of GFE-Net. This fully demonstrates the outstanding performance of our method in terms of model complexity.

Applications

Our enhancement method can be used as the pre-processing for cataract retinal vessel segmentation tasks. We use the U-Net[16] trained for vessel segmentation on the DRIVE[46] and STARE[47] datasets. The results are shown in Fig. 7. It can be observed that before enhancement, the structure of the retina fundus is relatively blurred, making it difficult to effectively segment vessels, resulting in less than ideal segmentation outcomes. However, after our enhancement network, the visibility of vessels in the retinal image is significantly improved, and the segmentation results are noticeably enhanced, effectively improving the ability of retinal vessel segmentation. More vessel structures can be segmented, which is beneficial for further medical diagnosis.

Furthermore, our approach aslo can contribute to automatic disease diagnosis. We use the retinal fundus multi-disease dataset (RFMiD)[48] to train the ConvNeXt network[49] for automatic detection. RFMiD is created for training automatic classification methods for both common and rare diseases, camprising 3200 fundus images captured by three different fundus cameras. Among them, 317 images are labeled with media haze, which may hinder disease diagnosis. We perform automatic disease detection on the original images and the enhanced images in this chapter. The results are shown in Fig. 7. Among them, DR represents diabetic retinopathy, ARMD represents age-related macular degeneration, MH represents media haze, BRVO represents branch retinal vein occlusion, ODC represents optic disc cupping, and ODE represents optic disc edema. We use recall and F1 metrics to evaluate the classification results. The results are shown in Table 7. It can be seen that the images enhanced by our model can achieve better classification results than the original images. Specifically, for ARMD, the images enhanced by our method show improvements of 0.36 and 0.39 in recall and F1 metrics, respectively. For ODE, the enhanced images show improvements of 0.47 and 0.29 in recall and F1 metrics, respectively. Especially for images with MH, the images enhanced by our method can achieve recall and F1 metrics of 0.94 and 0.96. This also demonstrates that our enhancement model can contribute to further automated disease detection and diagnosis, while effectively preserving the structural information of the fundus during the enhancement process, thus enhancing the accuracy of automated diagnosis.

Conclusion

In this paper, we have proposed a two-stage multi-scale attention-based network (TSMSA-Net) for weakly supervised cataract fundus image enhancement. To obtain more paired cataract fundus images which are close to the realistic scenarios, we have proposed a real-like cataract fundus image synthesis stage. To better utilize the features of the fundus images, we have proposed a multi-scale attention-based cataract fundus image enhancement stage, which extracts the structural features from different scales to facilitate better image enhancement. Extensive experiments have demonstrated that our TSMSA-Net favors against state-of-the-art cataract fundus image enhancement approaches. Furthermore, TSMSA-Net can improve the results of blood vessel segmentation and automatic disease detection tasks and can improve the accuracy of classification. So it can be used as a pre-processing of computer-aided algorithms for the facilitate diagnosis of ocular diseases. In future work, we will pay more attention to optic disc besides vessels to make the diseases easier for ophthalmologists to distinguish.

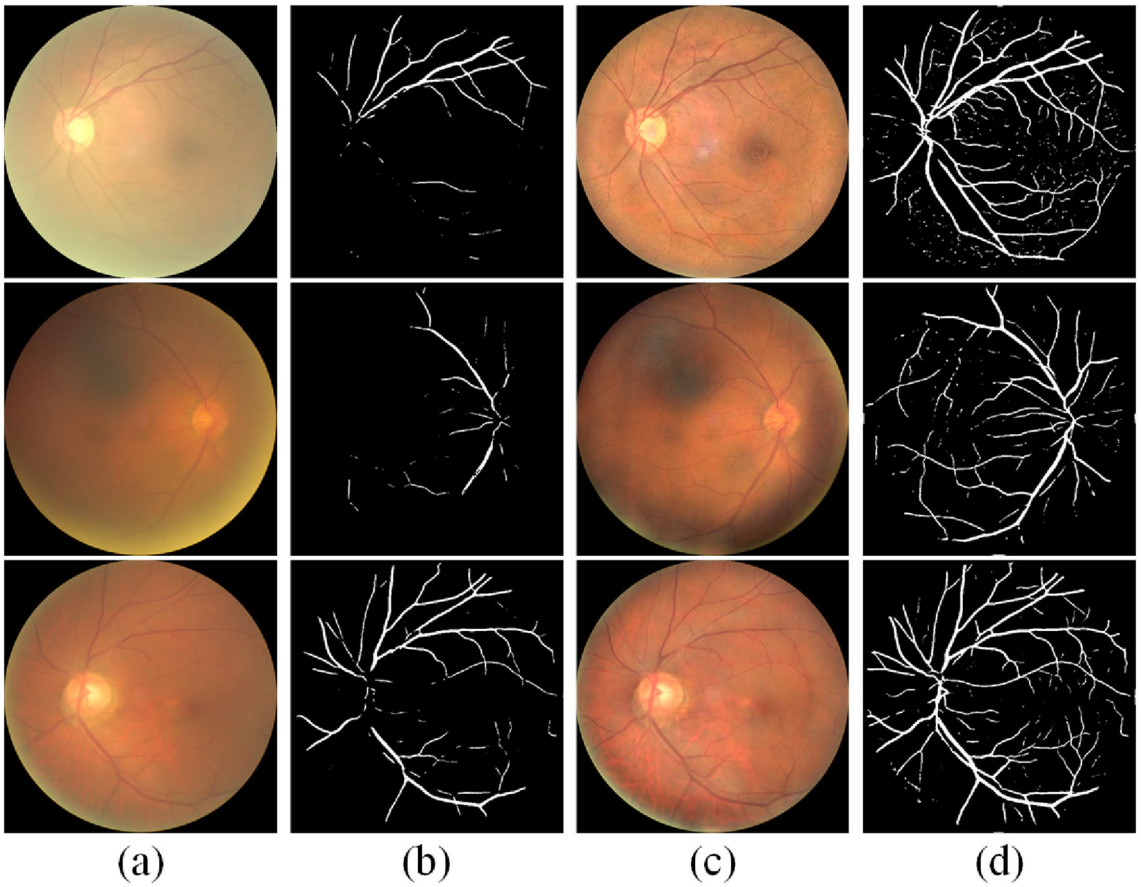

Fig. 7. Application in vessel segmentation (**a**) Low-quality image, (**b**) segmentation result of the low-quality image, (**c**) enhanced image, (**d**) segmentation result of the enhanced image.

	Original image		Enhanced image	
Disease	Recall	F1	Recall	F1
DR	0.61	0.73	0.65	0.77
ARMD	0.06	0.12	0.42	0.51
MH	0.79	0.81	0.94	0.96
BRVO	0.39	0.56	0.61	0.72
ODC	0.26	0.36	0.36	0.49
ODE	0.29	0.43	0.76	0.72

Table 7. Accuracy of automatic disease diagnosis.

Data availability

Te datasets analyzed during the current study are available at ODIR-5K dataset [https://github.com/linhandev/dataset].

Received: 27 July 2024; Accepted: 15 July 2025
Published online: 29 July 2025

References

1. Özbay, E. An active deep learning method for diabetic retinopathy detection in segmented fundus images using artificial bee colony algorithm. *Artif. Intell. Rev.* **56**, 3291–3318 (2023).
2. Peli, E. & Schwartz, B. Enhancement of fundus photographs taken through cataracts. *Ophthalmology* **94**, 10–13 (1987).
3. Alwazzan, M. J., Ismael, M. A. & Ahmed, A. N. A hybrid algorithm to enhance colour retinal fundus images using a Wiener filter and Clahe. *J. Digit. Imaging* **34**, 750–759 (2021).
4. Dash, S. et al. Guidance image-based enhanced matched filter with modified thresholding for blood vessel extraction. *Symmetry* **14**, 194 (2022).

5. Li, T. et al. Applications of deep learning in fundus images: A review. *Med. Image Anal.* **69**, 101971 (2021).
6. Priyadharsini, C. et al. Retinal image enhancement based on color dominance of image. *Sci. Rep.* **13** (2023).
7. Yu, J., Chen, K. H., Lucero, R., Ambrosi, C. M. & Entcheva, E. Cardiac optogenetics: Enhancement by all-trans-retinal. *Sci. Rep.* **5** (2015).
8. Wu, H.-T. et al. Fundus image enhancement via semi-supervised GAN and anatomical structure preservation. In *IEEE Transactions on Emerging Topics in Computational Intelligence* (2023).
9. Li, H. et al. An annotation-free restoration network for cataractous fundus images. *IEEE Trans. Med. Imaging* **41**, 1699–1710 (2022).
10. Li, H. et al. A generic fundus image enhancement network boosted by frequency self-supervised representation learning. *Med. Image Anal.* **90**, 102945 (2023).
11. Abbasi, M. M., Iqbal, S., Aurangzeb, K., Alhussein, M. A. & Khan, T. M. Lmbis-net: A lightweight bidirectional skip connection based multipath cnn for retinal blood vessel segmentation. *Sci. Rep.* **14** (2024).
12. Liu, H. et al. Degradation-invariant enhancement of fundus images via pyramid constraint network. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 507–516 (2022).
13. Zhang, W. et al. A fundus image enhancer based on illumination-guided attention and optic disc perception GAN. *Optik* **279**, 170729 (2023).
14. Li, H. et al. Structure-consistent restoration network for cataract fundus image enhancement. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 487–496 (Springer, 2022).
15. Yang, B. et al. Retinal image enhancement with artifact reduction and structure retention. *Pattern Recognit.* **133**, 108968 (2023).
16. Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. 234–241 (2015).
17. Pizer, S. M. et al. Adaptive histogram equalization and its variations. *Comput. Vis. Graph. Image Process.* **39**, 355–368 (1987).
18. Zhou, M., Jin, K., Wang, S., Ye, J. & Qian, D. Color retinal image enhancement based on luminosity and contrast adjustment. *IEEE Trans. Biomed. Eng.* **65**, 521–527 (2017).
19. Ke, A., Luo, J. & Cai, B. Unet-like network fused swin transformer and cnn for semantic image synthesis. *Sci. Rep.* **14**(1), 16761 (2024).
20. Ding, J.-T., Peng, Y.-Y., Huang, M. & Zhou, S.-J. Agrigan: Unpaired image dehazing via a cycle-consistent generative adversarial network for the agricultural plant phenotype. *Sci. Rep.* **14** (2024).
21. Sharma, A. M. et al. Enhanced low-light image fusion through multi-stage processing with Bayesian analysis and quadratic contrast function. *Sci. Rep.* **14** (2024).
22. Peng, L., Zhu, C. & Bian, L. U-shape transformer for underwater image enhancement. In *IEEE Transactions on Image Processing* (2023).
23. Gao, G. et al. Ctcnet: A cnn-transformer cooperation network for face image super-resolution. *IEEE Trans. Image Process.* **32**, 1978–1991 (2023).
24. Jiang, Y. et al. Enlightengan: Deep light enhancement without paired supervision. *IEEE Trans. Image Process.* **30**, 2340–2349 (2021).
25. Deng, Z. et al. Rformer: Transformer-based generative adversarial network for real fundus image restoration on a new clinical benchmark. *IEEE J. Biomed. Health Inform.* **26**, 4645–4655 (2022).
26. Raj, A., Shah, N. A. & Tiwari, A. K. A novel approach for fundus image enhancement. *Biomed. Signal Process. Control* **71**, 103208 (2022).
27. Shen, Z., Fu, H., Shen, J. & Shao, L. Modeling and enhancing low-quality retinal fundus images. *IEEE Trans. Med. Imaging* **40**, 996–1006 (2020).
28. Luo, Y. et al. Dehaze of cataractous retinal images using an unpaired generative adversarial network. *IEEE J. Biomed. Health Inform.* **24**, 3374–3383 (2020).
29. Vasa, V. K. et al. Context-aware optimal transport learning for retinal fundus image enhancement. arXiv preprint arXiv:2409.07862 (2024).
30. Guo, E., Fu, H., Zhou, L. & Xu, D. Bridging synthetic and real images: A transferable and multiple consistency aided fundus image enhancement framework. *IEEE Trans. Med. Imaging* **42**, 2189–2199 (2023).
31. Shakhnoza, M., Umirzakova, S., Sevara, M. & Cho, Y. Enhancing medical image denoising with innovative teacher-student model-based approaches for precision diagnostics. *Sensors* **23**, 9502 (2023).
32. Zhu, J.-Y., Park, T., Isola, P. & Efros, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*. 2223–2232 (2017).
33. Peli, E. & Peli, T. Restoration of retinal images obtained through cataracts. *IEEE Trans. Med. Imaging* **8**, 401–406 (1989).
34. Johnson, J., Alahi, A. & Fei-Fei, L. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*. 694–711 (Springer, 2016).
35. Isola, P., Zhu, J.-Y., Zhou, T. & Efros, A. A. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1125–1134 (2017).
36. Li, C. & Wand, M. Precomputed real-time texture synthesis with Markovian generative adversarial networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*. 702–716 (Springer, 2016).
37. Yao, L., Lin, Y. & Muhammad, S. An improved multi-scale image enhancement method based on Retinex theory. *J. Med. Imaging Health Inform.* **8**, 122–126 (2018).
38. Wang, Y., Li, Y., Wang, G. & Liu, X. Multi-scale attention network for single image super-resolution. arXiv preprint arXiv:2209.14145 (2022).
39. Guo, J., Pang, Z., Yang, F., Shen, J. & Zhang, J. Study on the method of fundus image generation based on improved GAN. *Math. Probl. Eng.* **2020**, 1–13 (2020).
40. Zhao, H., Yang, B., Cao, L. & Li, H. Data-driven enhancement of blurry retinal images via generative adversarial networks. In *Medical Image Computing and Computer Assisted Intervention*. 75–83 (2019).
41. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014).
42. Zhou, P., Lu, C., Feng, J., Lin, Z. & Yan, S. Tensor low-rank representation for data recovery and clustering. *IEEE Trans. Pattern Anal. Mach. Intell.* **43**, 1718–1732 (2021).
43. Singh, A., Chougule, A., Narang, P., Chamola, V. & Yu, F. R. Low-light image enhancement for UAVs with multi-feature fusion deep neural networks. *IEEE Geosci. Remote. Sens. Lett.* **19**, 1–5 (2022).
44. Xian, Y., Zhao, G., Wang, C., Chen, X. & Dai, Y. A novel hybrid retinal blood vessel segmentation algorithm for enlarging the measuring range of dual-wavelength retinal oximetry. *Photonics* **10** (2023).
45. Likassa, H. T., Chen, D., Chen, K., Wang, Y. & Zhu, W. Robust PCA with lw,$_*$ and l2,1 norms: A novel method for low-quality retinal image enhancement. *J. Imaging* **10**, 151 (2024).
46. Staal, J., Abràmoff, M. D., Niemeijer, M., Viergever, M. A. & Van Ginneken, B. Ridge-based vessel segmentation in color images of the retina. *IEEE Trans. Medical Imaging* **23**, 501–509 (2004).

47. Hoover, A., Kouznetsova, V. & Goldbaum, M. Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response. *IEEE Trans. Med. Imaging* **19**, 203–210 (2000).
48. Pachade, S. et al. Retinal fundus multi-disease image dataset (rfmid): A dataset for multi-disease detection research. *Data* **6**, 14 (2021).
49. Liu, Z. et al. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11976–11986 (2022).

Acknowledgements
This work was supported in part by Hunan Provincial Natural Science Foundation of China (Grant No. 2022JJ30017), Hunan Provincial Natural Science Foundation of China (Grant No. 2025JJ70174), National Key Research and Development Program of China (Grant No. 2021ZD0112400), National Natural Science Foundation of China (Grant No. U1908214), the Program for Innovative Research Team in University of Liaoning Province (Grant No. LT2020015), the Support Plan for Key Field Innovation Team of Dalian (2021RT06), the Support Plan for Leading Innovation Team of Dalian University (XLJ202010), 111 Center (No. D23006), Interdisciplinary project of Dalian University (Grant No. DLUXK-2025-QNLG-001).

Author contributions
Xiaoyong Fang wrote the main manuscript text, Yue Wang and Xiangyu Li conducted the experiments, Wanshu Fan and Dongsheng Zhou analysed the results. All authors reviewed the manuscript.

Declarations

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to W.F. or D.Z.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.