Original Article

Dual contrastive learning for synthesizing unpaired fundus fluorescein angiography from retinal fundus images

Jiashi Zhao[1,2]^, Haiyi Huang[1,2], Cheng Wang[3]^, Miao Yu[1,2]^, Weili Shi[1,2]^, Kensaku Mori[3,4]^, Zhengang Jiang[1,2]^, Jianhua Liu[5]

[1]School of Computer Science and Technology, Changchun University of Science and Technology, Changchun, China; [2]School of Computer Science and Technology, Zhongshan Institute of Changchun University of Science and Technology, Zhongshan, China; [3]Graduate School of Informatics, Nagoya University, Nagoya, Japan; [4]Research Center for Medical Bigdata, National Institute of Informatics, Tokyo, Japan; [5]Department of Radiology, The Second Hospital of Jilin University, Changchun, China

Contributions: (I) Conception and design: J Zhao, H Huang, C Wang, K Mori; (II) Administrative support: J Zhao, Z Jiang; (III) Provision of study materials or patients: J Liu; (IV) Collection and assembly of data: H Huang; (V) Data analysis and interpretation: J Zhao, H Huang; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

Correspondence to: Zhengang Jiang, PhD. School of Computer Science and Technology, Changchun University of Science and Technology, 7186 Satellite Road, Changchun 130022, China; School of Computer Science and Technology, Zhongshan Institute of Changchun University of Science and Technology, 16 Huizhan Dong Road, Torch Development Zone, Zhongshan 528436, China. Email: jiangzhengang@cust.edu.cn.

Background: Fundus fluorescein angiography (FFA) is an imaging method used to assess retinal vascular structures by injecting exogenous dye. FFA images provide complementary information to that provided by the widely used color fundus (CF) images. However, the injected dye can cause some adverse side effects, and the method is not suitable for all patients.

Methods: To meet the demand for high-quality FFA images in the diagnosis of retinopathy without side effects to patients, this study proposed an unsupervised image synthesis framework based on dual contrastive learning that can synthesize FFA images from unpaired CF images by inferring the effective mappings and avoid the shortcoming of generating blurred pathological features caused by cycle-consistency in conventional approaches. By adding class activation mapping (CAM) to the adaptive layer-instance normalization (AdaLIN) function, the generated images are made more realistic. Additionally, the use of CAM improves the discriminative ability of the model. Further, the Coordinate Attention Block was used for better feature extraction, and it was compared with other attention mechanisms to demonstrate its effectiveness. The synthesized images were quantified by the Fréchet inception distance (FID), kernel inception distance (KID), and learned perceptual image patch similarity (LPIPS).

Results: The extensive experimental results showed the proposed approach achieved the best results with the lowest overall average FID of 50.490, the lowest overall average KID of 0.01529, and the lowest overall average LPIPS of 0.245 among all the approaches.

Conclusions: When compared with several popular image synthesis approaches, our approach not only produced higher-quality FFA images with clearer vascular structures and pathological features, but also achieved the best FID, KID, and LPIPS scores in the quantitative evaluation.

Keywords: Fundus fluorescein angiography (FFA); adaptive layer-instance normalization (AdaLIN); unsupervised; image synthesis; dual contrastive learning

^ ORCID: Jiashi Zhao, 0000-0001-7756-1226; Cheng Wang, 0000-0003-0604-5245; Miao Yu, 0000-0001-9513-3292; Weili Shi, 0000-0001-6347-739X; Kensaku Mori, 0000-0002-0100-4797; Zhengang Jiang, 0000-0001-6315-243X.

Submitted Sep 15, 2023. Accepted for publication Dec 21, 2023. Published online Feb 23, 2024.
doi: 10.21037/qims-23-1304
View this article at: https://dx.doi.org/10.21037/qims-23-1304

Introduction

Fundus fluorescein angiography (FFA) is a significant diagnostic tool for understanding pathophysiological mechanisms and guiding treatment (1). Sodium fluorescein is injected into the human body through a vein and the structure of the retinal blood vessels is imaged through the blood circulation. However, the application of this technology causes varying degrees of damage to patients' bodies. Further, fluorescein leakage may occur if the procedure is not performed correctly. It is also unsuitable for some patients, such as the elderly, those with poor kidney function, hyperglycemia, hypertension, and allergies. Therefore, it is not easy to obtain high-quality FFA images.

Optical coherence tomography (OCT) (2) has become increasingly popular in recent years, as it is radiation-free, non-invasive, high resolution, and affordable. High resolution cross-sectional imaging of the retina is possible; however, it requires a high level of patient cooperation during image acquisition and cannot provide direct information about blood flow. Further, if patients are not careful, severe motion artifacts or data loss may occur. Optical coherence tomography angiography (OCTA) is an emerging, non-invasive imaging modality that provides information about the vascular structure of the retina and choroid, which can provide a detailed view of vascular occlusion and vascular compromise by the quantitative assessment of the severity of diabetic retinopathy (3). Further, OCTA can also provide high-resolution vascular images that display vascular abnormalities in the macular region, including choroidal neovascularization in patients with wet age-related macular degeneration. However, due to the high cost of the equipment, OCTA is not widely available at some small-sized hospitals.

At present, color fundus (CF) images (4) are the main tool for the diagnosis of retinal diseases, as the CF method is more accessible than most other methods and the equipment for this method is cheaper than that required by other methods. However, due to the low resolution of CF images, it is not entirely reliable as a diagnostic tool for retinal diseases.

The appearance of CF and FFA images differ significantly. As existing FFA image synthesis methods cannot accurately represent the pathological location and corresponding pathological structure of the retina, their clinical application value is low. In addition, due to the insufficient number and type of data sets, the models have poor generalizability. A more effective method of converting CF images to higher quality FFA images in unpaired data sets, without increasing the risk to patients, may reduce the need for authentic FFA images and improve the diagnostic accuracy of retinal disease.

Medical image synthesis is mainly used to: (I) increase the data diversity, which traditional data enhancement methods cannot effectively do; (II) reduce the acquisition time and the risk of side effects for patients; and (III) convert multimodal images to unimodal images before image registration to avoid the problems caused by inconsistent correspondence features and improve the registration accuracy to some extent.

Synthesizing FFA images from CF images is equivalent to image-to-image (I2I) translation (i.e., translation from the CF domain to the FFA domain). With the emergence of deep learning, generative adversarial networks (GANs) (5) have been widely used to solve I2I problems, such as image style transfer (6), image denoising (7), and image super-resolution (8). GANs have also been applied to medical imaging, such as computed tomography (CT) synthesis from magnetic resonance imaging (MRI) (9,10), super-resolution cardiac MRI (11), and positron emission tomography to CT translation (12).

The proposed FFA image synthesis approaches can be split into supervised (paired) and unsupervised (unpaired) approaches, depending on whether the data sets are paired or not. Researchers have proposed supervised approaches (13,14) that learn the direct mapping between two domains based on the U-Net structure (15). However, such approaches are unsatisfactory, as the methods can only learn the pixel-to-pixel mapping. In addition, multimodal image registration followed by image generation can complicate the task of image synthesis. Kamran *et al.* proposed a conditional GAN (CGAN)-based (16) approach for the generation of FFA images from a coarse-to-fine framework (17). Li *et al.* developed a separate representation method (18) based on the CGAN, and while their evaluation index results were better than others, the generated retinal vascular structure and pathological features were inaccurate.

Subsequently, Tavakkoli *et al.* (19) increased the number of discriminators from two to four. Based on this, Kamran *et al.* added an attention mechanism (20). After which, Kamran *et al.* proposed the image generation method using vision transformers (21) to replace the discriminator as an improvement (22). The generated normal retinal image was clearer than before; however, there is still room for improvement in the generation of the vascular structure and pathological features of the abnormal retina.

In relation to the unsupervised approaches, the methods proposed by researchers (23,24) have been largely similar to the cycle-consistent GAN (CycleGAN) (25), and the pathological features in the generated images are relatively fuzzy. Cai *et al.* introduced a triple multiscale structure to the CycleGAN to strengthen the similarity between the CF and FFA domains (26). However, due to the cycle-consistent property, the fake images have similar global features to the real images in the target domain, preventing the generation of accurate pathological structures that can be directly used for disease diagnosis. Further, in the absence of supervisory information, it is difficult to obtain better results if the data sets are not large enough.

The pathological features of the retinal images generated by the supervised methods are superior to those generated by the unsupervised methods; however, public paired retinal data sets are scarce, and lack expert annotation, which has limited the value of supervised methods in clinical applications. The CycleGAN structure is the mainstay of current unsupervised FFA synthesis approaches. The results obtained by these unsupervised methods are unsatisfactory in terms of the vascular structure and pathological characteristics, and thus cannot be applied to the clinical diagnosis of retinal diseases.

Several unsupervised I2I translation approaches based on cycle-consistency have been proposed. Based on the CycleGAN, the unsupervised generative attentional network with adaptive layer-instance normalization for image-to-image translation (U-GAT-IT) (27) model developed for better feature embedding that relies on class activation mapping (CAM) (28) and the adaptive layer-instance normalization (AdaLIN) function (27) and has all the advantages of the adaptive instance normalization (AdaIN) (29), instance normalization (IN) (30) and layer normalization (LN) (31). However, the U-GAT-IT can only ensure that the basic shape of real and fake images is not significantly changed. Additionally, too many discriminators may generate some unnecessary parts. However, the CycleGAN and U-GAT-IT rely on the pixel information of the input images to achieve image reconstruction, which may introduce some errors. The geometry-consistent GAN (32) replaced the cycle-consistency with the geometry-consistency, which may help the model to reduce semantic errors during the I2I translation.

On the strength of contrastive learning, some approaches have been put forward to address the problems of I2I translation. Contrastive learning for unpaired I2I translation (CUT) (33) introduced contrastive learning to maximize the mutual information shared between corresponding patches of real and fake images. CUT achieves the efficiency of contrastive learning by adopting an embedding approach between two domains that focuses only on what the two domains have in common and ignores the differences between them. Based on the CycleGAN and CUT, the dual contrastive learning GAN (DCLGAN) (34) learns the correspondence between real and fake image patches to maximize mutual information using a separate embedding block, which avoids the disadvantage of the cycle-consistency.

Models can sometimes be improved by the addition of some attention mechanisms, such as the commonly used channel attention mechanisms self-attention (35), the squeeze-and-excitation (SE) block (36), and the spatial and channel squeeze-and-excitation (scSE) block (37). However, these methods only focus on the information on the channel and ignore other information. The convolutional block attention module (CBAM) (38) focuses on the information on the channel, height, and weight, but may yield some ambiguous information. The coordinate attention (CA) block (39) takes both horizontal and vertical information into consideration, which may help the model generate more accurate information than the aforementioned methods.

The main contributions of this study are summarized as follows: (I) inspired by the DCLGAN and lightweight unsupervised generative attentional network with AdaLIN for I2I translation (L-U-GAT-IT), we first developed an unsupervised dual contrastive learning framework for FFA image synthesis, called the dual contrastive learning attention GAN (DCLAGAN), which also addressed the issue of multimodal retinal image registration. The correspondence between the real and generated image patches can be learned using a separate embedding block to maximize mutual information and derive effective mappings between unpaired images. Dual contrastive learning can avoid the fuzzy pathological features caused by cycle-consistency in conventional unsupervised approaches and improve the reference value of the generated images

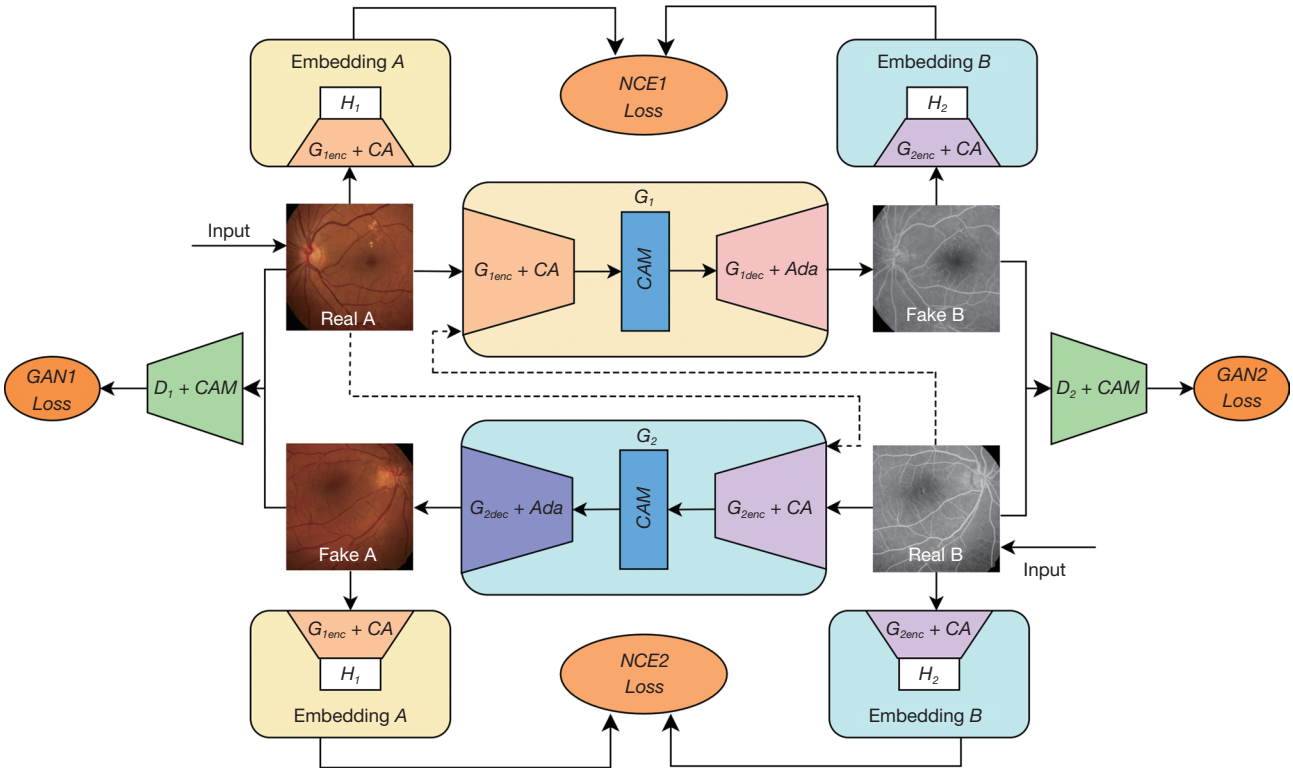

Figure 1 The overall framework of the DCLAGAN. G_1 and G_2, generators. D_1 and D_2, discriminators. G_{1enc} and G_{2enc}, the encoders of G_1 and G_2, respectively. G_{1dec} and G_{2dec}, the decoders of D_1 and D_2, respectively. H_1 and H_2, two-layer MLP projections. NCE, noise contrastive estimation; CA, coordinate attention block; CAM, class activation mapping; GAN, generative adversarial network; Ada, ResnetAdaLINBlocks.

in retinal diseases; (II) DCLAGAN used the CAM with an AdaLIN function (CAM-A) based on global average pooling (GAP) and global max pooling (GMP) as an attention-guiding model for the generator to distinguish whether an area is vital or not and to make the generated images more realistic. A similar CAM structure was also added to the discriminator to help the discriminator to more accurately distinguish between the generated and real images; (III) DCLAGAN appended CA blocks to the encoder in the generator to improve the process of dual contrastive learning, which can achieve more useful positional information for feature extraction and more accurately locate the corresponding vascular structures in the generated images.

Methods

Network architecture

The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). *Figure 1* depicts the proposed DCLAGAN, which includes domains *A* (CF images) and *B* (FFA images), generators G_1 and G_2, and discriminators D_1 and D_2. We define the encoders of G_1 and G_2 as G_{1enc}, and G_{2enc}, respectively, and the decoders of G_1 and G_2 as G_{1dec} and G_{2dec}, respectively. Generators are used to translate the input images into another image domain. To ensure that the generated images are in the correct domain, discriminators are used to determine whether the input images are real or fake.

We set G_{1enc} with H_1 [two-layer multilayer perceptron (MLP) projection] as embedding *A*, and G_{2enc} with H_2 (two-layer MLP projection) as embedding *B*. Image features can be extracted from the second and third encoder layers, and the first and last residual neural network blocks (ResnetBlocks). The information extracted from the four layers can be sent to H_1 or H_2. We define the CA block as *CA*, the ResnetAdaLINBlocks as *Ada*, the patch noise contrastive estimation (PatchNCE) loss as NCE1 and NCE2 loss, and the adversarial loss as GAN1 and GAN2 loss.

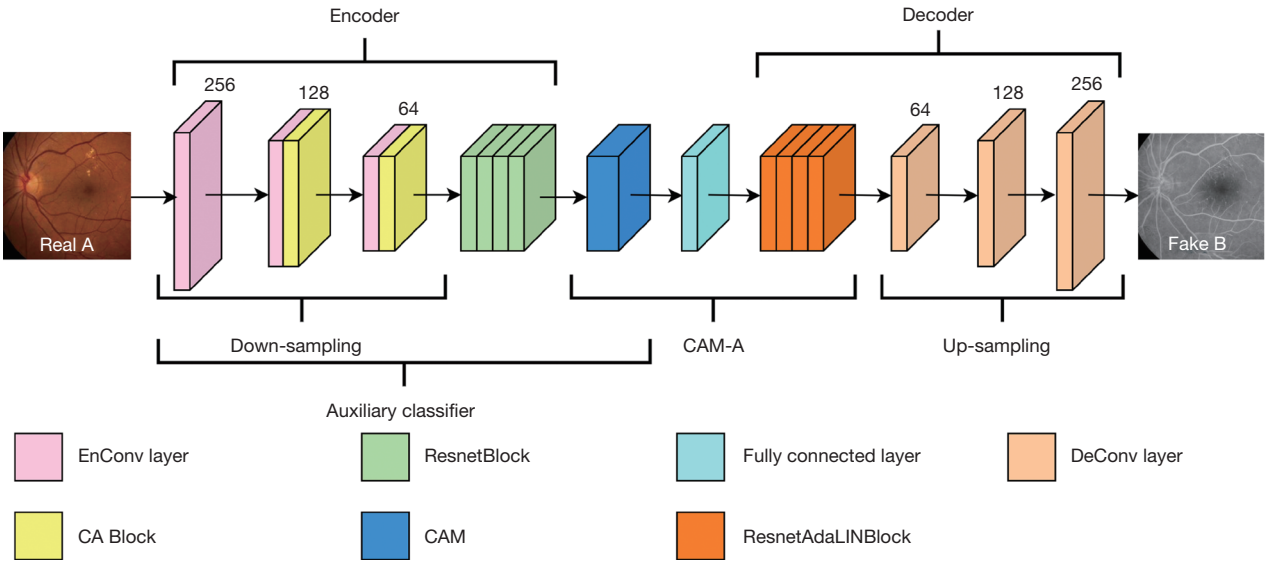

Figure 2 The structure of the generator G_1. EnConv layer, the convolution layer in the encoder; DeConv layer, the convolution layer in the decoder; CA Block, coordinate attention block; CAM, class activation mapping; CAM-A, class activation mapping with adaptive layer-instance normalization function.

Real images are fed into the generator and the other generator with identity loss to produce the fake and identity images, respectively. By the dual contrastive learning of two mappings (G_1: $A \rightarrow B$ and G_2: $B \rightarrow A$), the translation between the CF and FFA images can be realized.

Generator

Figure 2 uses the mapping direction of G_1: $A \rightarrow B$ as an example. We input the real A image into the four ResnetBlocks and perform down-sampling to output the encoder feature map. We use the encoder feature map as input for the CAM-A, including the CAM, fully connected (FC) layers, and ResnetAdaLINBlocks. After which, we generate the attention feature map based on the auxiliary classifier η_1. Through GAP and the FC layers, the parameters of γ and β are input into the ResnetAdaLINBlocks. Ultimately, the generated images (fake B) are produced by up-sampling.

Figure 3A shows the detailed structure of the CAM. Specifically, the working process of CAM can be split into two main steps. First, through the GAP, GMP, GMP_FC, and GAP_FC, we can obtain the auxiliary classifier, which can calculate the probability that the input images are from the CF domain or FFA domain. Second, by multiplying the encoder feature map and the weights of each encoder feature map generated by the auxiliary classifier, we can obtain the GMP and GAP attention feature map, and join them together based on the auxiliary classifier. Finally, we can obtain the results of the attention feature maps through the convolution layer and the activation function. CAM can determine the significance of image regions by using its' superior localization ability. The attention feature map output from the CAM is used as the input to the FC layer, and the output provides two parameters, γ and β. Next, the obtained parameters are substituted into the ResnetAdaLINBlocks, and the synthesized image is finally generated by the decoder. In this work, the attention guidance model CAM-A not only supports the model by ensuring better feature extraction by focusing on the vital information, but also makes the generated images look more realistic.

Figure 3B,3C show the structures of ResnetBlock and ResnetAdaLINBlock. Among them, the ReflectionPad, Convolution and InstanceNormed layers are included in the ResnetBlock structure. The ReflectionPad layers, the Convolution layers and the AdaLIN functions are contained in the structure of the ResnetAdaLINBlock. The AdaLIN function in the ResnetAdaLINBlock can flexibly maintain or change the information of the images according to the actual situation by dynamically adjusting the ratio of IN to LN; IN is a normalization operation for each channel in

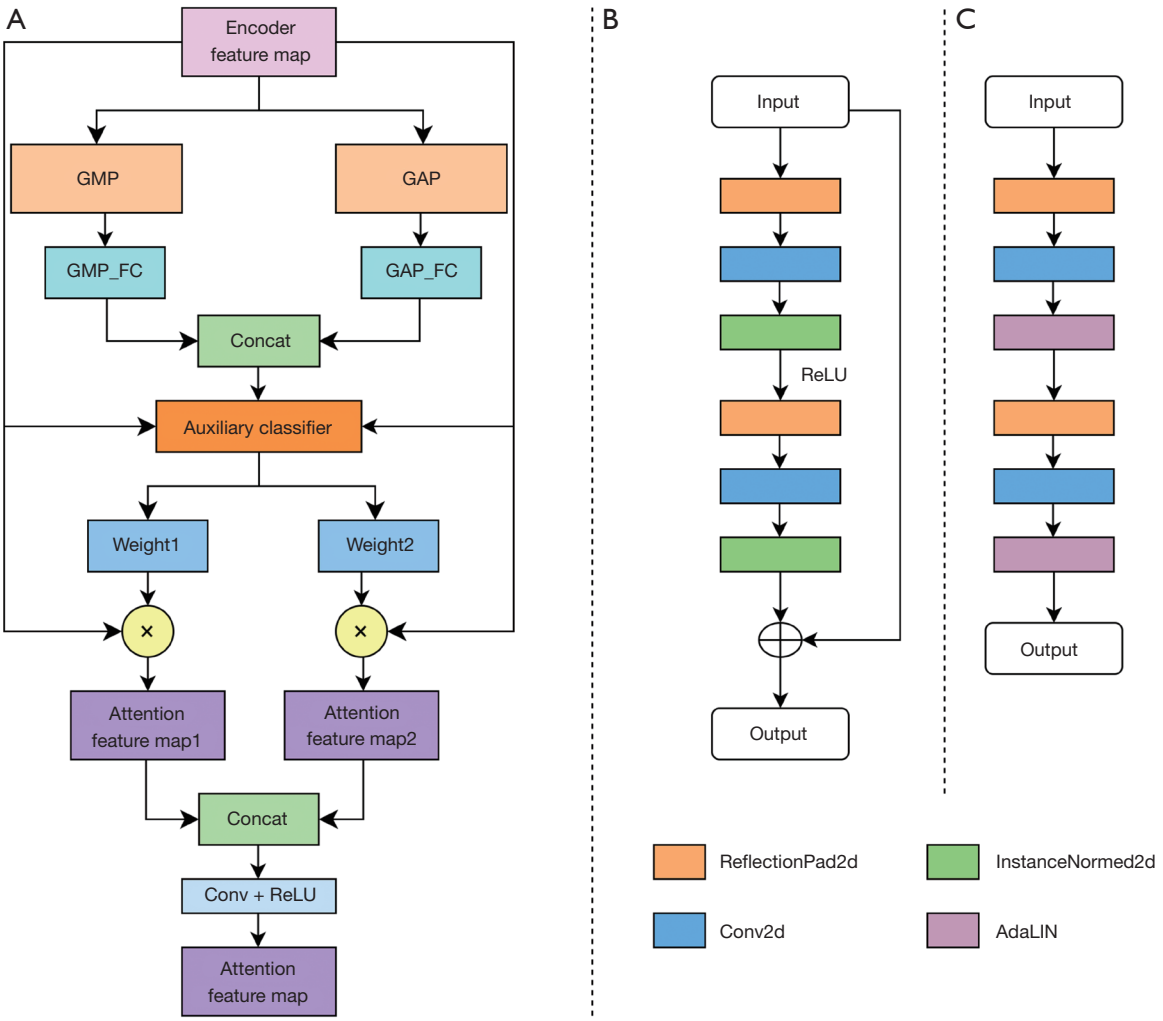

Figure 3 The structures of the CAM, ResnetBlock, and ResnetAdaLINBlock. (A) The structure of CAM. (B) The structure of ResnetBlock. (C) The structure of ResnetAdaLINBlock. GMP, global max pooling; GAP, global average pooling; FC, the fully connected layer; Concat, the concatenation operation; Weight, the weight of each encoder feature map generated by the auxiliary classifier; Conv, the convolution layer; AdaLIN, adaptive layer-instance normalization; ResnetBlocks, residual neural network blocks; ReLU, the rectified linear unit activation function; CAM, class activation mapping.

each image, which can maintain the independence between each image and preserve the content information in the source images; LN is a normalization operation for different channels of the same layer in each image, which can preserve different features in different channels and ensure that the different features between the generated images can be attenuated simultaneously. Specifically, LN allows more thorough style translation of the generated images than IN.

The AdaLIN function is expressed as:

$$AdaLIN(\alpha,\gamma,\beta)=\gamma\cdot\left(\rho\cdot\hat{\alpha}_{IN}+(1-\rho)\cdot\hat{\alpha}_{LN}\right)+\beta \quad [1]$$

$$\hat{\alpha}_{IN}=\frac{\alpha-\mu_{IN}}{\sqrt{\sigma_{IN}^2+\chi}},\quad \hat{\alpha}_{LN}=\frac{\alpha-\mu_{LN}}{\sqrt{\sigma_{LN}^2+\chi}},\quad \rho\leftarrow clip_{[0,1]}(\rho-\eta\Delta\rho) \quad [2]$$

where μ_{IN}, μ_{LN}, and σ_{IN}^2, σ_{LN}^2 represent the mean and variance by normalizing the input images at the channel

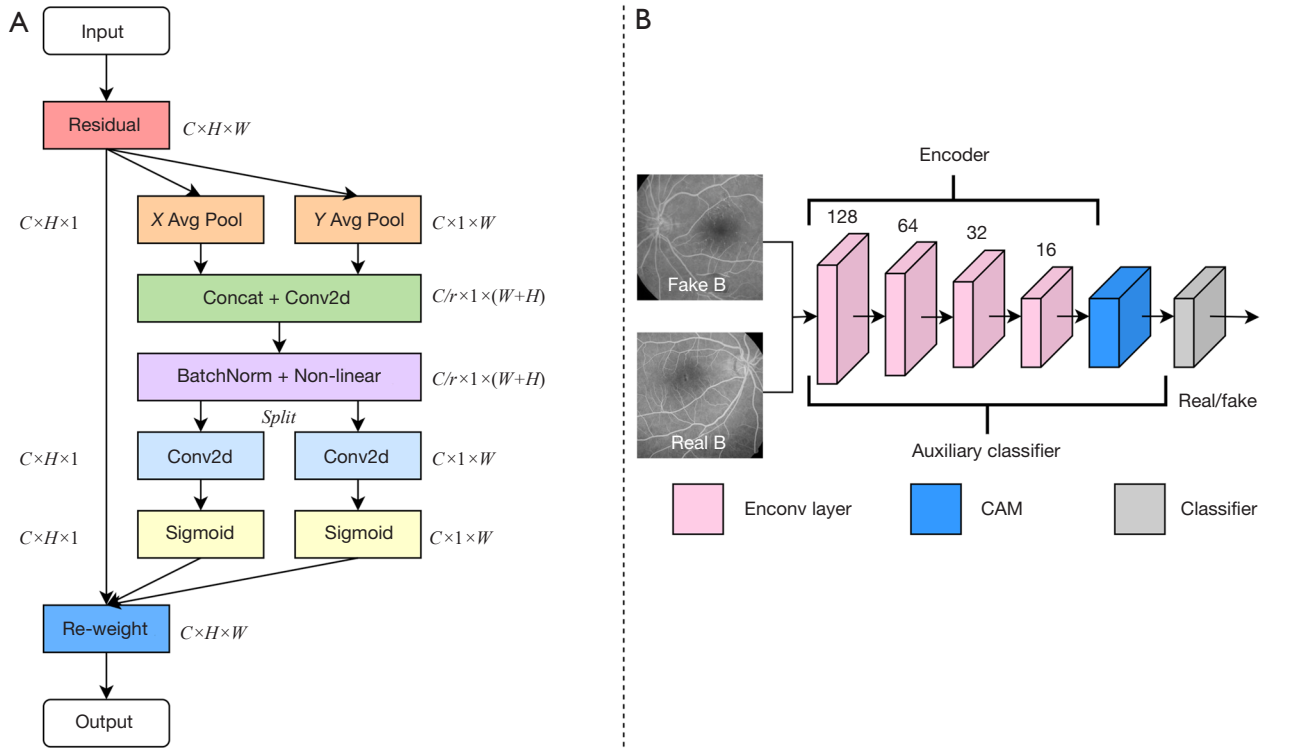

Figure 4 The structures of the CA block and discriminator, D_2. (A) The structure of the coordinate attention block. (B) The structure of the discriminator D_2. X Avg Pool, 1D horizontal global pooling; Y Avg Pool, 1D vertical global pooling; Conv2d, 2D convolution layer; BatchNorm, batch normalization; Non-linear, non-linear activation function; Sigmoid, sigmoid activation function; Enconv layer, the convolution layer in encoder; CAM, class activation mapping; C, the channel; H, the height; W, the width; r, the reduction ratio.

and layer levels, respectively. The parameters of γ and β are dynamically obtained from the output of FC. The ρ is the probability value from 0 to 1 that can adaptively balance the ratio between IN and LN. The initial value of ρ in ResnetAdaLINBlocks is 1, while the initial value in up-sampling is 0. η represents the learning rate. $\Delta\rho$ represents the update vector (27).

The realization of the CA block (*Figure 4A*) can be split into two steps. First, the coordinate information is embedded to obtain accurate position information. The output channels of the c-th at h or w are expressed as:

$$p_c^h(h)=\frac{1}{W}\sum_{0\le m\le W}x_c(h,m) \quad [3]$$

$$p_c^w(w)=\frac{1}{H}\sum_{0\le n\le H}x_c(n,w) \quad [4]$$

Second, the coordinated information is generated. A concatenation operation and a convolution function F are used to reduce the dimension of C and integrate the feature maps of the W and H orientations. Next, the weight information of each dimension can be obtained by using batchnorm2d and the non-linear activation function δ as follows:

$$f=\delta\left(F\left(\left[p^h,p^w\right]\right)\right),\quad f\in R^{C/r\times(H+W)} \quad [5]$$

where $[\cdot,\cdot]$ represents the concatenation operation. The $f\in R^{C/r\times(H+W)}$ represents the feature map that can encode spatial information in the H and W directions. The r can be used to control the size of the block. Further, the split function is used to split f into two tensors, using the F function and the activation function σ to generate each channel of weight along the H or W orientations, which are expressed as:

$$g^h=\sigma\left(F_h\left(f^h\right)\right),\quad f^h\in R^{C/r\times H} \quad [6]$$

$$g^w=\sigma\left(F_w\left(f^w\right)\right),\quad f^w\in R^{C/r\times W} \quad [7]$$

The entire function is expressed as:

$$y_c(m,n)=x_c(m,n)\times g_c^h(m)\times g_c^w(n) \quad [8]$$

Unlike the widely used channel attention mechanisms, the self-attention and SE blocks focus only on channel attention and ignore information about H and W orientations. When these attention mechanisms are added to the proposed model, the vascular structures in the generated images become fuzzier than before. Some attention mechanisms, such as the scSE block and CBAM, combined with the information from H, W, and C, cannot obtain the information of the exact locations. The CA block can embed spatial coordinate information into the channel attention mechanism (39). In this way, generated images with more accurate locations of vascular structures and pathological features can be obtained.

Discriminator

Figure 4B uses D_2 as an example. D_2 is a multi-scale structure that includes the encoder, the auxiliary classifier ηD_2, and the classifier. By inputting real B and fake B images, the D_2 encoder can output the encoder feature map. The encoder feature map is then used as the input for the CAM. The output of the attention feature map from the CAM based on ηD_2 can optimize the classifier by concentrating on the difference between the real and generated images in domain B.

Loss functions

The proposed approach has the following four loss functions: GAN loss, PatchNCE loss, CAM loss, and identity loss. The GAN loss ensures that the fake images are as similar as possible to the target domain. For the mapping G_1: $A \rightarrow B$, the GAN1 loss is expressed as:

$$L_{GAN1}(G_1, D_2, A, B) = E_{b\sim B}\left[\log D_2(b)\right] + E_{a\sim A}\left[\log\left(1 - D_2(G_1(a))\right)\right] \quad [9]$$

where G_1 encourages $G_1(a)$ to look similar to the target domain, while D_2 tries to distinguish $G_1(a)$ from real images in domain B. Analogously, for the mapping G_2: $B \rightarrow A$, the GAN2 loss is expressed as:

$$L_{GAN2}(G_2, D_1, A, B) = E_{a\sim A}\left[\log D_1(a)\right] + E_{b\sim B}\left[\log\left(1 - D_1(G_2(b))\right)\right] \quad [10]$$

The contrastive learning problems consist of three signals; that is, the query q, positive p^+, and N negative p^-. q is associated with p^+ and then compared with the signal p^-. q, p^+, and N p^- are mapped to D-dimensional vectors. They are denoted as h, $h^+ \in R^D$, and $h^- \in R^{N\times D}$. The samples are normalized to form an $(N + 1)$ classification problem. The cross-entropy loss is expressed as:

$$\ell\left(h, h^+, h^-\right) = -\log \frac{\exp\left(sim\left(h, h^+\right)/t\right)}{\exp\left(sim\left(h, h^+\right)/t\right) + \sum_{n=1}^{N}\exp\left(sim\left(h, h_n^-\right)/t\right)} \quad [11]$$

where the distance between q and the other samples is scaled by the temperature parameter t, q refers to the output, and p^+ and p^- refer to the corresponding and non-corresponding inputs, respectively.

To ensure that the corresponding patches between fake and real images are as similar as possible and different from other patches, we exploit the noisy contrastive estimation framework (40), which can extract features from domains A and B without sharing weights between the two parts of the embedding blocks for A and B. Thus, we can learn better embedding information and obtain variability in two different image domains A and B. Next, the images are embedded into the stack of features $\{w_l\}_L = \left\{H_1^l\left(G_{1enc}^l(a)\right)\right\}_L$, where G_{1enc}^l represents the output of layer l.

The result of each feature represents an image patch, and each feature has the goal of matching the corresponding patches between real and fake images. The spatial positions in the selected four layers are denoted as $s \in \{1, ..., S_l\}$. The corresponding positive features are denoted as $w_l^s \in R^z$, $z = (S_l - 1)\times C_l$. For $G_1(a)$, we use the embedding B and obtain other stacks of features $\{\hat{w}_l\}_L = \left\{H_2^l\left(G_{2enc}^l\left(G_1(a)\right)\right)\right\}_L$ (34). The mapping G_1: $A \rightarrow B$ PatchNCE1 loss is expressed as:

$$L_{PatchNCE1}(G_1, H_1, H_2, A) = E_{a\sim A}\sum_{l=1}^{L}\sum_{s=1}^{S_l}\ell\left(\hat{w}_l^s, w_l^s, w_l^{S/s}\right) \quad [12]$$

Similarly, the mapping G_2: $B \rightarrow A$ PatchNCE2 loss is expressed as:

$$L_{PatchNCE2}(G_2, H_1, H_2, B) = E_{b\sim B}\sum_{l=1}^{L}\sum_{s=1}^{S_l}\ell\left(\hat{w}_l^s, w_l^s, w_l^{S/s}\right) \quad [13]$$

The attention feature map generated by CAM-A in the generators or CAM in the discriminators are used to learn the areas in which they need to improve or to better distinguish the real images from the fake images. The losses of CAM in the generators are expressed as:

$$L_{GCAM1}(A) = -\left(E_{a\sim A}\left[\log\left(\eta_1(a)\right)\right] + E_{b\sim B}\left[\log\left(1 - \eta_1(a)\right)\right]\right) \quad [14]$$

$$L_{GCAM2}(B) = -\left(E_{b\sim B}\left[\log\left(\eta_2(b)\right)\right] + E_{a\sim A}\left[\log\left(1 - \eta_2(b)\right)\right]\right) \quad [15]$$

Where the η_1 and η_2 refer to the auxiliary classifier in G_1 and G_2. The CAM losses in the discriminators are

expressed as:

$$L_{DCAM1}(G_1, D_2, A, B) = E_{b\sim B}\left(\left(\eta D_2(b)\right)^2\right) + E_{a\sim A}\left[\left(1-\eta D_2\left(G_1(a)\right)\right)^2\right] \quad [16]$$

$$L_{DCAM2}(G_2, D_1, A, B) = E_{a\sim A}\left[\left(\eta D_1(a)\right)^2\right] + E_{b\sim B}\left[\left(1-\eta D_1\left(G_2(b)\right)\right)^2\right] \quad [17]$$

where ηD_1 and ηD_2 are defined as the auxiliary classifiers in D_1 and D_2 respectively.

The identity loss is added to avoid unnecessary changes generated by the generators and to encourage the mappings to maintain the similar colors between the real and fake images in the same domain. The loss of identity is expressed as:

$$L_{\text{Identity}}(G_1, G_2) = E_{a\sim A}\left[\left\|G_2(a)-a\right\|_1\right] + E_{b\sim B}\left[\left\|G_1(b)-b\right\|_1\right] \quad [18]$$

The entire loss function is expressed as:

$$\begin{aligned} &L(G_1, G_2, D_1, D_2, H_1, H_2) \\ &= \lambda_{GAN}\left(L_{GAN1} + L_{DCAM1} + L_{GAN2} + L_{DCAM2}\right) \\ &+ \lambda_{NCE}\left(L_{PatchNCE1} + L_{PatchNCE2}\right) \\ &+ \lambda_{CAM}\left(L_{GCAM1} + L_{GCAM2}\right) + \lambda_{ID} L_{Identity} \end{aligned} \quad [19]$$

Implementation details

We ran our proposed approach in Pytorch. We trained our model and other comparative models on NVIDIA A100 GPU. We set λ_{GAN} =1, λ_{NCE} =1, λ_{CAM} =1,000, λ_{ID} =10, batch size =1. We chose the Adam optimizer with β_1 =0.5 and β_2 =0.999. The model was trained using the following learning rate: 0.0001 up to 200 epochs, and a linearly decaying zero up to 400 epochs.

Data sets

CF-FFA

We compared the DCLAGAN with other comparison models using the public data set provided by Isfahan medical image and signal processing (MISP) (41), which comprises 59 pairs of unaligned CF and FFA images with a resolution of 720×576. Among them, 29 pairs were healthy retinal images from 29 healthy individuals, and 30 pairs were retinal images from 30 individuals with diabetes. We selected 42 pairs as the training data set and extracted eight pairs with a crop size of 512×512 in each image. After data augmentation and image resizing, the training data set comprised 1,680 images with a resolution of 256×256. We used the other 17 pairs as the testing data set and extracted 16 with a random crop in each image. Ultimately, the testing data set comprised 272 pairs of images with a resolution of 256×256. All of the images from the training and testing data sets were converted to the red green blue (RGB) format.

OCT-OCTA

The data set comprised 200 OCT and 200 OCTA projection maps with a resolution of 304×304, including 160 normal and 40 abnormal pairs. After data augmentation and image resizing, we had 800 pairs for training and 200 pairs for testing with a resolution of 256×256. All the data were acquired from the OCTA-500 data set (with a 3-mm view) (42).

Horse-Zebra

The data set comprised 1,068 horse and 1,335 zebra images for training and 120 horse and 140 zebra images for testing with a resolution of 256×256. All the data were acquired from ImageNet (43).

Cat-Dog

The data set comprised 5,153 cat and 4,739 dog images for training and 500 cat and 500 dog images for testing with a resolution of 512×512. All the data were acquired from StarGAN2 (44). The images were resized with a resolution of 256×256 before training and testing.

Metrics

Fréchet inception distance (FID)

The feature distance between real and fake images (45) was measured using the following formula:

$$FID(a,b) = \left\|\mu_a - \mu_b\right\|^2 + Tr\left(\sigma_a + \sigma_b - 2\sqrt{\sigma_a \times \sigma_b}\right) \quad [20]$$

where μ_a and μ_b refer to the mean feature value of the real and generated images, σ_a and σ_b refer to the covariance of the real and generated images, respectively, and *Tr* represents the trace of the matrix. Each image generated 2,048 feature vectors through the Inception Net-V3 (45). The more realistic the generated images, the smaller the FID value.

Kernel inception distance (KID)

The square of the maximum mean discrepancy between the distributions of P_a and P_b was calculated to assess the quality of the fake b images (46) using the following formulas:

$$KID(a,b)=E_{a,a'\sim P_a}\left[k(a,a')\right]-2E_{a\sim P_a,b\sim P_b}\left[k(a,b)\right]+E_{b,b'\sim P_b}\left[k(b,b')\right] \quad [21]$$

$$k(a,b)=\left(\frac{1}{d}a^{T}b+1\right)^{3} \quad [22]$$

where k is the feature kernel, d =2,048 is the dimension of the feature vector, and a' and b' are the new samples obtained after adding polynomial features. The smaller the KID, the closer the two distributions, and the better the quality of the images produced by the model.

Learned perceptual image patch similarity (LPIPS)

The distance between the real a and fake b images were computed at the perceptual level. The lower the value, the higher the degree of similarity between the images. The feature extraction process was performed by the trained network, and the distance of the feature maps between the a and b images in the layer l was output. Next, the vector W was used for scaling, and the L2 distance was calculated at the channel level (47). The distance function is expressed as follows:

$$d(a,b)=\sum_{l}\frac{1}{H_l W_l}\sum_{h,w}\left\|w_l\odot\left(\hat{a}^{l}_{hw}-\hat{b}^{l}_{hw}\right)\right\|_2^2,\quad \hat{a}^{l},\hat{b}^{l}\in R^{H_l\times W_l\times C_l},\quad w^{l}\in R^{C_l} \quad [23]$$

Baselines

The current unsupervised FFA image synthesis approaches are mainly based on the strength of the CycleGAN. Therefore, we used the CycleGAN as our comparison model. In addition, several popular unsupervised image synthesis approaches related to the DCLAGAN were selected as the contrast models, including the U-GAT-IT, L-U-GAT-IT, GcGAN, CUT, and DCLGAN.

Results

There were three main stages to the experiments for which the CF-FFA data set was used. First, we analyzed the effectiveness of the CAM-A in the generator and the CAM in the discriminator (*Figure 5* and *Table 1*). Second, we compared the effectiveness of different attention mechanisms using the proposed model without CA blocks as a baseline (*Figure 6* and *Table 2*). Third, the DCLAGAN was compared with other popular image synthesis approaches using the same training and testing samples (*Figures 7*,*8* and *Table 3*). We also used the OCT-OCTA, Horse-Zebra and Cat-Dog data sets to verify the generalization and robustness of the DCLAGAN (*Figures 7*,*9*,*10*). In addition, to facilitate the qualitative analysis of the synthesized retinal images, the synthesized FFA and OCTA images were classified into the following two categories: normal and abnormal images (*Figures 8*,*9*).

Discussion

We chose CAM-A as an attention-guiding model for the generators to help the model embed stylistic features into content features more accurately, and we used CAM in the discriminators to improve the discriminative capacity of model. In the absence of the CA block, we verified the effectiveness of the CAM-A in the generators and the CAM in discriminators using the metrics of the FID, KID, and LPIPS. As detailed in *Table 1*, when we used the CAM-A, the mean values of the FID, KID, and LPIPS were decreased by 2.278, 0.00053, and 0.011, respectively. When we used the CAM in the discriminators, the FID, KID, and LPIPS were reduced by 2.173, 0.00358, and 0.012, respectively. Further, if we added the CAM-A and the CAM at the same time, the three metrics were reduced by 4.981, 0.00565, and 0.009, respectively.

The effectiveness of CAM-A and CAM can also be seen in the comparisons displayed in *Figure 5*. Notably, the magnified images in *Figure 5*, F2 are more similar to the real FFA images in *Figure 5*, B2 than the other images, and the location of the vessels is clearer than in *Figure 5*, C2,D2,E2. Additionally, the pathological characteristics in the first row of *Figure 5*, F2 are more obvious than those in the first rows of *Figure 5*, C2,D2,E2. The CAM fine-tunes the model by helping it to determine what to focus on or ignore; however, the location of vessels in *Figure 5*, E2 is not as accurate as that in *Figure 5*, F2 when compared with *Figure 5*, A2. From the attention maps generated by the generator in *Figure 5*, D3,F3, the darker the color of the attention maps, the more attention the items receive. We can see that while the CAM-A can guide the model to focus on the area of the retina and ignore the area without the retina, the location of the retinal edge vision is not distinguished more clearly than the model in which the CAM-A and CAM are used together.

To verify the effectiveness of the CA block, some attention mechanisms, including the channel attention mechanisms (the self-attention and SE blocks), and the channel combined with the spatial attention mechanisms (the scSE block and CBAM) were chosen for the comparisons. As detailed in *Table 2*, the mean scores of the

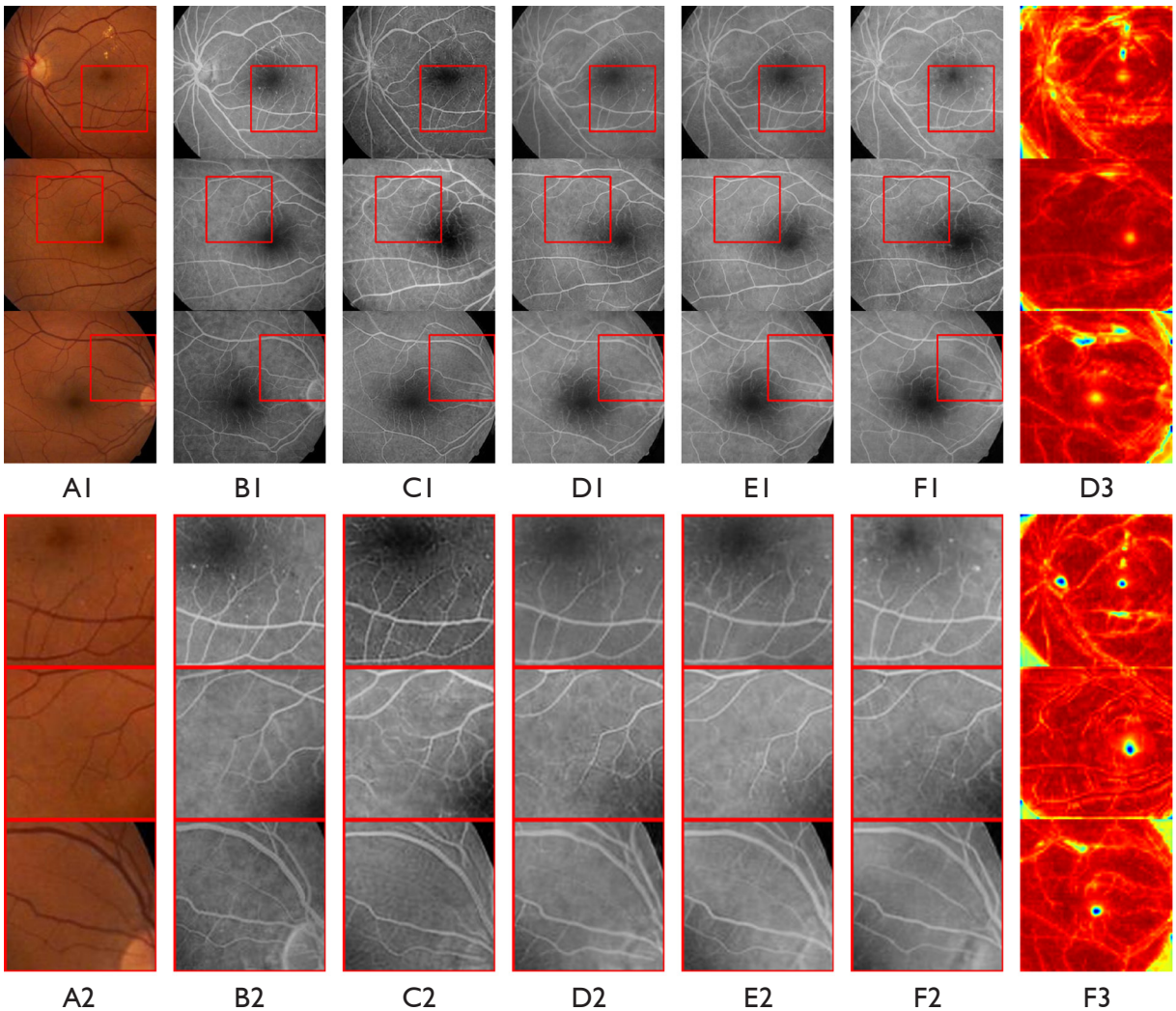

Figure 5 Ablation analysis of the CAM-A and CAM. (A1,A2) real CF images. (B1,B2) Real FFA images. (C1,C2) Synthesized images of DCLGAN. (D1,D2) Synthesized images of DCLGAN with CAM-A. (E1,E2) Synthesized images of DCLGAN with CAM. (F1,F2) Synthesized images of DCLGAN with CAM-A and CAM. (D3) Attention maps generated by the generator using CAM-A. (F3) Attention maps generated by the generator using CAM-A and CAM. The prominent pathological features or key vascular locations in each image are magnified with red frames. CAM, class activation mapping; CAM-A, class activation mapping with adaptive layer-instance normalization function; CF, color fundus; FFA, fundus fluorescein angiography; DCLGAN, dual contrastive learning generative adversarial network.

Table 1 The effects of the CAM-A and CAM

Methods	FID↓	KID×100↓	LPIPS↓
DCLGAN	58.914±1.854	2.453±0.169	0.256±0.013
DCLGAN + CAM-A	56.636±0.915	2.400±0.216	0.245±0.004
DCLGAN + CAM	56.741±0.032	2.095±0.308	0.244±0.008
DCLGAN + CAM-A + CAM	53.933±0.338	1.888±0.200	0.247±0.006

The values are presented as the mean ± standard deviation. ↓ means smaller numbers are better. CAM-A, class activation mapping with adaptive layer-instance normalization function; CAM, class activation mapping; FID, Fréchet inception distance; KID, kernel inception distance; LPIPS, learned perceptual image patch similarity; DCLGAN, dual contrastive learning generative adversarial network.

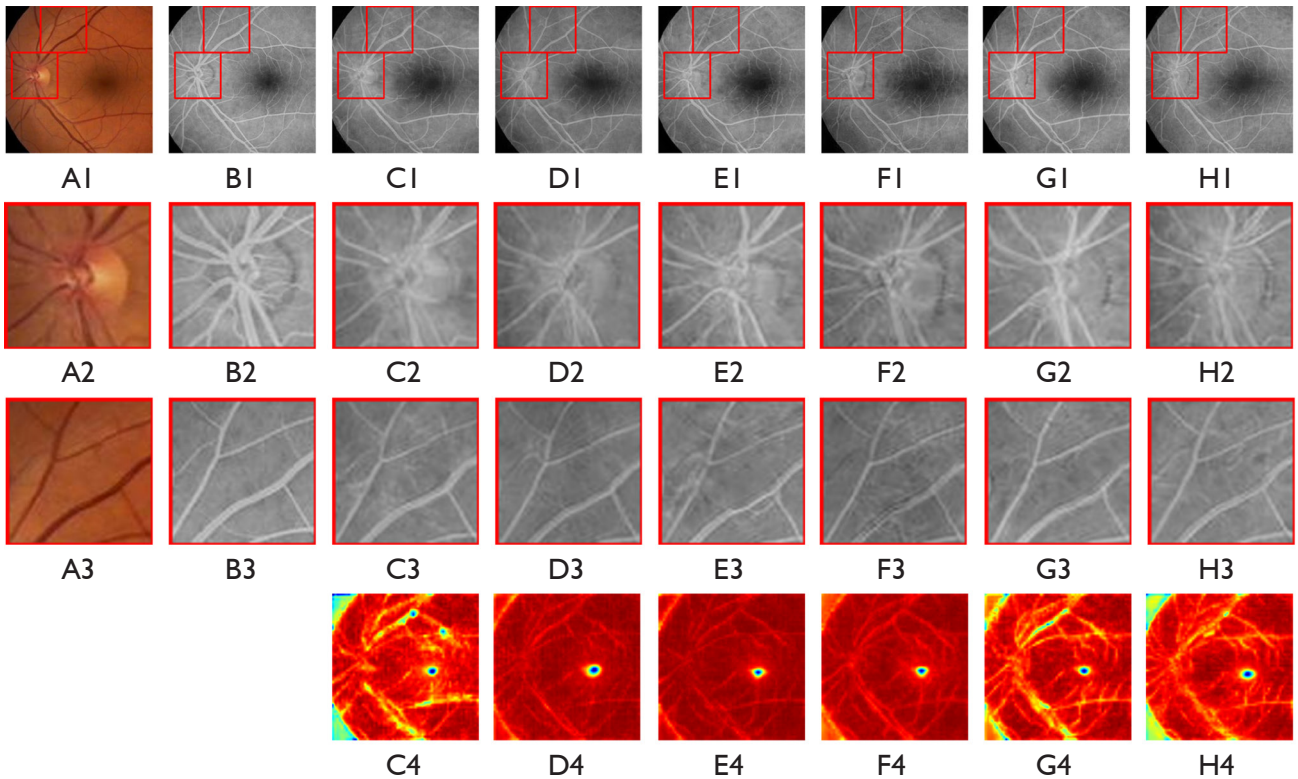

Figure 6 Visualization of the synthesized images and attention maps with different attention mechanisms. Baseline, DCLAGAN without CA block. (A1-A3) Real CF images. (B1-B3) Real FFA images. (C1-C4) Synthesized image and attention map generated by baseline. (D1-D4) Synthesized image and attention map generated by baseline with self-attention. (E1-E4) Synthesized image and attention map generated by baseline with SE block. (F1-F4) Synthesized image and attention map generated by baseline with scSE block. (G1-G4) Synthesized image and attention map generated by baseline with CBAM. (H1-H4) Synthesized image and attention map generated by baseline with CA block (DCLAGAN). The local vessels in each image are enlarged with red frames. DCLAGAN, dual contrastive learning attention generative adversarial network; CA, coordinate attention; CF, color fundus; FFA, fundus fluorescein angiography; SE, squeeze-and-excitation; scSE, spatial and channel squeeze-and-excitation; CBAM, convolutional block attention module.

Table 2 Comparisons of different attention mechanisms using the DCLAGAN without the CA block as the baseline

Methods	FID↓	KID×100↓	LPIPS↓
Baseline	53.933±0.338	1.888±0.200	0.247±0.006
Baseline + SA	59.822±2.983	2.494±0.523	0.251±0.004
Baseline + SE	52.235±0.527	1.771±0.121	0.243±0.010
Baseline + scSE	54.803±1.969	1.955±0.223	0.245±0.006
Baseline + CBAM	56.587±3.298	2.098±0.181	0.246±0.008
Baseline + CA	50.490±1.270	1.529±0.210	0.245±0.007

Values are presented as the mean ± standard deviation. ↓ means smaller numbers are better. Baseline, the DCLAGAN (our proposed model) without the CA block; CA, coordinate attention block; FID, Fréchet inception distance; KID, kernel inception distance; LPIPS, learned perceptual image patch similarity; SA, self-attention; SE, squeeze-and-excitation block; ScSE, spatial and channel squeeze-and-excitation block; CBAM, convolutional block attention module; DCLAGAN, dual contrastive learning attention generative adversarial network.

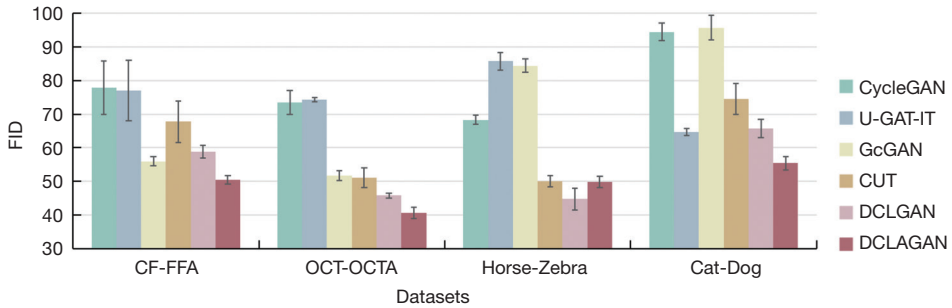

Figure 7 Comparisons of the DCLAGAN with other popular unsupervised image synthesis methods using histograms, the metric of the FID, and four data sets (CF-FFA, OCT-OCTA, Horse-Zebra, and Cat-Dog). FID, Fréchet inception distance; CF, color fundus; FFA, fundus fluorescein angiography; OCT, optical coherence tomography; OCTA, optical coherence tomography angiography; CycleGAN, cycle-consistent generative adversarial network; U-GAT-IT, unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation; GcGAN, geometry-consistent generative adversarial networks; CUT, contrastive learning for unpaired image-to-image translation; DCLGAN, dual contrastive learning generative adversarial network; DCLAGAN, dual contrastive learning attention generative adversarial network.

FID and KID improved greatly when we the CA block was chosen, but worsened when the self-attention block, scSE block, and CBAM were added. However, the mean score of the LPIPS obtained by the addition of the CA block was slightly worse than that obtained by the addition of the SE block.

As *Figure 6* shows, the vascular structure of *Figure 6*, H2,H3 was clearer than the other attention mechanisms with the baseline model, and the vascular structures of the baseline model with the CA block were closer to the real CF image than the images of others. Thus, the image synthesis model with the CA block improved the reliability of the fake images by providing more accurate information about the location of the vascular structures. From the attention maps generated by the generator in row 4, we can see that the location of the retinal edge in *Figure 6*, H4 was clearer than that in other attention maps generated by other attention mechanisms. *Figure 6*, C4 distinguishes more clearly between the retinal and extraretinal areas; however, the location is not as smooth and clear as the attention map in *Figure 6*, H4.

We performed quantitative evaluations for the proposed model DCLAGAN against several widely used unsupervised image synthesis models, including the CycleGAN, U-GAT-IT, L-U-GAT-IT, GcGAN, CUT, and DCLGAN. To make fair comparisons, we used the same training samples [1,680] and testing samples [272] for the evaluations. The quantitative image synthesis results are set out in *Table 3*, and our method gained 8.424 improvements in the FID, 0.00924 improvements in the KID, and 0.011 improvements in LPIPS compared with the baseline method of DCLGAN. As is well known, the current unsupervised FFA image synthesis methods are mainly based on the CycleGAN. Compared with the CycleGAN, our method gained 27.319 improvements in the FID, 0.04002 improvements in the KID, and 0.07 improvements in LPIPS.

Further, as *Figure* 7 vividly and intuitively shows, our DCLAGAN achieved the best FID compared to any of the comparison baselines using the CF-FFA data set, while the GcGAN achieved the next best FID. However, as *Figure 8F* shows, the FFA images generated by the GcGAN were not realistic and some information about the vessels was lost.

Additionally, as *Figure 8I* shows, the visual effects of our DCLAGAN were also better than those of others. Despite the fact that the L-U-GAT-IT required less GPU and training time than the U-GAT-IT, the basic shape of the generated images in *Figure 8E* was not as accurate as the basic shape in the images in *Figure 8D* when compared with *Figure 8A*. The GcGAN tries to ensure the basic shape of the source images are kept as similar as possible; however, the style characteristics of the target images cannot be preserved in the generated images any better than they can be by the DCLAGAN.

To verify the generalization and robustness of the DCLAGAN, we selected three other data sets; that is, the OCT-OCTA, Horse-Zebra and Cat-Dog data sets. Among them, the OCT and OCTA images in the OCT-OCTA data set have been popular retinal imaging modalities in recent years. The Horse-Zebra and Cat-Dog data sets are commonly used for image synthesis tasks. In the first

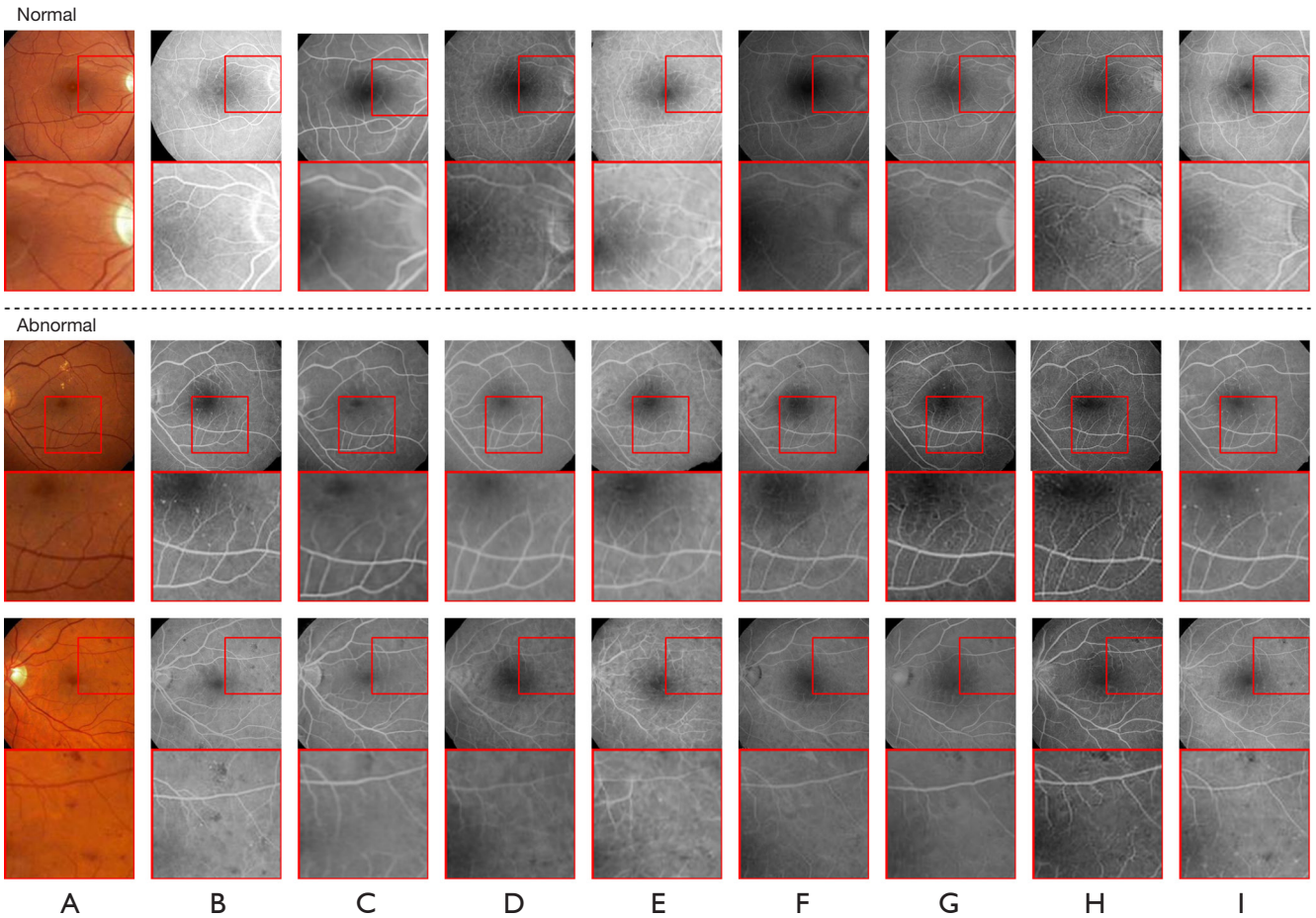

Figure 8 Comparisons of the DCLAGAN with other advanced methods using the CF-FFA data set: (A) real CF images; (B) real FFA images; (C) synthesized images generated by CycleGAN; (D) synthesized images generated by the U-GAT-IT; (E) synthesized images generated by the lightweight U-GAT-IT; (F) synthesized images generated by the GcGAN; (G) synthesized images generated by the CUT; (H) synthesized images generated by the DCLGAN; (I) synthesized images generated by the DCLAGAN. The location of the optic disc in each normal image is magnified with a red frame. The location of the prominent pathological features in each abnormal image is magnified with a red frame. DCLAGAN, dual contrastive learning attention generative adversarial network; CF, color fundus; FFA, fundus fluorescein angiography; CycleGAN, cycle-consistent generative adversarial network; U-GAT-IT, unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation; GcGAN, geometry-consistent generative adversarial networks; CUT, contrastive learning for unpaired image-to-image translation; DCLGAN, dual contrastive learning generative adversarial network.

row of positions (*Figure 9*), the synthesized OCTA images generated by the CycleGAN (*Figure 9C*), DCLGAN (*Figure 9G*), and DCLAGAN (*Figure 9H*) retained most of the blood vessel information. Comparing the CycleGAN synthesized OCTA images (*Figure 9C*) with the real OCT (*Figure 9A*) and real OCTA images (*Figure 9B*), we can clearly see that in the CycleGAN synthesized OCTA images, the accuracy of the vessel location was not as good as the results obtained using the DCLGAN and DCLAGAN methods. In the last row of positions, it is clear that there was more redundant information in the synthesized OCTA images generated by the U-GAT-IT, CUT, and DCLGAN. *Figure 7* shows the average results of the FID by using the OCT-OCTA data set; our proposed DCLAGAN had the best average score for the FID.

As *Figure 10* shows, the zebra images generated by the CUT, DCLGAN, and DCLAGAN were closer to the zebra in reality. The average results of the FID by using the Horse-Zebra data set can also be used as an aid to validate the above observation (*Figure 7*). However, the visualization

Table 3 The quantitative comparisons of the DCLAGAN against the mentioned baselines

Methods	FID↓	KID×100↓	LPIPS↓
CycleGAN (25)	77.809±7.973	5.531±1.230	0.315±0.006
U-GAT-IT (27)	76.986±8.937	4.333±0.576	0.302±0.017
L-U-GAT-IT (27)	99.787±4.801	7.432±0.438	0.328±0.013
GcGAN (32)	55.993±1.373	1.639±0.055	0.266±0.004
CUT (33)	67.793±6.140	3.112±0.531	0.291±0.003
DCLGAN (34)	58.914±1.854	2.453±0.169	0.256±0.013
DCLAGAN	50.490±1.270	1.529±0.210	0.245±0.007

Values are presented as the mean ± standard deviation. ↓ means smaller numbers are better. DCLAGAN, dual contrastive learning attention generative adversarial network; FID, Fréchet inception distance; KID, kernel inception distance; LPIPS, learned perceptual image patch similarity; CycleGAN, cycle-consistent generative adversarial network; U-GAT-IT, unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation; L-U-GAT-IT, lightweight U-GAT-IT; GcGAN, geometry-consistent generative adversarial network; CUT, contrastive learning for unpaired image-to-image translation; DCLGAN, dual contrastive learning generative adversarial network.

of the synthesized zebra images and the FID results all indicate that the quality of the synthesized zebra images generated by the DCLAGAN was slightly lower than that of the DCLGAN.

In relation to the synthesized dog images in *Figure 10*, it is clear that the images generated by the U-GAT-IT, DCLGAN, and DCLAGAN were more realistic than those generated by the other methods. The synthesized dog images generated by the CycleGAN and GcGAN retained some of the cat characteristics. The average results of the FID by using the Cat-Dog data set show that the quality of the synthesized dog images generated by the U-GAT-IT, DCLGAN, and DCLAGAN was better than that generated by the other method (*Figure* 7). In addition, the average FID score of the DCLAGAN using the Cat-Dog data set was the lowest among all the approaches.

In relation to the category of abnormal images shown in *Figure 8*, it is clear that the synthetic FFA images generated by the DCLAGAN (*Figure 8I*) retained more pathological features than the images generated by other methods when compared to the real FFA images (*Figure 8B*), but there was still a small proportion of pathological features that were not shown. The pathological features were barely visible in the abnormal FFA images generated by the U-GAT-IT (*Figure 8D*), L-U-GAT-IT (*Figure 8E*), and GcGAN (*Figure 8F*). The abnormal FFA images generated by the CycleGAN (*Figure 8C*) preserved some pathological features. However, the vessel location was not accurate. Further, the synthesized FFA images generated by the CUT and DCLGAN, while preserving some pathological features, also had some redundant vessels whose locations were inaccurate.

In relation to the category of abnormal images shown in *Figure 9*, it is obvious that the OCTA images generated by the DCLAGAN also ensured that the vessel locations were largely accurate and the vast majority of pathological features were retained, but a small number of vessels and pathological features were still blurred. The OCTA images generated by the CUT and DCLGAN had some obvious pathological features; however, they also had some more redundant vessels. The OCTA images generated by the CycleGAN, U-GAT-IT, and GcGAN showed clear blood vessels; however, the pathological features were largely lost.

Conclusions

We developed an unsupervised image synthesis framework DCLAGAN based on dual contrastive learning that can synthesize FFA images from unpaired CF images. Further, we also revised the structure of the generators to increase the effectiveness of the dual contrastive learning. Our extensive experimental results demonstrate the superiority of the DCLAGAN for unsupervised FFA image synthesis over the other methods mentioned, both in terms of our quantitative evaluation results and the visual effects. We believe that unsupervised FFA image synthesis will be widely used for the diagnosis of retinal diseases in the near future. The FFA images synthesized by the DCLAGAN

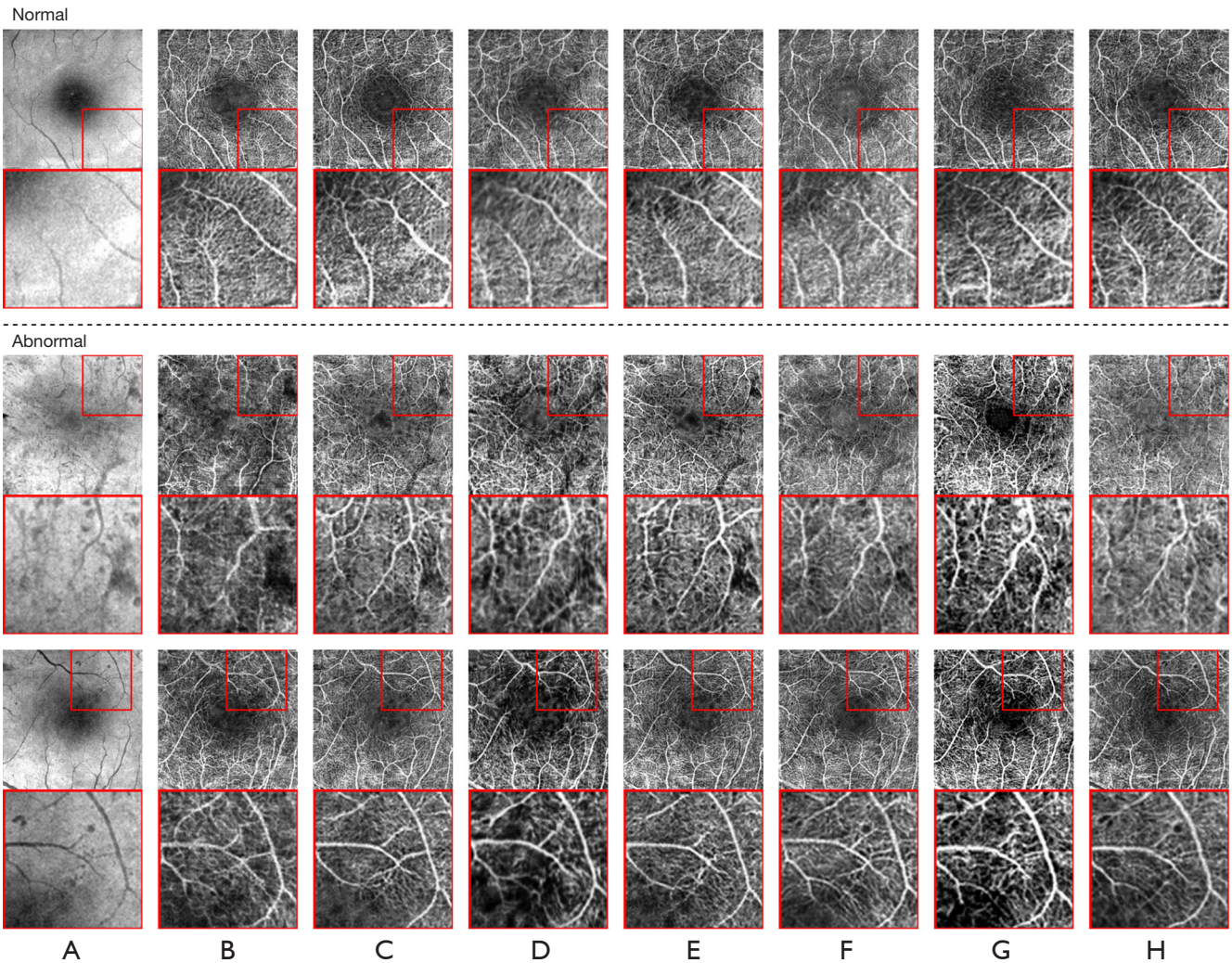

Figure 9 Comparisons of the DCLAGAN with other advanced methods using the OCT-OCTA data set: (A) real OCT images; (B) real OCTA images; (C) synthesized images generated by CycleGAN; (D) synthesized images generated by the U-GAT-IT; (E) synthesized images generated by the GcGAN; (F) synthesized images generated by the CUT; (G) synthesized images generated by the DCLGAN; (H) synthesized images generated by the DCLAGAN. The local vessel in each normal image is enlarged with red frame. The location of prominent pathological features in each abnormal image is magnified with a red frame. DCLAGAN, dual contrastive learning attention generative adversarial network; OCT, optical coherence tomography; OCTA, optical coherence tomography angiography; CycleGAN, cycle-consistent generative adversarial network; U-GAT-IT, unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation; GcGAN, geometry-consistent generative adversarial networks; CUT, contrastive learning for unpaired image-to-image translation; DCLGAN, dual contrastive learning generative adversarial network.

retained most of the pathological features; however, a small amount of redundant information still inevitably appeared in the images. In addition, the DCLAGAN may work better if data sets are used in which the positions of the content to be transformed in the source and target images are close to each other. In the future, we will seek to improve the quality of FFA images generated by image synthesis by attempting to incorporate a diffusion model with the DCLAGAN, and

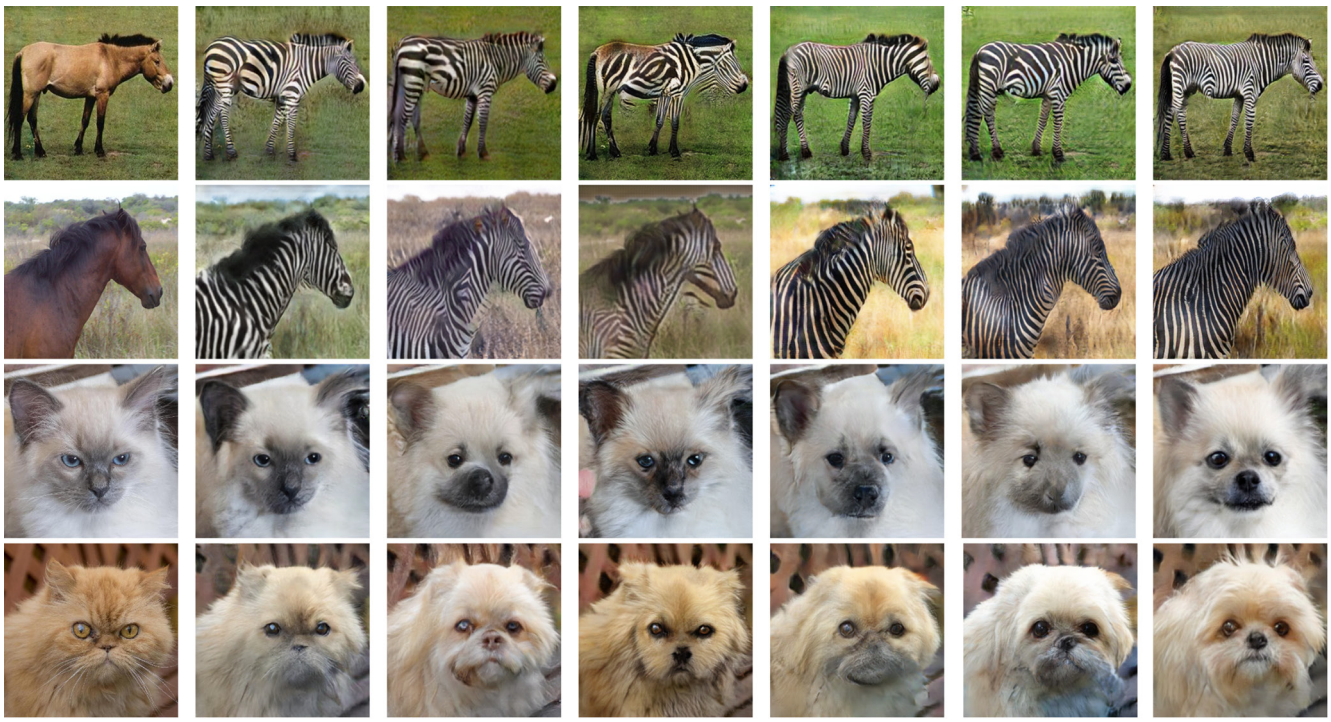

Figure 10 Comparisons of the DCLAGAN with other advanced methods using the Horse-Zebra and Cat-Dog data sets: (A) real horse and cat images; (B) synthesized images generated by CycleGAN; (C) synthesized images generated by the U-GAT-IT; (D) synthesized images generated by the GcGAN; (E) synthesized images generated by the CUT; (F) synthesized images generated by the DCLGAN; (G) synthesized images generated by the DCLAGAN. DCLAGAN, dual contrastive learning attention generative adversarial network; CycleGAN, cycle-consistent generative adversarial network; U-GAT-IT, unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation; GcGAN, geometry-consistent generative adversarial networks; CUT, contrastive learning for unpaired image-to-image translation; DCLGAN, dual contrastive learning generative adversarial network.

to ameliorate the unsupervised approaches of multimodal retinal registration based on the DCLAGAN. We will also improve CAM-A and CA to improve the accuracy of the content and location of the generated pathological features.

Acknowledgments

Funding: This work was supported by the Science and Technology Research Planning Project of the Education Department of Jilin Province: Research on Automatic Registration of Multimodal Retinal Images (No. JJKH20230843KJ).

Footnote

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://qims.amegroups.com/article/view/10.21037/qims-23-1304/coif). The authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013).

References

1. Yannuzzi LA, Rohrer KT, Tindel LJ, Sobel RS, Costanza MA, Shields W, Zang E. Fluorescein angiography complication survey. Ophthalmology 1986;93:611-7.
2. Fercher AF. Optical coherence tomography. J Biomed Opt 1996;1:157-73.
3. Chalam KV, Sambhav K. Optical Coherence Tomography Angiography in Retinal Diseases. J Ophthalmic Vis Res 2016;11:84-92.
4. Abràmoff MD, Garvin MK, Sonka M. Retinal imaging and image analysis. IEEE Rev Biomed Eng 2010;3:169-208.
5. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial nets. Advances in neural information processing systems. 2014;27.
6. Lin T, Ma Z, Li F, He D, Li X, Ding E, Wang N, Li J, Gao X. Drafting and revision: Laplacian pyramid network for fast high-quality artistic style transfer. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 20-25 June 2021; Nashville, TN, USA. IEEE; 2021:5141-50.
7. Fuentes-Hurtado F, Sibarita JB, Viasnoff V. Generalizable Denoising of Microscopy Images using Generative Adversarial Networks and Contrastive Learning. arXiv:2303.15214 [Preprint]. 2023. Available online: https://doi.org/10.48550/arXiv.2303.15214
8. Chan KCK, Xu X, Wang X, Gu J, Loy CC. GLEAN: Generative Latent Bank for Image Super-Resolution and Beyond. IEEE Trans Pattern Anal Mach Intell 2023;45:3154-68.
9. Liu Y, Chen A, Shi H, Huang S, Zheng W, Liu Z, Zhang Q, Yang X. CT synthesis from MRI using multi-cycle GAN for head-and-neck radiation therapy. Comput Med Imaging Graph 2021;91:101953.
10. Zhao B, Cheng T, Zhang X, Wang J, Zhu H, Zhao R, Li D, Zhang Z, Yu G. CT synthesis from MR in the pelvic area using Residual Transformer Conditional GAN. Comput Med Imaging Graph 2023;103:102150.
11. Zhao M, Liu X, Liu H, Wong KKL. Super-resolution of cardiac magnetic resonance images using Laplacian Pyramid based on Generative Adversarial Networks. Comput Med Imaging Graph 2020;80:101698.
12. Upadhyay U, Chen Y, Hepp T, Gatidis S, Akata Z. Uncertainty-guided progressive GANs for medical image translation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24. Springer International Publishing; 2021:614-24.
13. Hervella ÁS, Rouco J, Novo J, Ortega M. Retinal image understanding emerges from self-supervised multimodal reconstruction. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I. Springer International Publishing; 2018:321-8.
14. Hervella ÁS, Rouco J, Novo J, Ortega M. Self-supervised multimodal reconstruction of retinal images over paired data sets. Expert Systems with Applications 2020;161:113674.
15. 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. Springer International Publishing; 2015:234-41.
16. Mirza M, Osindero S. Conditional generative adversarial nets. arXiv:1411.1784 [Preprint]. 2014. Available online: https://doi.org/10.48550/arXiv.1411.1784
17. Kamran SA, Fariha Hossain K, Tavakkoli A, Zuckerbrod S, Baker SA, Sanders KM. Fundus2Angio: a conditional GAN architecture for generating fluorescein angiography images from retinal fundus photography. In: Advances in Visual Computing: 15th International Symposium, ISVC 2020, San Diego, CA, USA, October 5–7, 2020, Proceedings, Part II 15. Springer International Publishing; 2020:125-38.
18. Li W, Kong W, Chen Y, Wang J, He Y, Shi G, Deng G. Generating fundus fluorescence angiography images from structure fundus images using generative adversarial networks. arXiv:2006.10216 [Preprint]. 2020. Available online: https://doi.org/10.48550/arXiv.2006.10216
19. Tavakkoli A, Kamran SA, Hossain KF, Zuckerbrod SL. A novel deep learning conditional generative adversarial network for producing angiography images from retinal fundus photographs. Sci Rep 2020;10:21580.
20. Kamran SA, Hossain KF, Tavakkoli A, Zuckerbrod SL. Attention2angiogan: Synthesizing fluorescein angiography from retinal fundus images using generative adversarial networks. 2020 25th International Conference on Pattern

Recognition (ICPR); 10-15 January 2021; Milan, Italy. IEEE; 2021:9122-9.
21. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv:2010.11929 [Preprint]. 2020. Available online: https://doi.org/10.48550/arXiv.2010.11929
22. Kamran SA, Hossain KF, Tavakkoli A, Zuckerbrod SL, Baker SA. VTGAN: Semi-supervised retinal image synthesis and disease prediction using vision transformers. 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW); 11-17 October 2021; Montreal, BC, Canada. IEEE; 2021:3235-45.
23. Schiffers F, Yu Z, Arguin S, Maier A, Ren Q. Synthetic fundus fluorescein angiography using deep neural networks. In: Bildverarbeitung für die Medizin 2018: Algorithmen-Systeme-Anwendungen. Proceedings des Workshops vom 11. bis 13. März 2018 in Erlangen. Springer Berlin Heidelberg; 2018:234-8.
24. Hervella ÁS, Rouco J, Novo J, Ortega M. Deep multimodal reconstruction of retinal images using paired or unpaired data. In2019 International Joint Conference on Neural Networks (IJCNN); 14-19 July 2019; Budapest, Hungary. IEEE; 2019:1-8.
25. Zhu JY, Park T, Isola P, Efros AA. Unpaired image-to-image translation using cycle-consistent adversarial networks. 2017 IEEE International Conference on Computer Vision (ICCV); 22-29 October 2017; Venice, Italy. IEEE; 2017:2223-32.
26. Cai Z, Xin J, Wu J, Liu S, Zuo W, Zheng N. Triple Multi-scale Adversarial Learning with Self-attention and Quality Loss for Unpaired Fundus Fluorescein Angiography Synthesis. Annu Int Conf IEEE Eng Med Biol Soc 2020;2020:1592-5.
27. Kim J, Kim M, Kang H, Lee K. U-gat-it: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation. arXiv:1907.10830 [Preprint]. 2019. Available online: https://doi.org/10.48550/arXiv.1907.10830
28. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 27-30 June 2016; Las Vegas, NV, USA. IEEE; 2016:2921-9.
29. Huang X, Belongie S. Arbitrary style transfer in real-time with adaptive instance normalization. 2017 IEEE International Conference on Computer Vision (ICCV); 22-29 October 2017; Venice, Italy. IEEE; 2017:1501-10.
30. Ulyanov D, Vedaldi A, Lempitsky V. Instance normalization: The missing ingredient for fast stylization. arXiv:1607.08022 [Preprint]. 2016. Available online: https://doi.org/10.48550/arXiv.1607.08022
31. Ba JL, Kiros JR, Hinton GE. Layer normalization. arXiv:1607.06450 [Preprint]. 2016. Available online: https://doi.org/10.48550/arXiv.1607.06450
32. Fu H, Gong M, Wang C, Batmanghelich K, Zhang K, Tao D. Geometry-Consistent Generative Adversarial Networks for One-Sided Unsupervised Domain Mapping. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2019;2019:2422-31.
33. Park T, Efros AA, Zhang R, Zhu JY. Contrastive learning for unpaired image-to-image translation. In: Vedaldi A, Bischof H, Brox T, Frahm JM. editors. Computer Vision–ECCV 2020. Springer International Publishing; 2020:319-345.
34. Han J, Shoeiby M, Petersson L, Armin MA. Dual contrastive learning for unsupervised image-to-image translation. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); 19-25 June 2021; Nashville, TN, USA. IEEE; 2021:746-55.
35. Zhang H, Goodfellow I, Metaxas D, Odena A. Self-attention generative adversarial networks. Proceedings of the 36th International Conference on Machine Learning, PMLR 2019;97:7354-63.
36. Hu J, Shen L, Sun G. Squeeze-and-excitation networks. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 18-23 June 2018; Salt Lake City, UT, USA. IEEE; 2018:7132-41.
37. Roy AG, Navab N, Wachinger C. Concurrent spatial and channel 'squeeze & excitation'in fully convolutional networks. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I. Springer International Publishing; 2018:421-9.
38. Woo S, Park J, Lee JY, Kweon IS. CBAM: Convolutional block attention module. In: Ferrari V, Hebert M, Sminchisescu C, Weiss Y. editors. Computer Vision – ECCV 2018. Springer, Cham; 2018:3-19.
39. Hou Q, Zhou D, Feng J. Coordinate attention for efficient mobile network design. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 20-25 June 2021; Nashville, TN, USA. 2021:13713-22.
40. Oord AV, Li Y, Vinyals O. Representation learning with contrastive predictive coding. arXiv:1807.03748 [Preprint].

2018. Available online: https://doi.org/10.48550/arXiv.1807.03748
41. Hajeb Mohammad Alipour S, Rabbani H, Akhlaghi MR. Diabetic retinopathy grading by digital curvelet transform. Comput Math Methods Med 2012;2012:761901.
42. Li M, Huang K, Xu Q, Yang J, Zhang Y, Ji Z, Xie K, Yuan S, Liu Q, Chen Q. OCTA-500: A Retinal Data set for Optical Coherence Tomography Angiography Study. arXiv:2012.07261 [Preprint]. 2020. Available online: https://doi.org/10.48550/arXiv.2012.07261
43. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. ImageNet: A large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition; 20-25 June 2009; Miami, FL, USA. IEEE; 2009:248-55.
44. Choi Y, Uh Y, Yoo J, Ha JW. StarGAN v2: Diverse image synthesis for multiple domains. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 13-19 June 2020; Seattle, WA, USA. IEEE; 2020:8188-97.
45. Heusel M, Ramsauer H, Unterthiner T, Nessler B, Hochreiter S. GANs trained by a two time-scale update rule converge to a local nash equilibrium. 31st International Conference on Neural Information Processing Systems31st Conference on Neural Information Processing Systems (NIPS 2017); Long Beach, CA, USA. 2017:6629-40.
46. Fatras K, Zine Y, Majewski S, Flamary R, Gribonval R, Courty N. Minibatch optimal transport distances; analysis and applications. arXiv:2101.01792 [Preprint]. 2021. Available online: https://doi.org/10.48550/arXiv.2101.01792
47. Zhang R, Isola P, Efros AA, Shechtman E, Wang O. The unreasonable effectiveness of deep features as a perceptual metric. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 18-23 June 2018; Salt Lake City, UT, USA. IEEE; 2018:586-95.

Cite this article as: Zhao J, Huang H, Wang C, Yu M, Shi W, Mori K, Jiang Z, Liu J. Dual contrastive learning for synthesizing unpaired fundus fluorescein angiography from retinal fundus images. Quant Imaging Med Surg 2024;14(3):2193-2212. doi: 10.21037/qims-23-1304

