

VSG-GAN: A High Fidelity Image Synthesis Method with Semantic Manipulation in Retinal Fundus Image

Junjie Liu^{a,b,c,d,e}, Shixin Xu^e, Ping He^{a,b}, Sirong Wu^{a,b,c}, Xi Luo^{a,b,c}, Yuhui Deng^{a,b,*}, and Huaxiong Huang^{f,g,a,b,*}

^aGuangdong Provincial Key Laboratory of Interdisciplinary Research and Application for Data Science, Zhuhai 519087, China

^bBNU-HKBU United International College, Zhuhai 519087, China

^cFaculty of Science, Hong Kong Baptist University, Hong Kong SAR 999077, China

^dTrinity College Dublin, Dublin 2, Ireland

^eData Science Research Center, Duke Kunshan University, 8 Duke Ave, Kunshan, Jiangsu, China

^fResearch Center for Mathematics, Beijing Normal University at Zhuhai, 519087, China

^gDepartment of Mathematics and Statistics, York University, Toronto, Ontario, Canada

*Correspondence: ivandeng@uic.edu.cn, hhuang@uic.edu.cn

Abstract

In recent years, advancements in retinal image analysis, driven by machine learning and deep learning techniques, have enhanced disease detection and diagnosis through automated feature extraction. However, challenges persist, including limited dataset diversity due to privacy concerns and imbalanced sample pairs, hindering effective model training. To address these issues, we introduce the Vessel & Style Guided Generative Adversarial Network (VSG-GAN), an innovative algorithm building upon the foundational concept of GAN. In VSG-GAN, a generator and discriminator engage in an adversarial process to produce realistic retinal images. Our approach decouples retinal image generation into distinct modules: the vascular skeleton and background style. Leveraging style transformation and GAN inversion, our proposed Hierarchical Variational Autoencoder (HVAE) module generates retinal images with diverse morphological traits. Additionally, the Spatially-Adaptive De-normalization (SPADE) module ensures consistency between input and generated images. We evaluate our model on MESSIDOR and RITE datasets using various metrics, including Structural Similarity Index Measure (SSIM), Inception Score (IS), Fréchet Inception Distance (FID), and Kernel Inception Distance (KID). Our results demonstrate the superiority of VSG-GAN, outperforming existing methods across all evaluation assessments. This underscores its effectiveness in addressing dataset limitations and imbalances. Our algorithm provides a novel solution to challenges in retinal image analysis by offering diverse and realistic retinal image generation. Implementing the VSG-GAN augmentation approach on downstream diabetic retinopathy classification tasks has shown enhanced disease diagnosis accuracy, further advancing the utility of machine learning in this domain.

Statement of significance The proposed research significantly advances retinal image analysis by addressing critical challenges in machine learning-based diagnostics. By leveraging Conditional Generative Adversarial Networks (Conditional GANs), this study pioneers an innovative approach to generating diverse and realistic retinal images. This breakthrough overcomes limitations posed by restricted medical datasets due to privacy concerns, offering a solution to the scarcity of diverse samples and imbalanced data pairs. The generated images exhibit remarkable accuracy in simulating retinal morphology and colour style. Such advancements hold profound implications for enhanced disease diagnosis and treatment planning in ophthalmology. This research not only showcases superior performance over existing methods but also sets a new standard in retinal image generation, promising substantial contributions to improving healthcare practices and patient outcomes.

1 Introduction

The intricate structure of retinal blood vessels is key in diagnosing a range of eye, heart, and brain conditions, including glaucoma [6, 17, 43], macular telangiectasia type 2 [35], diabetic retinopathy [47], and Alzheimer’s disease [22]. Precisely identifying characteristics of these vessels, such as size, shape, intersections, branching, and angles, is essential for accurate diagnosis [33, 34]. Computer Vision, a significant area in Machine Learning and Deep Learning, leverages computers and algorithms for detecting objects, classifying instances, and segmenting images, whether in 2D or 3D. Deep Learning is notable for its exceptional capability in extracting non-linear features and combining features across scales and modes. This data-centric approach enables Deep Learning models to effectively learn and adapt to the patterns in data automatically. In medical imaging, Deep Learning models [1, 54] are extensively used for various tasks including segmentation and classification, across different imaging types (like X-rays, MRI, NMR, Ultrasound) and organs (brain, lung, liver, pancreas, prostate, colon, and others).

In both Machine Learning and Deep Learning, models are generally trained on set datasets and assessed on distinct test sets. However, limited or unbalanced training datasets can lead to overfitting or poor model generalization, a common issue in medical imaging due to often unavailable or small datasets. Data augmentation addresses this by generating new, varied examples from the same data distribution, thereby enlarging the dataset’s size and variety. This enhances the model’s ability to learn more diverse features and reduces overfitting. Image data augmentation includes transforming existing data and creating synthetic data. These methods involve manipulating images through geometric changes (like rotation, scaling, flipping) and pixel alterations (modifying brightness, contrast, saturation, hue, etc.), thereby generating varied versions of the original images. For example, an image can be transformed by vertical flipping, contrast adjustment, or Gaussian blurring to produce new, diverse examples.

However, these standard augmentation methods can usually generate plausible alternative data [2]. To achieve a more extensive range of augmentations, deep generative models can be employed. These generative models can produce artificial data not present in the original dataset as supplementary instances. Examples of such models include Generative Adversarial Network(GAN), Auto-Encoder(AE)-based methods, Variational Auto-Encoder(VAE) variant methods, and Feature-Mixing algorithms. The generated images may exhibit similarities to existing images but with variations, such as having the same vasculature architecture but different lesions or background colours. These variations can enrich the characteristic combination within the dataset. Additionally, the synthesis process can help us disentangle the image’s components, which can be helpful when investigating the correlation between diseases and image characteristics. Costa et al. [7, 8] suggested a synthesis approach relying on a conditional Generative Adversarial Network (GAN), employing the vessel tree as input for producing a synthetic fundus image. Additionally, they integrated an Adversarial Autoencoder (AAE) for synthesizing the vessel tree, explicitly enhancing the diversity in the resultant generation of fundus images. Burlina et al. [5] used ProGAN [27] architecture to generate 512×512 synthetic image for the discriminative tasks in ophthalmology. Guo et al. [16] introduced a retinal image generation model called RetiGAN, which is based on the image translation method Pix2Pix [23] and Pix2PixHD [49]. Diaz-Pinto et al. [10] joined VAE and DCGAN together for the fundus image generation, and they found that the latent space from the VAE can extract the information well for the generation. Mahapatra et al. [36] proposed a super-resolution method on retinal images by considering the local saliency loss between the generated high-resolution images and original low-resolution images. Kaplan [26] used the contrastive-learning technique with VAE to learn the latent representation from the data and then generate the coherence tomography (OCT) based on the latent space. Kim et al. [29] adopted the StyleGAN for fundus image synthesis and checked the validity of synthetic images in the downstream classification task.

The GAN-based structure is also widely recognized for its ability to encode and capture rich semantic information within the intermediate features or latent spaces of the network [55, 3, 11, 24]. We can manipulate the generated images to possess specific attributes by perturbing the input or latent code. Nevertheless, these manipulations are limited to GAN-generated images since the actual model lacks the inferential capacity for real-world images. To overcome this constraint, GAN Inversion has been introduced to map a given image into the latent space of a pre-trained GAN model. This mapping enables controllable manipulations as the input serves as guidance for generation. In other words, it is a process that, by considering the inversion of the generation, we can find the specific latent representation of the synthetic image. With the prevalence of StyleGAN [28], researchers intend to design an encoder for mapping the real-world input of the GAN’s latent space. Perarnau et al. [38] proposed a conditional GAN for the GAN inversion where they encode the real image and attribute vector into the GAN’s latent, and by adjusting the attribute vector, the

generator can provide a different synthesized image. Richardson et al. [40] introduced a Style Encoder called “Pixel2Style2Pixel”(pSp). The contribution of this model is introducing an encoder with pSp structure to extract the feature map of the input image and transfer the extracted feature map to style with “MAP2STYLE” module. Different from Richardson, Wei et al. [51], instead of training a complicated encoder for mapping, they proposed a shallow and efficient mapping with just average pooling and fully connected layers. Ren et al. [39] used GAN to generate medical images while training an encoder to align the real images’ latent codes to the synthetic ones. Golhar et al. [12] claimed that GAN inversion could be used for data augmentation as they invert image pairs with the same label into a semantically rich and disentangled latent space, enabling manipulation of latent representations to generate new synthetic images.

In our study, we thoroughly explore techniques for creating synthetic retinal images that mimic real ones in structure and appearance. We introduce VSG-GAN(Vessel & Style Guided - GAN), a new generative adversarial network based on the Conditional GAN framework. This network uses both the vasculature mask and the image’s latent representation as inputs, embedding these conditions at each generation step. The vessel tree provides content guidance, while the latent representation influences style and colour. VSG-GAN incorporates a UNet encoder for deep feature mapping and a Hierarchical VAE (HVAE) architecture as its latent encoder to capture various levels of image detail. The discriminator, based on the patchGAN model, focuses on local image patches to enhance detail recognition. An iterative optimization technique refines the network’s ability to segment arterial and venous structures, with alternating training of the generator and discriminator. This method enhances the realism and accuracy of generated vessel structures, aiding in overcoming class imbalance challenges in datasets, particularly for specific diseases. Our approach aims to improve data representation and boost the performance and generalization of our models through advanced image synthesis. We extensively train our model using publicly available fundus image datasets, specifically the DRIVE dataset by Staal et al. [46, 20] and the MESSIDOR [9] dataset by Decencière et al. (2014). Subsequently, we evaluate the trained model on a test set, providing a dependable benchmark for comparing our approach with current state-of-the-art techniques. Our evaluations show significant benefits for clinical diagnosis and screening.

The subsequent sections of the paper are structured as follows: Section 2 delves into our methodology, encompassing theories, associated technologies, and our innovative approach. In Section 3, we showcase the outcomes of our experiments, ablation studies and a potential downstream application, offering a comprehensive summary of our discoveries while outlining potential future directions in medical image synthesis. Our insights regarding the results and limitations of the proposed method are elaborated upon in Section 4.

2 Materials and methods

In medical image synthesis, ensuring the fidelity of generated images to actual anatomical structures and pathological conditions remains a critical challenge. Moreover, these synthetic images must exhibit both consistency and diversity. To address these imperatives, the utilization of generative adversarial networks (GANs) and GAN inversion processes emerges as a promising avenue. The GAN model, comprising a generator $\mathcal{G}$ and a discriminator $\mathcal{D}$ [14, 13], operates iteratively. The generator $\mathcal{G}$ takes random noise as input and produces synthetic data to deceive the discriminator $\mathcal{D}$, trained to differentiate between real and generated data. Effectively, $\mathcal{G}$ seeks to minimize the maximum loss of $\mathcal{D}$, while $\mathcal{D}$ strives to maximize its accuracy in distinguishing between the two domains. The iterative training aims for a Nash equilibrium, where $\mathcal{G}$ generates images indistinguishable from real data, making it challenging for $\mathcal{D}$ to differentiate between the two domains.

The process of GAN inversion significantly enhances the applicability and fidelity of the generated images. It involves mapping real-world data back into the latent space of the trained GAN, ensuring synthetic images not only exhibit high fidelity but also capture the complex and precise structures inherent in medical imagery. In our approach, we employ the Variational Autoencoder (VAE) [31] for this inversion process. VAE encodes the input into a lower-dimensional latent space, where each dimension follows a Gaussian distribution. The encoder module maps the input into the latent space $\mathbf{z}$, and the decoder reconstructs the input from the reparameterized sample. By utilizing VAE as the inversion model, we align the latent space $\mathbf{z}$ with the GAN’s latent space, effectively aligning real-world images with their synthetic counterparts.

In this study, we introduce an innovative model termed VSG-GAN, which harnesses the combined potential of two established generative frameworks: Variational Autoencoder (VAE) and Generative Adversarial Networks (GAN). VSG-GAN represents a refined evolution of the VAE-GAN architecture, specifically designed for generating high-fidelity retinal fundus images with precise vessel structures and distinct stylistic

characteristics. An overview of the VSG-GAN architecture is presented in Figure 1.

In essence, the original retinal fundus image r is initially processed through the segmentor module $\mathcal{S}$, resulting in the extraction of vessel segmentation information denoted as v , which further delineates artery and vein classifications. This segmented vessel information v is then utilized to train the generator module $\mathcal{G}$, facilitating the synthesis of a corresponding retinal fundus image, represented as $\mathcal{G}(v)$. The discriminator module $\mathcal{D}$ is presented with three sets of image pairs for evaluation. The first pair comprises the vessel segmentation v alongside the original fundus image r , denoted as (v, r) , establishing the real image pair. The second pair involves the vessel segmentation v juxtaposed with the generated synthetic image $\mathcal{G}(v)$, represented as $(v, \mathcal{G}(v))$. The final pair incorporates the vessel segmentation v along with the smoothed image, indicated as $(v, g(r))$, where the function $g(\cdot)$ denotes the Gaussian filter application. The inclusion of this filter aims to prevent the creation of overly blurred images during the synthesis process [42, 16]. These three image pairs, (v, r) , $(v, \mathcal{G}(v))$, and $(v, g(r))$ are fed into the discriminator $\mathcal{D}$ to distinguish between real and generated images, as well as into the VGG-19 model to guide the generation process in a high-level features aspect.

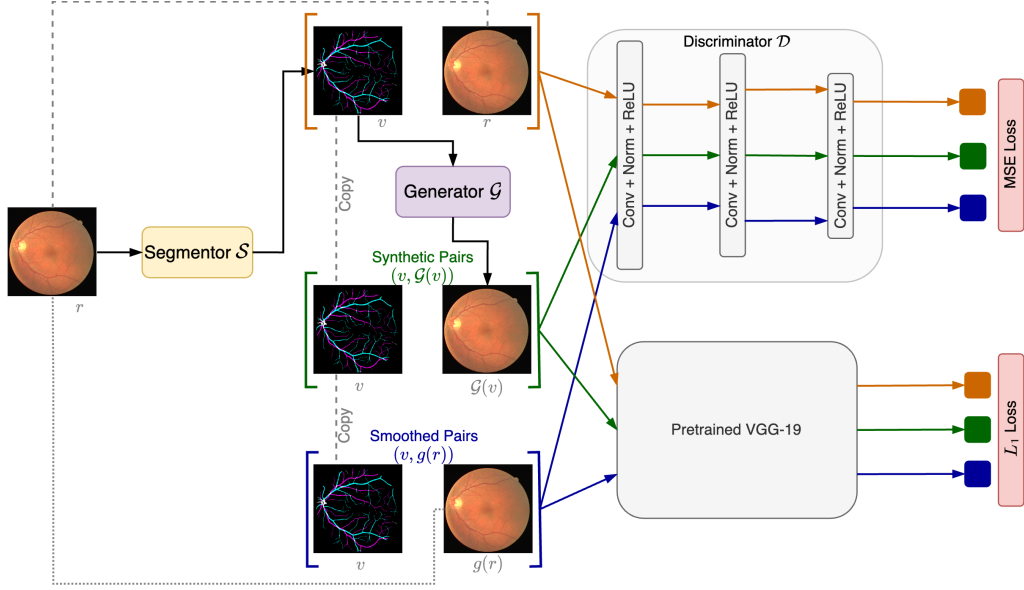


Figure 1: The overview of VSG-GAN. r denotes an original retinal image. Its vessel tree structure, denoted by v , is extracted from r through the segmentor $\mathcal{S}$. Paired with the original image r , the generated image $\mathcal{G}(v)$ through the generator $\mathcal{G}$, and the smoothed images $g(r)$ through a Gaussian filter, three sets of image pairs, (v, r) , $(v, \mathcal{G}(v))$, and $(v, g(r))$, are passed into the discriminator $\mathcal{D}$ to distinguish between real and generated images, as well as into the VGG-19 model to guide the generation process in a high-level features aspect.

Segmentor $\mathcal{S}$. For the vessel structure, we adopt the U-Net [41] as the vessel segmentation module. It is designed to work with fewer training samples but still achieve precise segmentation, which is crucial in medical applications. The model’s architecture is shown in Figure 2. The model consists of two paths. The first path is the contracting path. It consists of repeated application of convolutions, represented by green arrows, which are operations that extract features from images, followed by the rectified linear unit (ReLU) activation functions and max pooling operations that reduce the spatial dimensions of the feature maps. The second path is the expansive path, indicated by purple arrows. In this path, the spatial dimensions of the feature maps are increased through up-sampling. The up-sampling is followed by a convolution operation to refine the features, and ReLU activations are applied again. Between these two paths, there are skip connections, represented by blue arrows, that concatenate the feature maps in the contracting path to the upsampled feature maps in the expansive path. These connections help the network to localize and learn representations at various resolutions, which is critical for accurate segmentation. To increase the accuracy of the segmentation, before the segmentation process, we divide the original image into patches and restore it in the post-processing.

Generator $\mathcal{G}$. For the image generation module, to archive the idea of ensuring the synthetic image closely aligns with the given condition, we adopt the idea from GAN inversion [45, 53, 51]. We propose a novel hierarchical variational autoencoder (HVAE). It not only ensures that the reverse mapping aligns with the latent distribution of the actual data but also maps the original latent space to a multi-resolution latent space. This

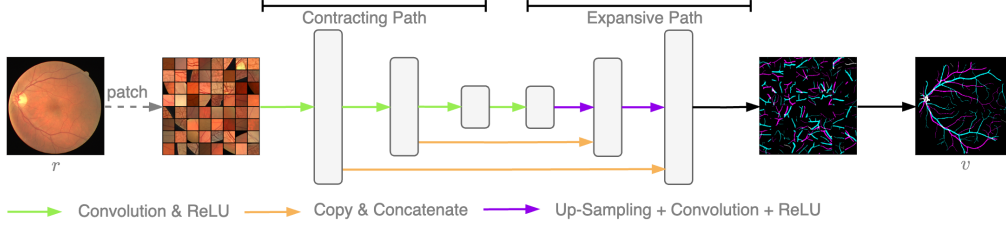


Figure 2: The segmentor $\mathcal{S}$ of VSG-GAN. It follows the U-Net’s architecture. As the input, the original image r is divided into patches and then passed through the contracting path and expansive path. Following this, the output patches are reassembled to produce the final vessel tree structure v .

multi-resolution latent space functions as a suitable representation for the synthesis of images. Furthermore, image style transfer is also considered to convert an image into an artistic painting, applying stylistic elements characteristic of various artistic styles or reference images. This procedure utilizes two input maps: the content map and the style map. The objective is to produce an output, referred to as the style-transferred image, that not only retains the content from the content map but also integrates the artistic style depicted in the style map. The proposed generator architecture is delineated in Figure 3. In the depicted generation process, a coarse-to-fine approach is followed. It begins with a randomly generated noise matrix serving as the initial input, which is then systematically upscaled across various stages of refinement. This upscaling is achieved through a series of computational blocks, Restricted ResBlock, with each block corresponding to a distinct refinement step. In each Restricted ResBlock, the vessel segmentation v provides spatial context through the Spatially-Adaptive De-normalization (SPADE) module, and the latent representation $z_i \in Z^s$, obtained from HVAE, provides extracted style through the Adaptive Instance Normalization (AdaIN) module. These inputs shape the synthesis process, allowing the model to improve the resolution and detail of the produced output $\mathcal{G}(v)$ progressively until it achieves a high-quality image. This systematic technique enables the development of detailed and cohesive visual material, with each layer building on the preceding one to boost the final image’s intricacy and depth.

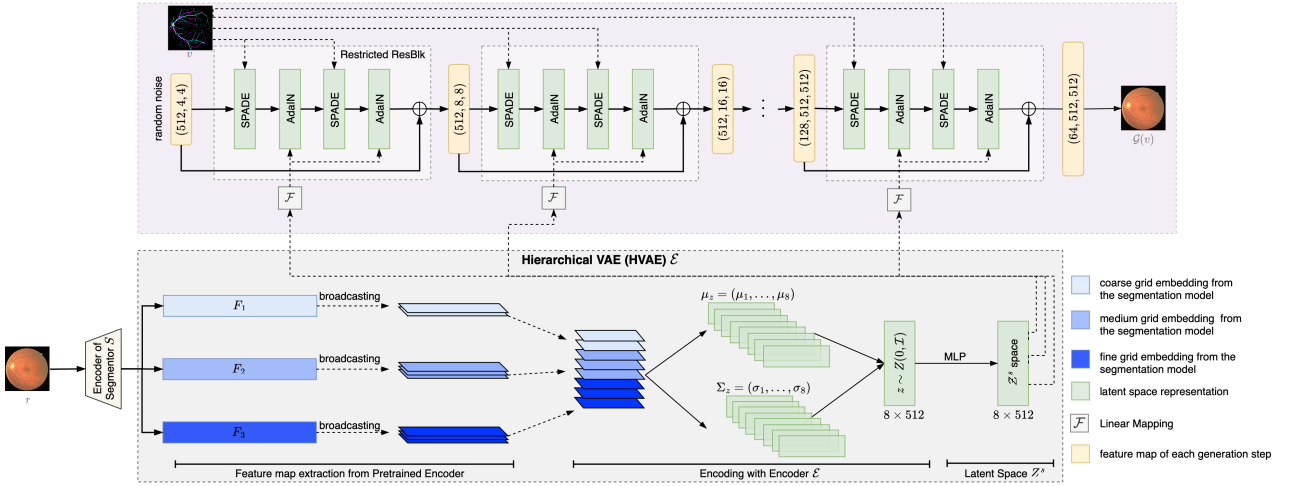


Figure 3: The generator $\mathcal{G}$ of VSG-GAN. It takes a random noise matrix as the initial input, which then undergoes successive stages of refinement through a series of Restricted ResBlocks. The latent representation extracted by the HVAE, along with the vessel segmentation v , are utilized in each block as conditions for the synthesis process of image $\mathcal{G}(v)$, employing normalization modules such as AdaIN and SPADE. These modules facilitate the model’s learning process from the specified conditions, as detailed in Equation (3) and Equation (4), respectively.

We employ HVAE [44, 48] for style extraction to preserve more of the style from the original image and capture a multi-scale latent representation, and further improve the stylization effect. This innovative approach is depicted in the lower half of Figure 3. The process begins with a pre-trained encoder that extracts feature

map $\mathbf{F} = \{F_1, \dots, F_k, \dots, F_K\}$ from the input data at different levels of granularity, each corresponding to a different resolution level. $\mathbf{F}$ is then passed through a convolutional feedforward network to generate a hierarchy of latent variables $\mathbf{z} = \{z_1, \dots, z_k, \dots, z_K\}$. These latent variables encode information at different resolutions, capturing various aspects of the input data’s structure and detail. In HVAE, the latent representation is split into K layers, as follows:

$$q(z) = \prod_{k=1}^K q(z_k|z_{<k}), \quad p(z|r) = \prod_{k=1}^K p(z_k|z_{<k}, F_k) \quad (1)$$

where each pair of $q(z_k|z_{<k})$ and $p(z_k|z_{<k}, r)$ are still follows the Gaussian Distribution. Thus, the $q(z)$ and $p(z|r)$ are the Autoregressive Gaussian process, and the KL divergence of the posterior is:

$$KL(q(z|r)||p(z)) = KL(q(z_K|r)||p(z_K)) + \sum_{k=1}^{K-1} \mathbb{E}_{q(z_{<k}|r)} \left[KL(q(z_k|z_{<k}, r)||p(z_k|z_{<k})) \right] \quad (2)$$

Unlike the HVAE approach employed in [44, 48], where it is treated as a recursive or recurrent process, our methodology leverages the precision of the segmentor $\mathcal{S}$ ’s encoder to delineate each feature map. In our approach, $\mathcal{S}$ follows the U-Net architecture, consisting of three tiers of encoder modules, as depicted in the contracting pathway of Figure 2. This stratification produces a hierarchy of features at coarse, medium, and fine resolutions, attributed to the depth of the encoding layers. By adopting this strategy, the need for introducing and training an additional encoder is eliminated, preserving the simplicity of the model and preventing any increase in its complexity. These feature maps are then used in a hierarchical style mapping mechanism to inform the generator at different levels of abstraction. The coarse grid broadcasting informs the early layers of the generator, contributing to the overall structure of the generated image. The medium grid broadcasting influences the mid-level features such as textures and patterns. Lastly, the fine grid broadcasting affects the fine details and final touches of the image output. This process allows the generator to create images with varying levels of detail, guided by latent representations. To incorporate the latent representations into the generative model’s domain, the latent space Z is expanded into the scaled latent space Z^s via a Multi-Layer Perceptron (MLP). Each latent representation is subsequently reshaped to a suitable form for subsequent processing using a corresponding linear transformation $\mathcal{F}$. This transformation facilitates the interaction between the reshaped latent representations and the feature maps derived from the input image, culminating in the production of the final image. This process ensures that the essential style and substance of the input are preserved in the generated images, while also incorporating new variations.

To embed the extracted styles $z_k^s \in Z^s$ into the generation process, we adopt the Adaptive Instance Normalization (AdaIN) algorithm proposed by Huang et al. [21], shown in Figure 4b. This algorithm leverages normalization alignment techniques to transfer the style of the style map to the content image. By aligning the normalization statistics, we can effectively transfer the desired image style to the content image. By employing the AdaIN algorithm for style transfer, we can successfully generate style-transferred fundus images that combine the content of the synthetic image with the artistic style of the real fundus image. This approach enhances the visual appeal and artistic quality of the generated images, contributing to advancements in fundus image synthesis techniques. Assuming that the style map corresponds to a real fundus image denoted as y and the content image corresponds to a synthesized image denoted as x^* , the output of AdaIN is formulated as follows:

$$\text{AdaIN}(x^*, y) = \sigma(\mathcal{F}(y)) \left(\frac{x^* - \mu(x^*)}{\sigma(x^*)} \right) + \mu(\mathcal{F}(y)) \quad (3)$$

where $\mu(\cdot)$ and $\sigma(\cdot)$ are the mean and standard deviation, $\mathcal{F}(y)$ denotes the latent representation of y extracted by a pre-trained model. we incorporate the AdaIN module in every Restricted ResBlock, as shown in the upper part of Figure 3.

In addition to the stylistic latent representation, the vessel structure also needs to be embedded in the generation process. We adopted the Spatially-Adaptive De-normalization (SPADE) module introduced by Park et al. in 2019 [37], represented in Figure 4a. Similar to the AdaIN, the SPADE module constitutes a form of normalization. However, after the data has undergone batch normalization, it transitions to denormalization—a form of affine transformation. We initially compute γ and β , which represent the parameters for scale and shift, respectively. These parameters align with the input feature map and are derived through a convolution operation involving the segmentation mask. In the context of semantic segmentation masks, each

pixel value signifies the pixel’s category. The convolution operation retains the spatial relationships between individual pixels without compromising their integrity. Consequently, the derived γ and β maintain spatial consistency within the input content mask. This aspect aids the generation process, as the content mask serves as prior information for the generation, enhancing the overall coherence and fidelity of the generated images. Assumes that the vessel tree structure v is given, and the input is x , the desired outcome of SPADE, denoted as x^* , is as follows:

$$x^* = \text{SPADE}(x, v) = \gamma \cdot \text{BN}(x) + \beta \quad (4)$$

where BN is the Batch Normalization, $\gamma = \text{Conv}_\gamma(\text{Conv}(v))$ and $\beta = \text{Conv}_\beta(\text{Conv}(v))$.

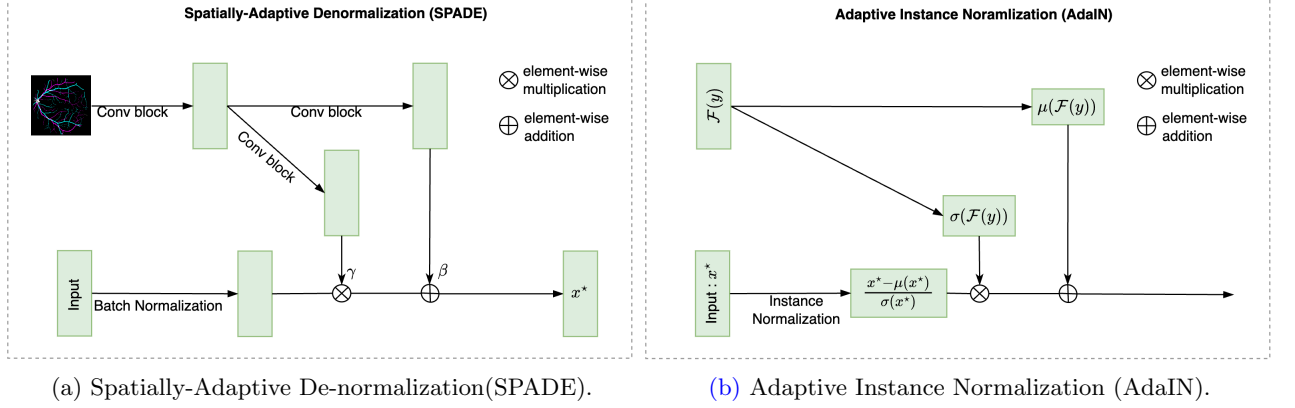


Figure 4: The overview of Spatial-Adaptive De-normalization (SPADE) module and Adaptive Instance Normalization (AdaIN). In the SPADE, as shown in Figure 4a, the vasculature is passed through two separated convolution blocks to compute the scale and shift parameters γ and β . In the AdaIN, as shown in Figure 4b, to eliminate the original background and acquire the new background, x^* is first subtracted from its mean and divided by its variance. Then, the result is multiplied by the variance of $\mathcal{F}(y)$ and added to the mean of $\mathcal{F}(y)$. Finally, the output is systematically upsampled across various Restricted Residual Blocks to the final generation of images.

By combining the AdaIN and SPADE, as illustrated in the upper section of Figure 3, depicts the Restricted ResBlock, the core module of the generation process is depicted. We incorporate two iterations of “structure restriction + style transfer” for each generation step. This approach provides the data with both content and style guidance, ensuring the production of a photorealistic image that aligns with the desired style and vasculature structure. The addition of a residual block is implemented to address gradient vanishing issues, thereby enhancing the model’s capacity to learn and generalize from the data.

Discriminator $\mathcal{D}$. Our discriminator adopts the PatchGAN’s architecture [32, 23], shown in the right upper corner in Figure 1 and the detailed is shown in Figure 5. Instead of classifying the whole image to a single “True/False” value, its output is a matrix, where each element inside the matrix corresponds to a patch of the receptive field or a patch of the input image. With this design, the discriminator can have a large receptive field to provide more accurate discrimination on the detail. To further improve the discrimination ability, we utilize multiple discriminators with different input scales: the expected output size and the downsampled by a factor of 2. With these two discriminators, we can control the global features of the generation but also the details of the images. What’s more, to avoid the discriminator accepting the blurred image, the blurred version of the original image has also been fed into the discriminator with the false label. This can improve the image sharpness of the generation.

Loss function. Multiple loss functions are used in training the HVAE ($\mathcal{E}$), generator ($\mathcal{G}$), and the discriminator ($\mathcal{D}$). For the HVAE and generator, the $\mathcal{L}_{\text{adv}}$ loss for the image reconstruction and $\mathcal{L}_{\text{KL}}$ as the constraint of the latent space. To measure the pixel-level difference between the generated image $\mathcal{G}(E(r))$ and the real image r , we adopt the perceptual loss $\mathcal{L}_{\text{per}}$ [25], where Gatys and colleagues discovered that the VGG image classification model’s learned filters are highly effective at general feature extraction. They suggested using the covariance statistics of these extracted features as a way to perceptually evaluate high-level image similarities. The purpose of using L_1 loss instead of the original L_2 is that L_1 is instrumental in preserving sharper edges and finer details, thereby effectively reducing the likelihood of blur. To stabilize the training process and

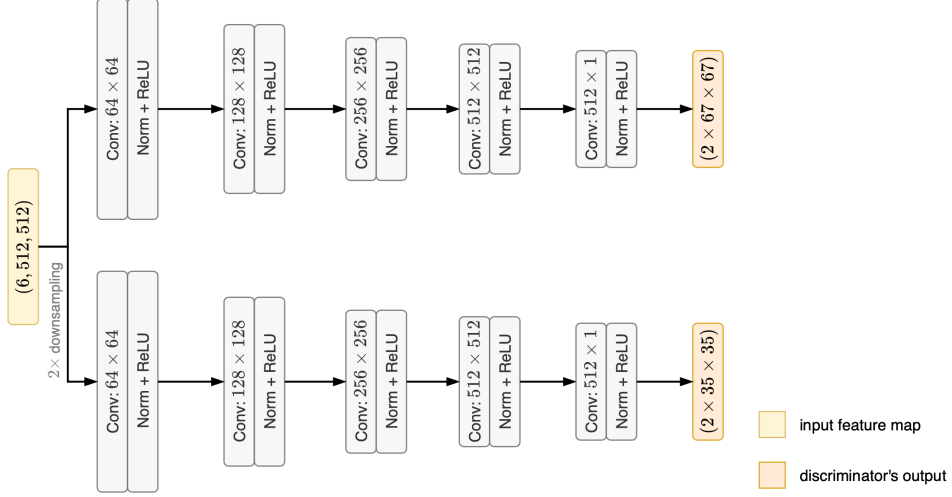


Figure 5: The discriminator $\mathcal{G}$ of VSG-GAN. The discriminator follows the PatchGAN’s architecture, where the input is the three sets of image pairs, (v, r) , $(v, \mathcal{G}(v))$, and $(v, g(r))$. The image pairs are concatenated and sent into the two distinct discriminators with or without downsampling, and the outputs are used for the loss calculation.

balance the generator and discriminator, we use a feature matching loss $\mathcal{L}_{\text{FM}}$, and adversarial loss $\mathcal{L}_{\text{adv}}$ is used to train the generator network by making it generate synthetic data that can fool the discriminator.

$$\mathcal{L}_{\mathcal{G}} = \lambda_0 \mathcal{L}_{\text{adv}} + \lambda_1 \mathcal{L}_{\text{FM}} + \lambda_2 \mathcal{L}_{\text{per}} + \lambda_3 \mathcal{L}_{\text{KL}} \quad (5)$$

Let v be the retina vasculature and r be the original retinal image, these loss functions can be denoted as follows,

1. $\mathcal{L}_{\text{adv}} = \|r, (\mathcal{G}(v))\|_2$, where v is the input of generator.
2. $\mathcal{L}_{\text{FM}} = \mathbb{E} \left[\sum_{i=1}^T \frac{1}{N_i} \|D_k^i(r), D_k^i(\mathcal{G}(v))\|_1 \right]$, where T is the total number of layers, N_i is the number of elements at layer i and D_k^i denotes the i -th layer in discriminator k .
3. $\mathcal{L}_{\text{per}} = \mathbb{E} \left[\sum_{i=1}^N \frac{1}{M_i} \|F^i(r) - F^i(\mathcal{G}(v))\|_1 \right]$, where F^i denotes the i -th layer with M_i elements of the VGG-19 network.
4. $\mathcal{L}_{\text{KL}} = \text{KL}(\mathcal{E}(r) \| N(0, I))$
5. λ_i are the corresponding weights of each loss function.

For the discriminator loss, the discriminator’s main task is not only to improve the ability to discriminate the generated from the real images but also to discriminate the smoothed and the clear images:

$$\mathcal{L}_{\mathcal{D}} = \mathbb{E} [\log (\mathcal{D}(r))] + \mathbb{E} [1 - \log (\mathcal{D}(\mathcal{G}(v)))] + \mathbb{E} [1 - \log (\mathcal{D}(\text{blur}(r)))] \quad (6)$$

Experiment preparation. The experimental platform’s detail is shown in Table 1. The computations presented in this paper were executed on the Bayes Cluster provided by the Department of Statistics and Data Science at BNU-HKBU United International College. Node A and Node B are employed to perform various tasks. Machine A is dedicated to data preprocessing, vessel segmentation, and classification experiments, while Machine B is utilized for fundus image generation.

For the segmentation, we utilized the Adam optimizer [30] with a fixed learning rate of 1×10^{-4} . The model optimizer operated with a batch size of 2 and an epoch setting of 50. Preceding the training phase, we randomly divided the image-label pairs into multiple 64×64 blocks and applied diverse random image augmentation techniques. Specifically, we employed the Albumentations package¹, incorporating flipping, rotation, CLAHE, random gamma adjustment, and random brightness/contrast modifications. Subsequently,

¹Albumentations is a Python package for image augmentation available at: <https://albumentations.ai/>

Table 1: Experiment Environment and Equipment

	Node A	Node B
Operating System	Ubuntu 20.04	Ubuntu 22.04
Python Environment	Python 3.10	Python 3.10
CUDA Version	11.7	11.7
Pytorch Version	2.0	2.0
CPU	Intel(R) Xeon(R) Silver 4114	Intel(R) Xeon(R) Gold 5218
Memory	128GB DDR4 ECC	256GB DDR4 ECC
GPU	NVIDIA TELS A P100 12GB	NVIDIA TESLA A40 48GB $\times$ 2

the augmented blocks were randomly integrated into the model for training. Regarding the RITE dataset, we utilized a pre-defined train-test split and employed a grouping method, assigning 20% of the data from the training set for validation. Concerning the LES-AV dataset, we randomly allocated 80% of the data for training (with 20% designated for validation) and used the remaining data for testing.

For the image synthesis, we employed the same optimizer configuration, maintaining a learning rate of 1×10^{-4} and β values of (0.99, 0.999) for both the generator and discriminator networks. The batch size was held constant at 2, while the epoch was set to 400. To ensure uniformity, we initially cropped the dataset to a square shape and resized it to 512×512 , refraining from any augmentation. The training and test sets were randomly partitioned using an 80:20 ratio. After completing the model training process as described earlier, we can proceed with sampling from the latent space $\mathcal{Z}+$ to generate fundus images. However, it is important to note that the latent space may include regions that are not part of the learned manifold during training. This implies that sampling points from $\mathcal{Z}+$, which are far from the latent representations of the training examples, may lead to the generation of unreliable images, such as synthetic fundus images with two discs.

To address this issue and find a meaningful path between two given points, such as z_1 and z_2 , interpolation can be employed. Linear interpolation(Lerp) is widely used for the generation; however, in the case where a Gaussian prior is used, T.White et al. [52] proposed a spherical interpolation (Slerp) method, which is particularly suitable for this scenario and yields improved results. The slerp calculation is performed using the following formula:

$$\begin{aligned} \text{Lerp}(z_1, z_2; t) &= (1 - t)z_1 + t \cdot z_2 \\ \text{Slerp}(z_1, z_2; t) &= \frac{\sin(1 - t)}{\sin(\theta)} z_1 + \frac{\sin(t\theta)}{\sin(\theta)} z_2 \end{aligned} \quad (7)$$

where the dot product $z_1 \cdot z_2 = \cos(\theta)$, and t represents the interpolation parameter ranging from 0 to 1. When $t = 0$, the slerp interpolation yields z_1 , and when $t = 1$, it produces z_2 . The slerp interpolation generates points along a geodesic on a sphere containing z_1 and z_2 for all intermediate values of t . By employing the slerp interpolation method, we can effectively explore the latent space between two given points, in our case, the points are two reference images, allowing for a smooth transition and generation of meaningful intermediate fundus images. This approach ensures that the generated images maintain the desired characteristics and adhere to the underlying distribution learned by the model during training. The result of interpolation is shown in the Section 3.

3 Results

We employ the Structural Similarity Index (SSIM) [50], Inception Score (IS) [42], Frechet Inception Distance(FID) [19] and Kernel Inception Distance(KID) [4] to quantitatively evaluate the image quality. Assumes that x and y are two given images,

1. SSIM measures the structural similarity between synthetic and real images by comparing their brightness, contrast, and structure.

$$\text{SSIM}(x, y) = \left[\frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \right]$$

where μ_x and σ_x^2 is the pixel mean and variance of x , μ_y and σ_y^2 is the pixel mean and variance of y , σ_{xy} is the cross-correlation of x and y , c_1 and c_2 are two variables to stabilize the division. A large value of

SSIM implies the synthetic image has a higher fidelity compared to the original image test set.

2. IS is a metric for evaluating the quality of the generated images based on the pre-trained Inception-V3 Network.

$$\text{IS}(G) = \exp(\mathbb{E}_{x \sim p_g} D_{\text{KL}}(p(y|x) \| p(y)))$$

where $x \sim p_g$ indicates that x is an generated image provided by the generator G , and D_{KL} is the KL-divergence. A large value of IS implies the synthetic image has more variety.

3. FID measures the features' distance between the generation and the real image by calculating the distance between these two data distributions.

$$\text{FID}(x, y) = \|\mu_x - \mu_y\|_2 + \text{Tr}(\text{Cov}_x + \text{Cov}_y) - 2(\text{Cov}_x \cdot \text{Cov}_y)^{\frac{1}{2}}$$

where Cov_x and Cov_y are the covariance matrices. A smaller FID value indicates the synthetic image is close to the original image.

4. Unlike the FID, KID uses squared maximum mean discrepancy with the polynomial kernel, $k(x, y) = (\frac{1}{d}x^T y + 1)^3$ where d is the representation dimension, to calculate the distance between the two datasets' inception representations, i.e., the feature maps extracted from the pre-trained inception models.

$$\text{KID}(x, y) = k(\phi(x), \phi(y))$$

where ϕ is the function that maps images to the Inception representation. A smaller KID value indicates the synthetic image is close to the original image.

Additionally, we utilize the image intensity histogram as a measurement to assess the colour restoration between the synthesized image and the original image. Image intensity pertains to the magnitude of light that constitutes an image. We incorporate the use of the Jensen-Shannon (JS) divergence as a means to evaluate the effectiveness of the image intensity histogram. The JS-divergence is defined as follows:

$$D_{JS}(p|q) = \frac{1}{2}D_{KL}\left(p \parallel \frac{p+q}{2}\right) + \frac{1}{2}D_{KL}\left(q \parallel \frac{p+q}{2}\right)$$

where D_{KL} is the KL-divergence, p and q are the probability distributions. The JS-divergence values span from 0 to 1, with the divergence being 0 when p equals to q .

To assess the performance of VSG-GAN in fundus image synthesis, we compare the results with other state-of-the-art methods. Two datasets, RITE [20] and MESSIDOR [9], are used in the training, validation, and testing processes. The RITE dataset consists of 40 RGB images, with 20 for training and 20 for testing. The image resolution is 565×584 . The MESSIDOR dataset contains 1200 RGB images with capturing resolutions of 440×960 , 2240×1488 , or 2304×1536 . The training set comprises 80% of the data, while the remaining 20% is used for testing. In our experiment, we crop all images into squares and reshape them to 512×512 for consistency. The RITE training set is used for training the vessel segmentation model, which is employed for semantic information extraction. The MESSIDOR training set is utilized for training the generator, discriminator, and HVAE. During the testing phase, we apply the trained model to the RITE and MESSIDOR test sets and evaluate the model using four indicators: SSIM, IS, FID, and KID.

3.1 Comparison Between VSG-GAN with other Methods

In Table 2, we present a comparative analysis of VSG-GAN against other state-of-the-art (SOTA) models concerning image synthesis performance.

The initial row showcases the method proposed by Costa et al. [8], which attained an SSIM score of 0.6636. Unfortunately, the IS score for this method is unavailable. The FID score reaches 81.81, indicating a relatively higher dissimilarity from real images. On the other hand, the KID score sits at 0.0918, suggesting a moderate similarity in terms of feature distributions.

Guibas et al.'s method yields a slightly elevated SSIM score of 0.6993 compared to Costa et al. However, akin to the former method, the IS score remains unavailable. The FID score demonstrates a decrease to 65.12, indicating reduced dissimilarity from real images compared to Costa et al. Correspondingly, the KID score shows an improved similarity in feature distributions, registering at 0.0713.

The third row showcases the results obtained through the RetiGAN method. It achieved a notably higher SSIM score of 0.8365, signaling better similarity to real images in contrast to the prior methods. Nevertheless, similar to the aforementioned methods, the IS score is unavailable. The FID score showcases a drop to 46.74, suggesting reduced dissimilarity from real images when compared to both Costa et al. and Guibas et al. Additionally, the KID score portrays increased similarity in feature distributions, reaching 0.0486.

VSG-GAN, in contrast, achieves the highest SSIM score of 0.9184, signifying the closest similarity to real images among all methods evaluated. With an IS score of 1.3101 and a standard deviation of 0.0358, it showcases good quality and diversity in the synthesized images. Moreover, the FID score diminishes to 21.57, denoting a closer resemblance to real images in statistical properties. However, the KID score escalates to 0.0104, indicating an enhanced similarity in feature distributions compared to the other methods.

In summary, the findings illustrate VSG-GAN’s superiority over other methods across SSIM, FID, and KID scores, indicating better similarity, closeness to real images, and feature distribution similarity. Additionally, VSG-GAN attains commendable quality and diversity in synthesized images, as evidenced by the IS score. Figure 6 presents 5 synthetic images generated using VSG-GAN with corresponding real images.

Table 2: Comparison Between VSG-GAN and other state-of-the-art methods with MESSIDOR dataset

Methods	SSIM $\uparrow$	IS $\uparrow$	FID $\downarrow$	KID $\downarrow$
Costa et al. [8]	0.6636	—	81.81	0.0918
Guibas et al. [15]	0.6993	—	65.12	0.0713
RetiGAN [16]	0.8365	—	46.74	0.0486
RetiGAN [16](<i>reproduced</i>)	0.8812	1.2662 ± 0.0257	33.7018	0.0264 ± 0.0021
VSG-GAN(proposed)	0.9231	1.3033 ± 0.0458	21.84	0.0111 ± 0.0014

Figure 6 also illustrates a comparison among three groups: the first group is the original images along with their respective intensity histogram of red, green, and blue channels; the second group showcases the outcomes produced by RetiGAN; and the third group reveals the results achieved through VSG-GAN. The average JS-divergence of RetiGAN with original images is $\{R : 0.3235, G : 0.3716, B : 0.4025\}$ and VSG-GAN with original images is $\{R : 0.2519, G : 0.2596, B : 0.5931\}$. Based on the given results, we can observe that VSG-GAN performs significantly better in the red(R) and green(G) channels, with lower JS-divergence values compared to the RetiGAN. However, in the blue(B) channels, both models have higher JS-divergence values, indicating that reproducing the blue channel accurately is more challenging for both models. Overall, based on the JS-divergence values, VSG-GAN appears to be closer to the original images in terms of colour distribution.

The JS-divergence in Figure 6 shows that the VSG-GAN can generate retinal images that are closer in the color distributions to the original dataset. What’s more, the horizontally aligned images for VSG-GAN indicate that this model not only produces images with a distribution of pixel intensities closer to the original images (in terms of the JS-divergence in the red and green channels), but it also maintains a diversity in the generated images. The diversity is essential for ensuring that the model does not simply memorize and reproduce a limited set of patterns, but rather can generate a variety of plausible images that are representative of the broader dataset.

3.2 Ablation Study of the proposed method

In order to assess the individual contributions of each module, we conducted a series of ablation studies, aiming at evaluating the performance of each additional module in isolation. There are two extra modules in the VSG-GAN, SPADE and HVAE. The SPADE module aims to integrate the semantic information of the given vasculature into the generation process, while the HVAE tries to inject the extracted context information from the given fundus image to enrich the style of generated images. In general, four experiments were performed, namely:

1. “Full Model”: In this configuration, we employed the original U-Net architecture while incorporating the HVAE and SPADE modules in the generator.
2. “Full Model *w.o.* HVAE”: This configuration employed the U-Net architecture without the HVAE module but still integrated the SPADE module in the generator.
3. “Full Model *w.o.* SPADE”: This configuration utilized the U-Net architecture without the SPADE module while incorporating the HVAE module in the generator.

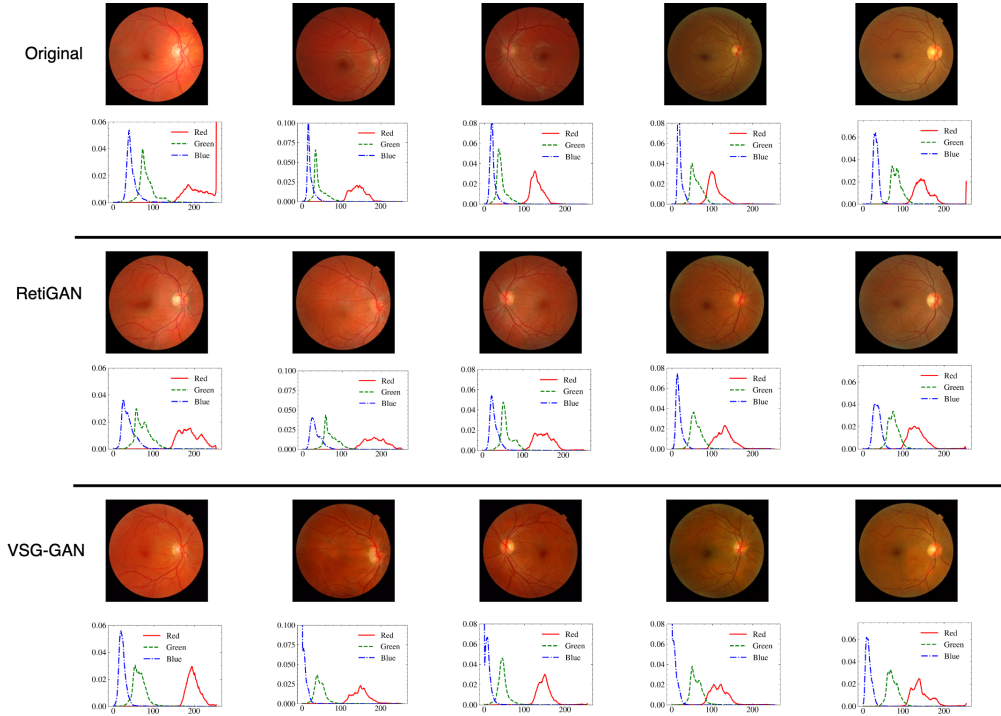


Figure 6: Synthesis comparison among images and corresponding intensity histogram from Original Ground Truth, RetiGAN and VSG-GAN. The figure shows that our VSG-GAN can generate more similar images to the original ones. The average JS-divergence of RetiGAN with original images is $\{R : 0.3235, G : 0.3716, B : 0.4025\}$ and VSG-GAN with original images is $\{R : 0.2519, G : 0.2596, B : 0.5931\}$.

4. “Full Model *w.o.* (HVAE & SPADE)”: In this configuration, both the HVAE and SPADE modules were excluded, using the original U-Net architecture without any additional modules in the generator.

Table 3 and Table 4 show the evaluation results of different configurations in both the MESSIDOR test set and RITE test set. Overall, the results demonstrate that the Full Model outperformed the other configurations across all metrics. Removing HVAE (Hierarchical Variational Autoencoders) or SPADE (Spatially-Adaptive Normalization) individually or together generally decreases performance, with the full model without both components performing the worst. The same trend is observed with the RITE test set, though the absolute metric scores differ, indicating dataset-specific performance. The HVAE appears to be critical for capturing the detailed textures and maintaining the structural integrity of the synthesized images and the SPADE helps in normalizing the feature maps in a spatially adaptive manner.

Table 3: Ablation study: Retinal Fundus Image Synthesis with VSG-GAN with MESSIDOR test set

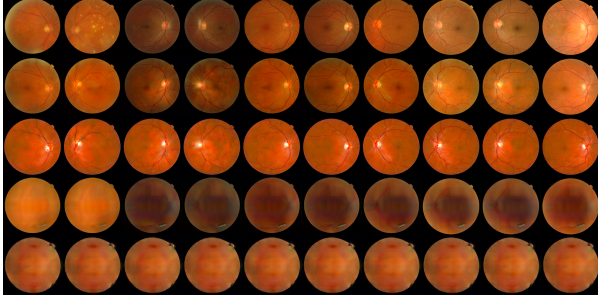
Ablation studies with MESSIDOR test set				
Exp ID	SSIM $\uparrow$	IS $\uparrow$	FID $\downarrow$	KID $\downarrow$
Full Model	0.9231	1.3033 $\pm$ 0.0458	21.84	0.0111 $\pm$ 0.0014
Full Model <i>w.o.</i> HVAE	0.7058	1.2370 $\pm$ 0.04315	53.88	0.0541 $\pm$ 0.0023
Full Model <i>w.o.</i> SPADE	0.7820	1.1511 $\pm$ 0.0339	188.85	0.2845 $\pm$ 0.0210
Full Model <i>w.o.</i> (HVAE & SPADE)	0.7622	1.0222 $\pm$ 0.0035	235.72	0.4059 $\pm$ 0.0164

Figure 7 shows the generated image by each ablated network in two distinct datasets. This can visually show the effectiveness of HVAE and SPADE modules.

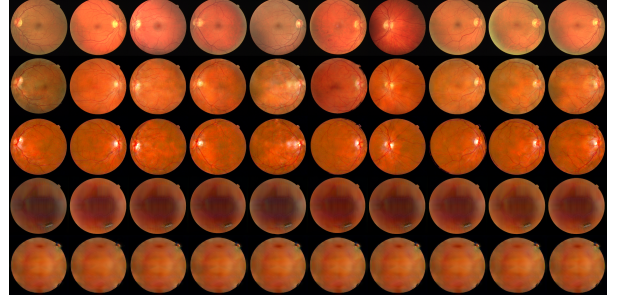
Another important aspect of generative models is their ability to generate latent spaces, which allows us to gain a deeper understanding of the generation process. In the case of VSG-GAN, the HVAE module can be viewed as an inversion of the GAN, as it provides the latent code as the initialization for the generation process. To analyze the inversion capability, we demonstrate examples of interpolation between two points in the latent space in Figure 8. The figure showcases the transformation between two given background style points, denoted

Table 4: Ablation study: Retinal Fundus Image Synthesis with VSG-GAN with RITE test set

Exp ID	Ablation studies with RITE test set			
	SSIM $\uparrow$	IS $\uparrow$	FID $\downarrow$	KID $\downarrow$
Full Model	0.6802	1.1174 $\pm$ 0.06872	63.59	0.0342 $\pm$ 0.0060
Full Model <i>w.o.</i> HVAE	0.5016	1.1277 $\pm$ 0.11742	102.25	0.1049 $\pm$ 0.011
Full Model <i>w.o.</i> SPADE	0.5385	1.0438 $\pm$ 0.02477	204.54	0.3348 $\pm$ 0.0149
Full Model <i>w.o.</i> (HVAE & SPADE)	0.6390	1.0111 $\pm$ 0.0053	224.68	0.3918 $\pm$ 0.0092



(a) Ablation Study of FundusGAN Synthesis with MESSIDOR test set



(b) Ablation Study of FundusGAN Synthesis with RITE test set

Figure 7: Ablation Study of Fundus Image Synthesis with VSG-GAN on MESSIDOR and RITE. The images are arranged from top to bottom as follows: ground truth (the 1st row), the output of VSG-GAN (the 2nd row), the output of Full Model *w.o.* HVAE (the 3rd row), the output of configuration Full Model *w.o.* SPADE (the 4th row), and output of Full Model *w.o.* (HVAE & SPADE) (the 5th row). It is evident that the results in the second row (VSG-GAN) closely resemble the ground truth depicted in the first row.

as z_A^s and z_B^s . The computation of the interpolated latent representation follows Equation (7) and let $z_{\text{new}}^s = \text{Slerp}(z_A^s, z_B^s; t)$ where $t \in [0, 0.2, 0.4, 0.6, 0.8, 1]$. Figure 8a illustrates three examples of the transformation with a fixed vessel tree structure but different background style destinations. Figure 8b showcases three examples of the transformation with a fixed background style destination but different vessel tree structures. As expected, by decoupling the implementation of the vessel architecture and the fundus background style, we can seamlessly transform an image from one style to another without compromising the vessel architecture. This separation enables a controllable generation process, where we have the flexibility to fix either the vessel tree or the background style, thus facilitating the generation of diverse variations of patients' images. This capability highlights the value of the VSG-GAN model as a valuable tool for generating supplementary images to enhance the original dataset.

3.3 Classification with Synthetic Image

We implement the use of synthetic images for downstream diabetic retinopathy grade classification tasks on the MESSIDOR dataset. This approach demonstrates the potential to generate various types of fundus images, thereby increasing the diversity of the dataset and, consequently, improving the accuracy of downstream classification tasks. The grades of diabetic retinopathy diagnoses are classified into four classes (level 0, 1, 2 and 3) based on the following measurements: the count of microaneurysms (μA), the count of hemorrhages (H), and a binary variable indicating the presence or absence of neovascularization (NV). To assess the efficacy of the suggested approach, two subsequent classification tasks are introduced. The initial task involves binary classification. We group together images classified as level 1, level 2, and level 3 as those depicting diabetic retinopathy, while images classified as level 0 are considered as representing normal retinas. The second task is the multi-class classification respective to the original four level diagnoses.

The classic image classification model ResNet[18] is selected to be employed for these tasks due to its proven effectiveness in various image recognition challenges. ResNet, known for its deep residual learning framework, addresses the vanishing gradient problem, allowing it to benefit from considerably deeper network architectures. The pretrained ResNet18 and ResNet50 are selected for their varying depths, offering a range of complexities and learning capabilities. We did not apply any preprocessing method except the basic resizing.

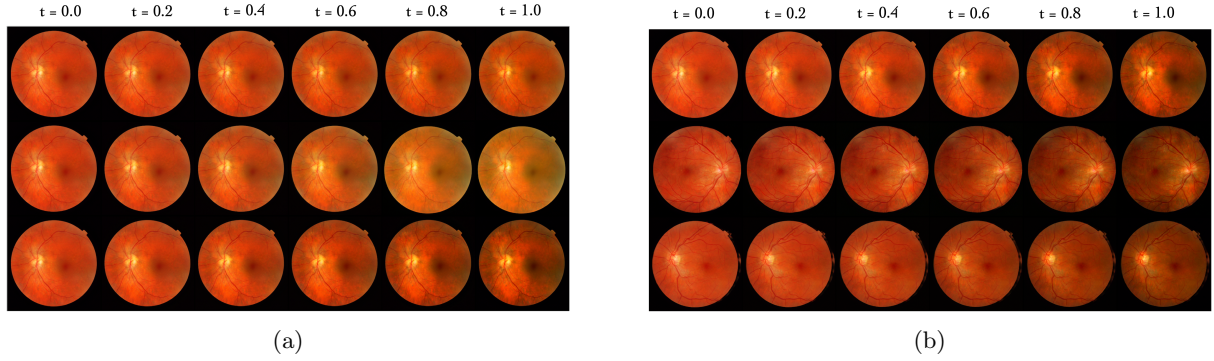


Figure 8: The examples of Spherical Interpolation. Following the Equation (7), where $t \in [0, 0.2, 0.4, 0.6, 0.8, 1]$, we implemented the interpolation in two aspects. Figure 8a illustrates three examples of the transformation from $t = 0$ to $t = 1$ with a fixed vessel tree structure but different background style destinations. The vessel tree structure can be well-preserved and not influenced by the shift of the reference image’s background style; Figure 8b showcases three examples of the transformation from $t = 0$ to $t = 1$ with a fixed background style destination but different vessel tree structures. The image’s style can be well-preserved and not influenced by the shift of the vessel tree structure. These examples show that the model can generate diverse images via interpolation with the consistency of both vasculature and background style.

This decision was made to preserve the original characteristics of the images as much as possible and aims to identify the synthetic image’s generalizability and robustness.

Synthetic images are generated by following Equation 7, with the parameter t set to the value 1. The synthesis process proceeds as follows: initially, two randomly chosen images, denoted as A and B , are selected. Their vascular structures (V_A, V_B) and background representations (S_A, S_B) are then extracted using the trained segmentor $\mathcal{S}$ and the HVAE module, respectively. To enhance the diversity of potential augmented data, the vascular structure and background representation of one image are blended with another image, resulting in two new pairs, (V_A, S_B) and (V_B, S_A) . Subsequently, these pairs are decoded by the trained generator $\mathcal{G}$ to generate two newly synthesized images. These synthesized images are characterized by their blended attributes, which are a direct consequence of the process guided by the principles of spherical interpolation. Finally, the newly synthesized images are labeled with the same label as the one that has its original background representation. That is, the newly synthesized image by (V_A, S_B) is labeled as the same label of B .

To address the class imbalance issue in the MESSIDOR dataset, we incorporated an additional 250 synthesized images into each class of level 1, level 2, and level 3 in the training image dataset using the synthesis process described above. The final classification results were obtained by employing two vanilla deep-learning models, ResNet18 and ResNet50. Comparative results for binary classification and multi-class classification can be found in Table 5 and Table 6, respectively. The results clearly demonstrate an improvement in overall accuracy for both classification tasks when synthetic images are included in the training set. This improvement highlights the effectiveness of utilizing synthetic images to enhance the performance of classifiers in the domain of medical image analysis.

Table 5: The classification accuracy comparison of Diabetic Retinopathy’s level classification (binary-classification). Two models (ResNet18 and ResNet50) are used to identify the synthetic images’ effectiveness.

Classification Report of Diabetic Retinopathy’s Level (Binary)										
	ResNet18					ResNet50				Support
	Level	Precision	Recall	F1-Score	Overall Accuracy	Precision	Recall	F1-Score	Overall Accuracy	
Original Train Set	0	0.73	0.86	0.79	0.79	0.76	0.86	0.81	0.81	110
	1	0.86	0.73	0.79		0.87	0.77	0.82		132
Augmented Train Set with VSG-GAN Synthesis	0	0.71	0.93	0.81	0.80	0.76	0.90	0.83	0.83	110
	1	0.92	0.69	0.79		0.90	0.77	0.83		132

4 Discussion

Our approach revolves around a Conditional-GAN framework that intertwines two crucial elements: the representation of image style, indicative of diverse diseases, and the intricate vascular structure. This duality

Table 6: The classification accuracy comparison of Diabetic Retinopathy’s level classification(multi-classification). Two models (ResNet18 and ResNet50) are used to identify the synthetic images’ effectiveness.

Classification Report of Diabetic Retinopathy’s Level (Multi-class)									
	ResNet18					ResNet50			
	Level	Precision	Recall	F1-Score	Overall Accuracy	Precision	Recall	F1-Score	Overall Accuracy
Original Train Set	0	0.71	0.81	0.76	0.65	0.75	0.87	0.81	0.68
	1	0.19	0.13	0.15		0.21	0.16	0.18	
	2	0.53	0.56	0.54		0.57	0.56	0.57	
	3	0.86	0.73	0.79		0.88	0.71	0.78	
Augmented Train Set with VSG-GAN Synthesis	0	0.71	0.93	0.81	0.67	0.76	0.89	0.82	0.71
	1	0.38	0.16	0.23		0.36	0.29	0.32	
	2	0.47	0.44	0.45		0.64	0.50	0.56	
	3	0.87	0.67	0.76		0.82	0.78	0.80	

guides the image generation process, employing both a label map and specific style cues. The inclusion of the SPADE module alongside the generator enables the utilization of spatial data for image generation aligned with the label map. Simultaneously, the HVAE model captures an encompassing style representation from the reference image. This segregation of semantic and contextual information empowers the generator to craft distinct representations at different stages of image creation, boosting its effectiveness in data augmentation. By maintaining a consistent vessel structure, our method generates a spectrum of image variations in various styles. Moreover, transferring styles across diverse vascular structures enriches the variety of synthetic images, facilitating dataset expansion and diversity through effective data augmentation.

Despite advancements in enhancing image synthesis quality and disentangling semantic and contextual elements, a few challenges persist. The HVAE module proficiently captures global context from reference images but lacks the inclusion of critical lesion details essential for creating fundus images. Additionally, a phenomenon termed “pattern collapse”, may observed during generative training—a concern associated with both the GAN framework and the HVAE—results in the generation of highly similar images. This collapse, akin to posterior collapse in Variational Autoencoders, restricts the model’s capacity to accurately represent the retina’s uniqueness, hampering the primary objective of generative models to diversify data. Addressing pattern collapse becomes pivotal to achieving the desired variety in synthesized retinal fundus images, ensuring an accurate representation of distinct features and variations present in real retinal images, thus enhancing the practicality and efficacy of generative models in retinal image synthesis.

Our future research endeavors will primarily focus on refining arteriovenous classification, exploring the correlation between vascular changes and eye diseases, and tackling feature decoupling in medical image synthesis. We intend to enhance our approach by integrating lesion segmentation, incorporating resulting label maps alongside vessel trees as supplementary inputs for image generation. Furthermore, we aim to delve into multi-modality translation, particularly in converting between different image types such as Optical Coherence Tomography (OCT), Scanning Laser Ophthalmoscopy (SLO), and fundus images. This becomes particularly critical in multi-task learning scenarios where diverse data types necessitate integration. Additionally, our goal is to leverage image synthesis to augment our original dataset. By assessing the impact of synthetic images on downstream tasks like blood vessel segmentation and disease classification, we aim to gain deeper insights into the utility of generated data. These initiatives align with our commitment to advancing retinal imaging technology and contributing significantly to the development of more effective diagnostic and research tools in the field of ophthalmology.

Author contributions

Conceptualization: Yuhui Deng, Huaxiong Huang

Formal analysis: Junjie Liu, Yuhui Deng, Ping He

Funding acquisition: Huaxiong Huang

Investigation: Junjie Liu, Yuhui Deng

Methodology: Junjie Liu, Yuhui Deng, Shixin Xu

Project administration: Yuhui Deng, Shixin Xu, Huaxiong Huang

Software: Junjie Liu, Yuhui Deng, Sirong Wu, Xi Luo

Supervision: Yuhui Deng, Shixin Xu, Huaxiong Huang

Validation: Junjie Liu, Yuhui Deng

Visualization: Junjie Liu, Yuhui Deng

Writing - original draft: Junjie Liu, Yuhui Deng, Shixin Xu

Writing - review & editing: Junjie Liu, Yuhui Deng, Shixin Xu, Huaxiong Huang

Declaration of interests

The authors declare no competing interests.

Acknowledgements

This work was partially supported by the National Natural Science Foundation of China (No. 12231004, 12071190); Guangdong Provincial Key Laboratory of Interdisciplinary Research and Application for Data Science, BNU-HKBU United International College, project code 2022B1212010006, UIC research grant R0400001-22; Guangdong University Innovation and Enhancement Programme Funds Featured Innovation Project 2018KTSCX278; UIC Research Grants R5201910, R201809, UICR0600036 and UICR0600048; European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 101002240). The computations in this paper were performed using the Bayes Cluster (USBC) provided by the Department of Statistics and Data Science, BNU-HKBU United International College.

References

- [1] Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, et al. The medical segmentation decathlon. *Nature communications*, 13(1):4128, 2022.
- [2] Antreas Antoniou, Amos Storkey, and Harrison Edwards. Data augmentation generative adversarial networks. *arXiv preprint arXiv:1711.04340*, 2017.
- [3] David Bau, Hendrik Strobelt, William Peebles, Jonas Wulff, Bolei Zhou, Jun-Yan Zhu, and Antonio Torralba. Semantic photo manipulation with a generative image prior. *arXiv preprint arXiv:2005.07727*, 2020.
- [4] Miłkołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018.
- [5] Philippe M Burlina, Neil Joshi, Katia D Pacheco, TY Alvin Liu, and Neil M Bressler. Assessment of deep generative models for high-resolution synthetic retinal image generation of age-related macular degeneration. *JAMA ophthalmology*, 137(3):258–264, 2019.
- [6] Stacey S Choi, Robert J Zawadzki, Michele C Lim, James D Brandt, John L Keltner, Nathan Doble, and John S Werner. Evidence of outer retinal changes in glaucoma patients as revealed by ultrahigh-resolution in vivo retinal imaging. *British journal of ophthalmology*, 95(1):131–141, 2011.
- [7] Pedro Costa, Adrian Galdran, Maria Inês Meyer, Michael David Abramoff, Meindert Niemeijer, Ana Maria Mendonça, and Aurélio Campilho. Towards adversarial retinal image synthesis. *arXiv preprint arXiv:1701.08974*, 2017.
- [8] Pedro Costa, Adrian Galdran, Maria Ines Meyer, Meindert Niemeijer, Michael Abramoff, Ana Maria Mendonça, and Aurélio Campilho. End-to-end adversarial retinal image synthesis. *IEEE transactions on medical imaging*, 37(3):781–791, 2017.
- [9] Etienne Decencière, Xiwei Zhang, Guy Cazuguel, Bruno Lay, Béatrice Cochener, Caroline Trone, Philippe Gain, Richard Ordonez, Pascale Massin, Ali Erginay, et al. Feedback on a publicly distributed image database: the messidor database. *Image Analysis & Stereology*, 33(3):231–234, 2014.
- [10] Andres Diaz-Pinto, Adrián Colomer, Valery Naranjo, Sandra Morales, Yanwu Xu, and Alejandro F Frangi. Retinal image synthesis for glaucoma assessment using dcgan and vae models. In *Intelligent Data Engineering and Automated Learning–IDEAL 2018: 19th International Conference, Madrid, Spain, November 21–23, 2018, Proceedings, Part I 19*, pages 224–232. Springer, 2018.
- [11] Lore Goetschalckx, Alex Andonian, Aude Oliva, and Phillip Isola. Ganalyze: Toward visual definitions of cognitive image properties. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5744–5753, 2019.
- [12] Mayank Golhar, Taylor L Bobrow, Saowanee Ngamruengphong, and Nicholas J Durr. Gan inversion for data augmentation to improve colonoscopy lesion classification. *arXiv preprint arXiv:2205.02840*, 2022.
- [13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.
- [14] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.
- [15] John T Guibas, Tejal S Virdi, and Peter S Li. Synthetic medical images from dual generative adversarial networks. *arXiv preprint arXiv:1709.01872*, 2017.
- [16] Xiaoxin Guo, Xinfeng Lu, Qifeng Lin, Jia Zhang, Xiaoying Hu, and Songtian Che. A novel retinal image generation model with the preservation of structural similarity and high resolution. *Biomedical Signal Processing and Control*, 78:104004, 2022.

- [17] Muhammad Salman Haleem, Liangxiu Han, Jano Van Hemert, and Baihua Li. Automatic extraction of retinal features from colour retinal images for glaucoma diagnosis: a review. *Computerized medical imaging and graphics*, 37(7-8):581–596, 2013.
- [18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [19] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.
- [20] Qiao Hu, Michael D Abràmoff, and Mona K Garvin. Automated separation of binary overlapping trees in low-contrast color retinal images. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013: 16th International Conference, Nagoya, Japan, September 22–26, 2013, Proceedings, Part II 16*, pages 436–443. Springer, 2013.
- [21] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510, 2017.
- [22] Herbert YH Hui, An Ran Ran, Jia Jia Dai, and Carol Y Cheung. Deep reinforcement learning-based retinal imaging in alzheimer’s disease: Potential and perspectives. *Journal of Alzheimer’s Disease*, (Preprint):1–12, 2023.
- [23] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.
- [24] Ali Jahanian, Lucy Chai, and Phillip Isola. On the” steerability” of generative adversarial networks. *arXiv preprint arXiv:1907.07171*, 2019.
- [25] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 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*, pages 694–711. Springer, 2016.
- [26] Sinan Kaplan and Lasse Lensu. Contrastive learning for generating optical coherence tomography images of the retina. In *International Workshop on Simulation and Synthesis in Medical Imaging*, pages 112–121. Springer, 2022.
- [27] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.
- [28] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019.
- [29] Mingyu Kim, You Na Kim, Miso Jang, Jeongeun Hwang, Hong-Kyu Kim, Sang Chul Yoon, Yoon Jeon Kim, and Namkug Kim. Synthesizing realistic high-resolution retina image by style-based generative adversarial network and its utilization. *Scientific Reports*, 12(1):17307, 2022.
- [30] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [31] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.
- [32] Chuan Li and Michael Wand. Combining markov random fields and convolutional neural networks for image synthesis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2479–2486, 2016.
- [33] Guansheng Li, Ting Ye, Zehong Xia, Sitong Wang, and Ziwei Zhu. Analysis and prediction of hematocrit in microvascular networks. *International Journal of Engineering Science*, 191:103901, 2023.

- [34] Guansheng Li, Ting Ye, Bo Yang, Sitong Wang, and Xuejin Li. Temporal-spatial heterogeneity of hematocrit in microvascular networks. *Physics of Fluids*, 35(2), 2023.
- [35] Jessica Loo, Cindy X Cai, John Choong, Emily Y Chew, Martin Friedlander, Glenn J Jaffe, and Sina Farsiu. Deep learning-based classification and segmentation of retinal cavitations on optical coherence tomography images of macular telangiectasia type 2. *British Journal of Ophthalmology*, 106(3):396–402, 2022.
- [36] Dwarikanath Mahapatra, Behzad Bozorgtabar, Sajini Hewavitharanage, and Rahil Garnavi. Image super resolution using generative adversarial networks and local saliency maps for retinal image analysis. In *Medical Image Computing and Computer Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part III 20*, pages 382–390. Springer, 2017.
- [37] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2337–2346, 2019.
- [38] Guim Perarnau, Joost Van De Weijer, Bogdan Raducanu, and Jose M Álvarez. Invertible conditional gans for image editing. *arXiv preprint arXiv:1611.06355*, 2016.
- [39] Zhihang Ren, X Yu Stella, and David Whitney. Controllable medical image generation via gan. *J. Percept. Imaging*, 5:000502–1, 2022.
- [40] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2287–2296, 2021.
- [41] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 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*, pages 234–241. Springer, 2015.
- [42] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.
- [43] Somayyeh Soltanian-Zadeh, Kazuhiro Kurokawa, Zhuolin Liu, Furu Zhang, Osamah Saeedi, Daniel X Hammer, Donald T Miller, and Sina Farsiu. Weakly supervised individual ganglion cell segmentation from adaptive optics oct images for glaucomatous damage assessment. *Optica*, 8(5):642–651, 2021.
- [44] Casper Kaae Sønderby, Tapani Raiko, Lars Maaløe, Søren Kaae Sønderby, and Ole Winther. Ladder variational autoencoders. *Advances in neural information processing systems*, 29, 2016.
- [45] Guoxian Song, Linjie Luo, Jing Liu, Wan-Chun Ma, Chunpong Lai, Chuanxia Zheng, and Tat-Jen Cham. Agilegan: stylizing portraits by inversion-consistent transfer learning. *ACM Transactions on Graphics (TOG)*, 40(4):1–13, 2021.
- [46] Joes Staal, Michael D Abràmoff, Meindert Niemeijer, Max A Viergever, and Bram Van Ginneken. Ridge-based vessel segmentation in color images of the retina. *IEEE transactions on medical imaging*, 23(4):501–509, 2004.
- [47] Nikos Tsiknakis, Dimitris Theodoropoulos, Georgios Manikis, Emmanouil Ktistakis, Ourania Boutsora, Alexa Berto, Fabio Scarpa, Alberto Scarpa, Dimitrios I Fotiadis, and Kostas Marias. Deep learning for diabetic retinopathy detection and classification based on fundus images: A review. *Computers in biology and medicine*, 135:104599, 2021.
- [48] Arash Vahdat and Jan Kautz. Nvae: A deep hierarchical variational autoencoder. *Advances in neural information processing systems*, 33:19667–19679, 2020.
- [49] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018.

- [50] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.
- [51] Tianyi Wei, Dongdong Chen, Wenbo Zhou, Jing Liao, Weiming Zhang, Lu Yuan, Gang Hua, and Nenghai Yu. E2style: Improve the efficiency and effectiveness of stylegan inversion. *IEEE Transactions on Image Processing*, 31:3267–3280, 2022.
- [52] Tom White. Sampling generative networks. *arXiv preprint arXiv:1609.04468*, 2016.
- [53] Weihao Xia, Yulun Zhang, Yujiu Yang, Jing-Hao Xue, Bolei Zhou, and Ming-Hsuan Yang. Gan inversion: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3121–3138, 2022.
- [54] Jiancheng Yang, Rui Shi, and Bingbing Ni. Medmnist classification decathlon: A lightweight automl benchmark for medical image analysis. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 191–195. IEEE, 2021.
- [55] Fangneng Zhan, Yingchen Yu, Rongliang Wu, Jiahui Zhang, Shijian Lu, Lingjie Liu, Adam Kortylewski, Christian Theobalt, and Eric Xing. Multimodal image synthesis and editing: A survey. *arXiv preprint arXiv:2112.13592*, 2021.

Appendix A The performance of the trained Segmentor $\mathcal{S}$

To evaluate the effectiveness of the segmentor $\mathcal{S}$, we utilized five different metrics for assessment:

1. ACC(Accuracy): This is the metrics that measures the overall correctness of the model. It is calculated as the ratio of correctly predicted observations (both true positives and true negatives) to the total number of observations:

$$\text{ACC} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

where “TP”, “FP”, “FN”, and “TN” represent True Positives, False Positives, False Negatives, and True Negatives, respectively.

2. DICE (Dice Coefficient): The dice coefficient compares the pixels in the segmented output with the pixels in a ground truth image, giving a measure of how similar the two images are. It ranges from 0 (no overlap) to 1 (perfect overlap):

$$\text{DICE} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}$$

3. SENS (Sensitivity): Sensitivity measures the proportion of actual positives correctly identified by the model. In medical terms, it indicates how well the test identifies patients with the condition.

$$\text{SENS} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

4. SPEC (Specificity): Specificity measure of the proportion of true negatives correctly identified. It reflects the model’s ability to identify negative results:

$$\text{SPEC} = \frac{\text{TN}}{\text{FP} + \text{TN}}$$

5. BACC(Balanced Accuracy): The BACC is similar to accuracy but adjusted for imbalanced datasets. It is the arithmetic mean of the proportion of correctly predicted positive cases and the proportion of correctly predicted negative cases:

$$\text{BACC} = \frac{\text{SENS} + \text{SPEC}}{2}$$

The Table 7 indicates that the trained segmentor $\mathcal{S}$ demonstrates strong performance in both vessel segmentation and the classification of arteries and veins.

Table 7: This table shows the performance of Segmentator $\mathcal{S}$ in both vessel segmentation task and artery/vein classification task.

	ACC $\uparrow$	BACC $\uparrow$	DICE $\uparrow$	SENS $\uparrow$	SPEC $\uparrow$
Vessel Segmentation	0.9674	0.8915	0.8087	0.9836	0.7993
Artery/Vein Classification	0.9160	0.9157	0.9264	0.9296	0.9017