*Biological Imaging* (2024), 4: e15
doi:10.1017/S2633903X24000163

RESEARCH ARTICLE

# The quest for early detection of retinal disease: 3D CycleGAN-based translation of optical coherence tomography into confocal microscopy

Xin Tian[1], Nantheera Anantrasirichai[1], Lindsay Nicholson[2] and Alin Achim[1]

[1]Visual Information Laboratory, University of Bristol, Bristol, UK
[2]Autoimmune Inflammation Research, University of Bristol, Bristol, UK
**Corresponding author**: Xin Tian; Email: xin.tian@bristol.ac.uk

**Received:** 02 January 2024; **Revised:** 18 August 2024; **Accepted:** 28 September 2024

**Keywords:** Image-to-image translation; CycleGAN; retinal image; OCT; confocal microscopy

## Abstract

Optical coherence tomography (OCT) and confocal microscopy are pivotal in retinal imaging, offering distinct advantages and limitations. *In vivo* OCT offers rapid, noninvasive imaging but can suffer from clarity issues and motion artifacts, while *ex vivo* confocal microscopy, providing high-resolution, cellular-detailed color images, is invasive and raises ethical concerns. To bridge the benefits of both modalities, we propose a novel framework based on unsupervised 3D CycleGAN for translating unpaired *in vivo* OCT to *ex vivo* confocal microscopy images. This marks the first attempt to exploit the inherent 3D information of OCT and translate it into the rich, detailed color domain of confocal microscopy. We also introduce a unique dataset, OCT2Confocal, comprising mouse OCT and confocal retinal images, facilitating the development of and establishing a benchmark for cross-modal image translation research. Our model has been evaluated both quantitatively and qualitatively, achieving Fréchet inception distance (FID) scores of 0.766 and Kernel Inception Distance (KID) scores as low as 0.153, and leading subjective mean opinion scores (MOS). Our model demonstrated superior image fidelity and quality with limited data over existing methods. Our approach effectively synthesizes color information from 3D confocal images, closely approximating target outcomes and suggesting enhanced potential for diagnostic and monitoring applications in ophthalmology.

**Impact Statement**
While OCT provides fast imaging, it can suffer from clarity issues; conversely, confocal microscopy offers cellular detailed views but at the cost of invasiveness. Our 3D deep learning image-to-image translation framework is the first to bridge optical coherence tomography (OCT) and confocal microscopy, offering rapid and noninvasive acquisition of high-resolution confocal images. This image-to-image translation method has the potential to significantly enhance diagnostic and monitoring practices in ophthalmology by overcoming the ethical and technical constraints of traditional methods.

## 1. Introduction

Multimodal retinal imaging is critical in ophthalmological evaluation, enabling comprehensive visualization of retinal structures through imaging techniques such as fundus photography, optical coherence tomography (OCT), fundus fluorescein angiography (FFA), and confocal microscopy[1–3]. Each imaging

modality manifests different characteristics of retinal structure, such as blood vessels, retinal layers, and cellular distribution. Thus, integrating images from these techniques can help with tasks such as retinal segmentation[4,5], image-to-image translation (I2I)[6,7], and image fusion[8,9]. Thereby improving the diagnosis and treatment of a wide range of diseases, from diabetic retinopathy (DR), and macular degeneration, to glaucoma[2,8].

Among these retinal imaging modalities, confocal microscopy, and OCT stand as preeminent methodologies for three-dimensional retinal imaging, each offering unique insights into the complexities of retinal anatomy. Confocal microscopy is a powerful ophthalmic imaging technique that generates detailed, three-dimensional images of biological tissues. It utilizes point illumination and point detection to visualize specific cells or structures. This allows for exceptional depth discrimination and detailed structural analysis. This technique is particularly adept at revealing the intricate cellular details of the retina, crucial for the detection of abnormalities or pathologies[10,11]. Although *in vivo* confocal microscopy enables noninvasive examination of the ocular surface, its application is confined to imaging superficial retinal layers and is constrained by a small field of view, as well as by the impact of normal microsaccadic eye movements on the quality of the images[12–14]. On the other hand, *ex vivo* confocal microscopy, requiring tissue removal from an organism, is invaluable in research settings as it offers enhanced resolution, no movement artifacts, deeper and detailed structural information, *in vitro* labeling of specific cell markers, and visualization of cellular-level pathology which is not achievable with *in vivo* methods[15,16]. The high-resolution capabilities of confocal microscopy allow for a more granular assessment of tissue health. This includes clearer visualization of changes in the appearance and organization of retinal pigment epithelium (RPE) cells, crucial in the pathogenesis of age-related macular degeneration (AMD). Moreover, it is vital for observing microvascular changes such as microaneurysms and capillary dropout in DR, and for monitoring neovascularization and its response to treatment, offering detailed insights into therapeutic effectiveness. These attributes make *ex vivo* confocal microscopy an essential tool for comprehensive retinal research. While *ex vivo* confocal microscopy can only be used to image human retina post-mortem, making it ineligible for use in regular clinical screening. Thus, it is notably beneficial to use murine retinal studies as the mouse retina shares significant anatomical and physiological similarities with the human retina[17,18]. However, *ex vivo* confocal imaging requires tissue removal with the potential to introduce artifacts through extracting and flattening the retina. Furthermore, the staining process can lead to over-coloring, uneven color distribution, or incorrect coloring, potentially complicating the interpretation of pathological features.

The OCT, on the other hand, is a noninvasive (*in vivo*) tomographic imaging technique that provides three-dimensional images of the retinal layers, offering a comprehensive view of retinal anatomy. It boasts numerous advantages, such as rapid acquisition times and the ability to provide detailed cross-sectional grayscale images, which yield structural information at the micrometer scale. Clinically, OCT is utilized extensively for its objective and quantitative measurements, crucial for assessing retinal layer thickness, edema, and the presence of subretinal fluids or lesions, thereby facilitating real-time retinal disease monitoring and diagnosis[1,19]. Although OCT provides substantial advantages for retinal imaging, it faces limitations such as diminished clarity under certain conditions and speckle noise, which manifests as a grainy texture due to the spatial-frequency bandwidth limitations of interference signals. These limitations can lead to artifacts, often exacerbated by patient movement, potentially obscuring critical details necessary for accurate diagnosis and research. However, these speckle patterns are not just noise; they are thought to contain valuable information about the retinal tissue's microstructure[20], which could be harnessed for detailed disease analysis and diagnosis.

In response to the need for a swift and noninvasive method of obtaining high-resolution, detailed confocal images, we turn to the burgeoning field of deep learning-based medical image-to-image translation (I2I)[21]. I2I is employed to transfer multimodal medical images from one domain to another, aiming to synthesize less accessible but informative images from available images. The translation supports further analytical tasks, utilizing imaging modalities to generate images that are difficult to acquire due to invasiveness, cost, or technical limitations[22–25]. Thus, it enhances the utility of existing

https://doi.org/10.1017/S2633903X24000163 Published online by Cambridge University Press

datasets and strengthens diagnostics in fields like ophthalmology[7,26], where multimodal approaches have shown advantages over uni-modal ones in the analysis and diagnosis of diabetic eye diseases (mainly DR), diabetic macular edema, and glaucoma[2,8,27–29]. In OCT to Confocal translation, I2I aims to transfer information that is challenging to visualize in OCT images into the clear, visible confocal domain, preserving the structure of OCT while enriching them with high-resolution and cellular-level details. By learning the relationship between confocal microscopy cell distribution and OCT speckle patterns, we aim to synthesize "longitudinal confocal images," revealing information traditionally obscured in OCT. This advance aids early disease detection and streamlines treatment evaluation, offering detailed retinal images without the ethical concerns or high costs of conventional confocal methods.

Common medical image-to-image translation approaches have evolved significantly with the advent of generative adversarial networks (GAN)[30,31]. For instance, the introduction of pix2pix[32], a supervised method based on conditional GANs, leveraging paired images as a condition for generating the synthetic image. However, obtaining such paired images can be challenging or even infeasible in many medical scenarios. Consequently, unpaired image-to-image translation methods, like CycleGAN[33], have emerged to fill this gap, addressing these limitations by facilitating the translation without the need for paired images. These methods have been successfully applied to modalities like MRI and CT scans[34–36], yet the challenge of translating between fundamentally different image domains, such as from 3D volumetric grayscale OCT to color confocal images at the cellular level remains relatively unexplored. Translations of this nature require not only volume preservation but also intricate cellular detail rendering in color, different from the grayscale to grayscale transitions typically seen in MRI-CT[37] or T1-weighted and T2-weighted MRI conversions[38]. This gap highlights the necessity for advanced translation frameworks capable of handling the significant complexity of OCT to confocal image translation, a domain where volumetric detail and cellular-level color information are both critical and yet to be thoroughly investigated.

In this paper, we propose a 3D modality transfer framework based on 3D CycleGAN to capture and transfer information inherent in OCT images to confocal microscopy. As registered ground truth is unavailable, the proposed framework is based on an unpaired training strategy. By extending the original CycleGAN approach, which processes 2D images slice-by-slice and often leads to spatial inconsistencies in 3D data, we incorporated 3D convolutions into our model. This adaptation effectively translates grayscale OCT volumes into rich, confocal-like colored volumes, maintaining three-dimensional context for improved consistency and continuity across slices. We also unveil the OCT2Confocal dataset, a unique collection of unpaired OCT and confocal retinal images, poised to be the first of its kind for this application. This manuscript builds upon our initial investigation of this topic, with preliminary results presented in[39]. In conclusion, the core contributions of our work are as follows:

1. We introduce a 3D CycleGAN framework that first addresses the unsupervised image-to-image translation from OCT to confocal images. The methodology exploits the inherent information of *in vivo* OCT images and transfers it to *ex vivo* confocal microscopy domain without the need for paired data.
2. Our framework effectively captures and translates three-dimensional retinal textures and structures, maintaining volumetric consistency across slices. The result shows enhanced interpretability of OCT images by synthesizing confocal-like details, which may potentially aid improved diagnostic processes without the constraints of traditional methods.
3. The introduction of the OCT2Confocal dataset, a unique collection of OCT and confocal retinal images, facilitates the development and benchmarking of cross-modal image translation research.

The remainder of this paper is organized as follows: Section 2 reviews relevant literature, contextualizing our contributions within the broader field of medical image translation. Section 3 outlines our methodological framework, detailing the architecture of our 3D CycleGAN and the rationale behind its design. Section 4 describes our novel OCT2Confocal dataset. Section 5 presents the experimental setup,

https://doi.org/10.1017/S2633903X24000163 Published online by Cambridge University Press

including the specifics of our data augmentation strategies, implementation details, and evaluation methods. Section 6 presents the results and analysis with ablation studies, dissecting the impact of various architectural choices and hyperparameter tunings on the model's performance, quantitative metrics, and qualitative assessments from medical experts. Finally, Section 7 concludes with a summary of our findings and an outlook on future directions, including enhancements to our framework and its potential applications in clinical practice.

## 2. Related work

The importance of image-to-image translation is increasingly recognized, particularly for its applications ranging from art creation and computer-aided design to photo editing, digital restoration, and especially medical image synthesis[40].

Deep generative models have become indispensable in this domain, with (i) VAEs (Variational AutoEncoders)[41] which encode data into a probabilistic latent space and reconstruct output from latent distribution samples, effectively compressing and decompressing data while capturing its statistical properties; (ii) diffusion models (DMs)[42,43] which are parameterized Markov chains, trained to gradually convert random noise into structured data across a series of steps, simulating a process that reverses diffusion or Brownian motion; and (iii) GANs[30] which employ an adversarial process wherein a generator creates data in an attempt to deceive a discriminator that is trained to differentiate between synthetic and real data. VAEs often produce blurred images lacking in detail[44], while DMs often fall short of the high standards set by GANs and are computationally slower[45]. GANs are particularly noted for their ability to generate high-resolution, varied, and style-specific images, making them especially useful in medical image synthesis[46–50]. In particular, those based on models such as StyleGAN[51] and pix2pix[32] architectures, offer significant improvements in image resolution and variety, although with certain limitations. StyleGAN, an unconditional generative adversarial network, performs well within closely related domains but falls short when faced with the need for broader domain translation. On the other hand, pix2pix operates as a conditional GAN that necessitates paired images for the generation of synthetic images. While powerful, this requirement often poses significant challenges in medical scenarios where obtaining precisely pixelwise matched, paired datasets is difficult or sometimes impossible.

Unpaired image-to-image translation methods, like CycleGAN[33], emerged to address the need for paired datasets. CycleGAN, equipped with two generators and two discriminators (two mirrored GANs), enforces style fidelity by training each generator to produce images indistinguishable from the target domain by mapping the statistical distributions from one domain to another. It utilizes the cycle consistency loss[52] to ensure the original input image can be recovered after a round-trip translation (source to target and back to source domain) to preserve the core content. This architecture has shown effectiveness in biological image-to-image translation[53] and medical image-to-image translation tasks, such as MRI and CT scan translations[34,35,54] and fluorescein angiography and retinography translations[55], demonstrating its utility in scenarios where direct image correspondences are not available and showing its capability of broader domain translation.

Notably, a significant gap remains in the translation of 3D medical images, where many existing methods simulate a 3D approach by processing images slice-by-slice rather than as complete volumes[38,56]. While some work has been done in the 3D CycleGAN space, such as in translating between diagnostic CT and cone-beam CT (CBCT)[57], these efforts have not ventured into the more complex task of translating between fundamentally different domains, such as from grayscale OCT images to full-color confocal microscopy. Such translations not only require the preservation of volumetric information but also a high-fidelity rendering of cellular details in color, distinguishing them from more common grayscale image-to-image translation.

In summary, translating OCT images into confocal is a novel problem in medical image-to-image translation. This process, which involves the translation from grayscale to full-color 3D data, has yet to be explored, particularly using a dedicated 3D network. This is the focus of our work.

https://doi.org/10.1017/S2633903X24000163 Published online by Cambridge University Press

## 3. Proposed methodology

### *3.1. Network architecture*

The proposed 3D CycleGAN method, an extension of the 2D CycleGAN architecture[33], employs 3D convolutions to simultaneously extract the spatial and depth information inherent in image stacks. Given an OCT domain $X$ and a Confocal domain $Y$, the aim of our model is to extract statistic information from both $X$ and $Y$ and then learn a mapping $G: X \rightarrow Y$ such that the output $\hat{y} = G(x)$, where $x \in X$ and $y \in Y$. An additional mapping $F$ transfers the estimated Confocal $\hat{y}$ back to the OCT domain $X$. The framework comprises two generators and two discriminators to map $X$ to $Y$ and vice versa. The input images are processed as 3D stacks, and all learnable kernels within the network are three-dimensional, as depicted in Figure 1.

### *3.2. Generators and discriminators of 3D CycleGAN*

#### *3.2.1. Generator*

The generator $G$ begins with a Convolution-InstanceNorm-ReLU layer, followed by three down-sampling layers, and nine residual blocks[58] that process the image features. It accepts an input of OCT cubes with dimensions $H \times W \times D \times C_1$, where $H, W,$ and $D$ represent height, width, and depth, respectively, and $C_1$ is the channel dimension, with $C_1 = 1$ indicating a single-channel grayscale format. Then, three fractional-strided convolution layers are used to increase the image size back to its original dimensions. Finally, the network concludes with a convolution layer that outputs the image in a 3-channel RGB format to construct confocal images, using reflection padding to avoid edge artifacts. Note that we have tested several settings, including U-Net architectures[59], WGAN-GP[60], and the nine residual blocks (ResNet 9) give the best results. The generator $F$ shares the identical architecture with the generator $G$, but its final convolution layer outputs the image in a single channel to reconstruct OCT images. It processes input dimensions $H \times W \times D \times C_2$, where $C_2 = 3$ correspond to the RGB color channels of the confocal microscopy images.

#### *3.2.2. Discriminator*

The discriminator networks in our framework are adaptations of the 2D PatchGAN[32] architecture. In our implementation, the 3D PatchGANs assess 70 × 70 × 9 voxel cubes from the 3D images to evaluate their authenticity. The key benefits of utilizing a voxel-level discriminator lie in its reduced parameter count relative to a full image stack discriminator.

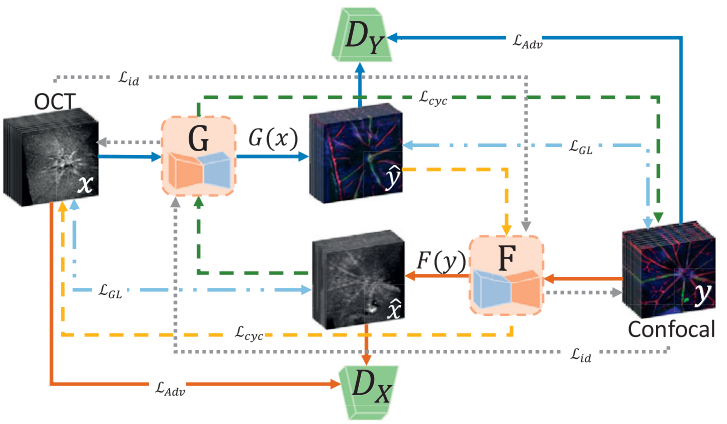



***Figure 1.*** *The proposed OCT-to-Confocal image translation method is based on 3D CycleGAN.*

https://doi.org/10.1017/S2633903X24000163 Published online by Cambridge University Press

### *3.3. The loss function*

The objective consists of four terms: (1) adversarial losses[30] for matching the distribution of generated images to the data distribution in the target domain, (2) a cycle consistency loss to prevent the learned mappings $G$ and $F$ from contradicting each other, (3) an identity loss to ensure that if an image from a given domain is transformed to the same domain, it remains unchanged, and (4) the gradient loss to enhance the textural and edge consistency in the translated images

1) **Adversarial loss:** In our model, the adversarial loss is based on the binary cross-entropy (BCE) loss, as used in traditional GANs[30]. It adapts the style of the source domain to match the target by encouraging the generators to produce outputs that are indistinguishable from the target domain images and is defined as follows:

$$\mathcal{L}_{\text{Adv}}(G, D_Y) = \mathbb{E}_{y\sim p_{\text{data}}(y)}[-\log D_Y(y)] + \mathbb{E}_{x\sim p_{\text{data}}(x)}[-\log(1 - D_Y(G(x)))], \tag{1}$$

where $G$ denotes the generator creating confocal images $G(x)$ that aim to be indistinguishable from real confocal images in domain $Y$, and $D_Y$ represents the discriminator, distinguishing between actual confocal $y$ and translated images $G(x)$. The BCE loss measures the discrepancy between the discriminator's predictions and the ground truth labels using a logarithmic function, which can be more sensitive to changes when the discriminator is making a decision. We use an equivalent adversarial BCE loss for the mapping function $F : Y \rightarrow X$ and its discriminator $D_X$ as $\mathcal{L}_{Adv}(F, D_X)$ to maintain the adversarial relationship in both translation directions. The adversarial losses ensure the translated images conform to the stylistic characteristics of the target domain.

2) **Cycle consistency loss:** Cycle consistency loss[52], defined in Equation (2), ensures the network learns to accurately translate an image $x$ from domain $X$ to domain $Y$ and back to $X$ via mappings $G$ and $F$ (forward cycle) and vice versa for an image $y$ (backward cycle), preserving the original image's integrity.

$$\mathcal{L}_{\text{cyc}}(G, F) = \mathbb{E}_{x\sim p_{\text{data}}(x)}[\|F(G(x)) - x\|_1] + \mathbb{E}_{y\sim p_{\text{data}}(y)}[\|G(F(y)) - y\|_1]. \tag{2}$$

The $L_1$ loss between the original and translated backed image minimizes information loss, ensuring that the transformed image retains essential details and the core content of the input image.

3) **Identity loss:** It was shown in[61] that adding identity losses can enhance the performance of the CycleGAN by preserving color consistency and other low-level information between the input and output, defined as follows:

$$\mathcal{L}_{\text{id}}(G, F) = \mathbb{E}_{x\sim p_{\text{data}}(x)}[\|F(x) - x\|_1] + \mathbb{E}_{y\sim p\text{data}(y)}[\|G(y) - y\|_1]. \tag{3}$$

The identity loss is calculated by taking the $L_1$ norm of the difference between a source domain image and its output after being passed through the generator designed for the opposite domain. For instance, if an OCT image is fed into a generator trained to translate confocal images to OCT images (opposite domain), the generator should ideally return the original OCT images unchanged. This process helps maintain consistent color and texture and indirectly stabilizes style fidelity.

4) **Gradient loss:** The gradient loss promotes textural fidelity and edge sharpness by minimizing the $L_1$ norm difference between the gradients of real and synthesized images[57], thereby preserving detail clarity and supporting both style rendering and information preservation through the enhancement of smooth transitions and the maintenance of edge details. The gradient loss is defined as follows:

$$\mathcal{L}_{\text{GL}}(G, F) = \mathbb{E}_{x\sim p_{\text{data}}(x)}[\|\nabla G(x) - \nabla y\|_1] + \mathbb{E}_{y\sim p_{\text{data}}(y)}[\|\nabla F(y) - \nabla x\|_1], \tag{4}$$

where $\nabla$ denotes the gradient operator. The term $\nabla G(y) - \nabla y$ represents the difference between the gradients of the generated image $G(y)$ and the real image $y$.

https://doi.org/10.1017/S2633903X24000163 Published online by Cambridge University Press

The total objective loss to minimize is the weighted summation of the four losses: the adversarial, the cyclic, the identity, and the gradient, given as follows:

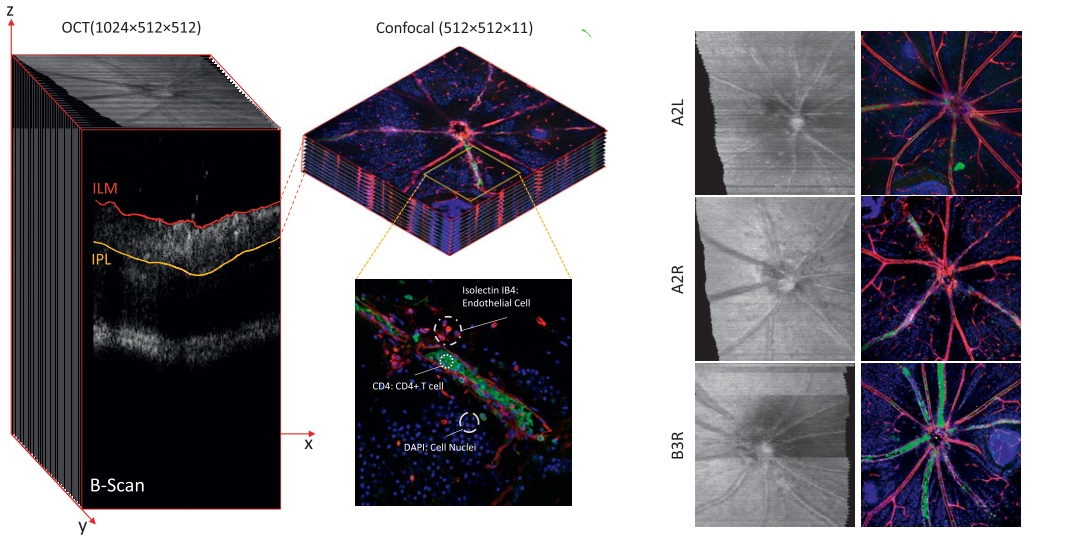
$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{Adv}}(G, D_Y) + \mathcal{L}_{\text{Adv}}(F, D_X) + \lambda_1 \mathcal{L}_{\text{cyc}} + \lambda_2 \mathcal{L}_{\text{id}} + \lambda_3 \mathcal{L}_{\text{GL}} \tag{5}$$

where $\lambda_1$, $\lambda_2$ and $\lambda_3$ are hyperparameters.

## 4. OCT2Confocal dataset

We introduce the OCT2Confocal dataset[62], to the best of our knowledge, the first to include *in vivo* grayscale OCT and corresponding *ex vivo* colored confocal images from C57BL/6 mice, a model for human disease studies[63,64], with induced autoimmune uveitis. Our dataset specifically features 3 sets of retinal images, designated as A2L, A2R, and B3R. These identifiers represent the specific mice used in the study, with "A2" and "B3" denoting the individual mice, and "L" and "R" indicating the left and right eyes, respectively. An example of the A2R data is shown in Figure 2 (a). It is important to note that although the training data consists of 3D volumes, for the sake of clarity in visualization and ease of understanding, throughout the paper we predominantly display 2D representations of the OCT and confocal images (Figure 2(b)).

a) **The *in vivo* OCT images** were captured at various time points (days 10, 14, 17, and 24) using the Micron IV fundus camera equipped with an OCT scan head and a mouse objective lens provided by Phoenix Technologies, California. The resolutions of mice OCT images are $512 \times 512 \times 1024$ ($H \times W \times D$) pixels, which is significantly smaller than human OCT images. Artifacts in OCT images, such as speckle noise and striped lines, can arise from motion artifacts, multiple scattering, attenuation artifacts, or beam-width artifacts. Volume scans, or serial B-scans (Figure 2(a)) defined at the x–z plane, were centered around the optic disc[1]. In this study, for image-to-image translation from OCT to Confocal microscopy, the OCT volumetric data captured on day 24 is utilized to align with the day when confocal microscopy images are acquired.




(a) The OCT cube with the confocal image stack of A2R

(b) The OCT projection and confocal of three mice

***Figure 2.*** *OCT2Confocal data. (a) The OCT cube with the confocal image stack of A2R, (b) The OCT projection and confocal of 3 mice.*

https://doi.org/10.1017/S2633903X24000163 Published online by Cambridge University Press

Specifically, the selected OCT volumes encompass the retinal layers between the inner limiting membrane (ILM) and inner plexiform layer (IPL) to align with the depth characteristics of the corresponding confocal microscopy images. The OCT B-scans are enhanced through linear intensity histogram adjustment and the adaptive-weighted bilateral filter (AWBF) denoising proposed by Anantrasirichai et al.[65]. The 2D OCT projection image, defined at the x-y plane (Figure 2(b)), is generated by summing up the OCT volume along the z-direction.

b) **The *ex vivo* confocal image**. After the OCT imaging phase, the mice were euthanized on day 24, and their retinas were extracted and prepared for confocal imaging. The retinas were flat-mounted, and sequential imaging was performed using adaptive optics with a Leica SP5-AOBS confocal laser scanning microscope connected to a Leica DM I6000 inverted epifluorescence microscope. The retinas were stained with antibodies attached to four distinct fluorochromes, resulting in four color channels (Figure 3):
   - Red (Isolectin IB4) channel (Figure 3(a)), staining endothelial cells lining the blood vessels. This is important as changes in retinal blood vessels can indicate a variety of eye diseases such as DR, glaucoma, and AMD.
   - Green (CD4) channel (Figure 3(b)), highlighting CD4+ T cells, which are critical in immune responses and can indicate an ongoing immune reaction in the retina.
   - Blue (DAPI) channel (Figure 3(c)), which stains cell nuclei, giving a clear picture of cell distribution.
   - White (Iba1) channel (Figure 3(d)), staining microglia and macrophages, providing insights into the state of the immune system in the retina.

   This specific representation of cell types and structures via distinct color channels referred to as the "color code," is critical for the interpretability and utility of the confocal images in retinal studies. Specifically, the blue channel represents the overall cell distribution within the retina, the green channel highlights areas of immune response, and the red channel delineates the contour of the vessels. Thus, combining these three channels, we create an RGB image encompassing a broader range of retina-relevant information, forming the training set, and providing comprehensive colored cellular detail essential for the model training process. These RGB confocal images, with their corresponding day 24 OCT images, were used for the training of the translation process. The confocal images include resolutions of A2L at 512 × 512 × 14 pixels, A2R at 512 × 512 × 11 pixels (shown in Figure 2(a)), and B3R at 512 × 512 × 14 pixels, all captured between the ILM and IPL layers.

Additionally, 23 OCT images without confocal matches from the retinal OCT dataset, also with induced autoimmune uveitis, introduced by Mellak et al.[66] were used to assess the model's translation performance

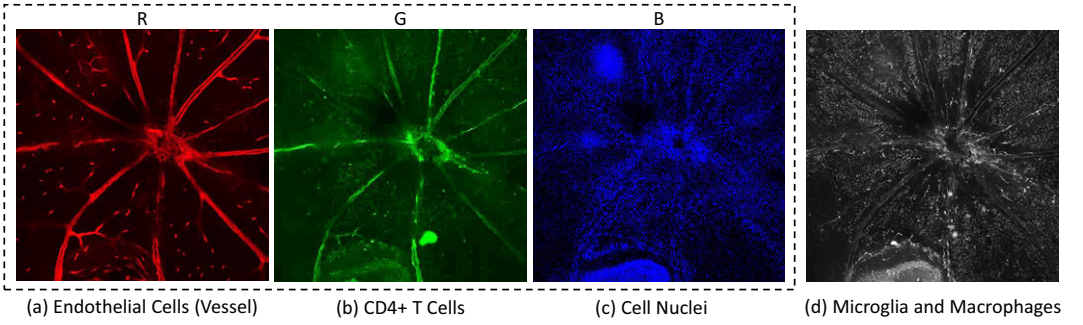



***Figure 3.*** *Example of one slice in an original four-color channel of retinal confocal image stack. The images show (from left to right):(a) Endothelial cells lining the blood vessels (red), (b) CD4+ T cells (green), (c) Cell nuclei stained with DAPI (blue), and (d) Microglia and macrophages (white).*

https://doi.org/10.1017/S2633903X24000163 Published online by Cambridge University Press

as a test dataset. The OCT images are derived from either different mice or the same mice on different days, which also makes the dataset suitable for longitudinal registration tasks as performed in[67]. This OCT2-Confocal dataset initiates the application of OCT-to-confocal image translation and holds the potential to deepen retinal analysis, thus improving diagnostic accuracy and monitoring efficacy in ophthalmology.

## 5. Experimental setup

### *5.1. Dataset augmentation*

Our dataset expansion employed horizontal flipping, random zooming (0.9–1.1 scale), and random cropping, which aligns with common augmentation practices in retinal imaging[68]. Horizontal flipping is justified by the inherent bilateral symmetry of the ocular anatomy, allowing for clinically relevant image transformations. Random zoom introduces a controlled variability in feature size, reflecting physiologic patient diversity encountered in clinical practice. Random cropping introduces translational variance and acts as a regularization technique, mitigating the risk of the model overfitting to the borders of training images. These augmentation strategies were specifically chosen to avoid the introduction of non-physiological distortions that could potentially affect clinical diagnosis.

### *5.2. Implementation details*

The implementation was conducted in Python with the PyTorch library. Training and evaluation took place on the BluePebble supercomputer[69] at the University of Bristol, featuring Nvidia V100 GPUs with 32 GB RAM, and a local workstation with RTX 3090 GPUs.

For our experiments, the OCT image cubes processed by the generator $G$ were sized $512 \times 512 \times 9 \times C_1$, with $C_1 = 1$ indicating grayscale images. Similarly, the confocal images handled by the generator $F$ had dimensions $512 \times 512 \times 9 \times C_2$, where $C_2 = 3$ represents the RGB color channels.

Optimization utilized the Adam optimizer[70] with a batch size of 1, and a momentum term of 0.5. The initial learning rate was set at $2 \times 10^{-5}$, with an input depth of 9 slices. Loss functions were configured with $\lambda_1 = 8$, $\lambda_2 = 0.1$, and $\lambda_3 = 0.1$. The 400-epoch training protocol maintained the initial learning rate for the first 200 epochs, then transitioned to a linear decay to zero over the next 200 epochs. Weights were initialized from a Gaussian distribution $\mathcal{N}(0,0.02)$, and model parameters were finalized at epoch 300 based on FID and KID performance.

### *5.3. Evaluation methods*

#### *5.3.1. Quantitative evaluation*

The quantitative evaluation of image translation quality is conducted employing Distribution-Based (DB) objective metrics[71] due to their ability to gauge image quality without necessitating a reference image. Specifically, the Fréchet inception distance (FID)[72] and KID scores[73] were utilized.

These metrics are distribution-based, comparing the statistical distribution of generated images to that of real images in the target domain. Their widespread adoption in GAN evaluations underscores their effectiveness in reflecting perceptual image quality. FID focuses on matching the exact distribution of real images using the mean and covariance of features, which can be important for capturing the precise details in medical images and the correct anatomical structures with the appropriate textures and patterns. KID, on the other hand, emphasizes the diversity and general quality of the generated images without being overly sensitive to outliers ensuring that the generated images are diverse and cover the range of variations seen in real medical images. Lower FID and KID scores correlate with higher image fidelity.

a) **FID**[72] is calculated as follows:

https://doi.org/10.1017/S2633903X24000163 Published online by Cambridge University Press

$$FID(r,g) = \|\mu_r - \mu_g\|_2^2 + \mathrm{Tr}(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{\frac{1}{2}}), \tag{6}$$

where $\mu_r$ and $\mu_g$ are the feature-wise mean of the real and generated images, respectively, derived from the feature vector set of the real image collection as obtained from the output of the Inception Net-V3[74]. Correspondingly, $\Sigma_r$ and $\Sigma_g$ are the covariance matrices for the real and generated images from the same feature vector set. Tr denotes the trace of a matrix, and $\|\cdot\|_2$ denotes the $L_2$ norm. A lower FID value implies a closer match between the generated distribution and the real image distribution. Specifically in this study, the higher-dimensional feature vector sets characterized by 768-dimensional (FID768) and 2048-dimensional (FID2048) vectors are utilized as they capture higher-level perceptual and semantic information, which is more abstract and complex compared to the direct pixel comparison done by lower-dimensional feature spaces. These higher-dimensional features are likely to include important biomarkers and tissue characteristics critical for accurate image translation.

b) **KID**[73] is calculated using the maximum mean discrepancy (MMD) with a polynomial kernel, as follows:

$$KID(r,g) = \frac{1}{m(m-1)}\sum_{i\neq j} k\left(x_i^r, x_j^r\right) + \frac{1}{n(n-1)}\sum_{i\neq j} k\left(x_i^g, x_j^g\right) - \frac{2}{mn}\sum_{i,j} k\left(x_i^r, x_j^g\right) \tag{7}$$

where $m$ and $n$ are the numbers of real and generated images, respectively, $x_i^r$ and $x_j^g$ are the feature vectors of the real and generated images, respectively, and $k(x,y)$ is the polynomial kernel function.

### *5.3.2. Qualitative evaluation*

The current objective metrics have been designed for natural images, limiting their performance when applied to medical imaging. Therefore, a subjective test leveraging a remote, crowd-based assessment was conducted for qualitative evaluation. This approach, contrasted with lab-based assessments, involved distributing the images to participants rather than hosting them in a controlled laboratory environment. The evaluation compared image-to-image translation results from five different methods: the UNSB diffusion model[43], 2D CycleGAN, and three variations of the proposed 3D CycleGAN approach. This evaluation involved a panel of experts comprising five ophthalmologists and five individuals specializing in medical image processing. Participants were tasked with evaluating and ranking five images in their relative quality score for 13 sets of images. The five images in each set are resulted from the translation process of retinal OCT to confocal image translation. To mitigate sequence bias, the order of images within each set was randomized. Scores collected from the subjective testing were quantified and expressed as a mean opinion score (MOS), which ranges from 1 to 100. Higher MOS values denote translations of greater authenticity and perceived quality. The evaluation was structured as follows:

1) Initial familiarization: The first three image sets included an original OCT image alongside its corresponding authentic confocal image and five translated confocal images from different methods and models, referred to as the with reference (W Ref) group. These were provided to acquaint the participants with the defining features of confocal-style imagery.
2) Blind evaluation: The subsequent ten sets presented only the original OCT and five translated confocal images, omitting any genuine confocal references to ensure an unbiased assessment of the translation quality, referred to as the without reference (W/O Ref) group.

https://doi.org/10.1017/S2633903X24000163 Published online by Cambridge University Press

***Table 1.*** *Correlation of selected DB and NR image quality metrics with MOS*

| | FID64 | FID192 | FID768 | FID2048 | KID | NIQE | NIQE_M | BRISQUE |
|---|---|---|---|---|---|---|---|---|
| SROCC | −0.3768 | −0.4235 | −0.7666 | −0.7823 | −0.8271 | 0.6346 | −0.5416 | 0.032 |
| LCC | −0.699 | −0.6894 | −0.7872 | −0.7813 | −0.8099 | 0.5955 | −0.5804 | −0.0446 |

*Note:* FID and KID metrics assess image quality, with lower values indicating better quality. NIQE and BRISQUE are no-reference image quality evaluators; lower NIQE scores suggest better perceptual quality, whereas BRISQUE evaluates image naturalness. SROCC and LCC measure the correlation between objective metrics and subjective MOS ratings. SROCC and LCC values closer to −1 or 1 indicate a strong correlation, with positive values suggesting a direct relationship and negative values an inverse relationship.

The participants were instructed to rank the images based on the following criteria:

- Authenticity: The degree to which the translated image replicates the appearance of a real confocal image.
- Color code preservation: Participants were advised to focus on the accuracy of color representation indicative of high-fidelity translation which are: (i) The presence of green and red color to represent different cell types, with blue indicating cell nuclei, (ii) The delineation of vessels by red, with green typically enclosed within these regions, (iii) The alternation of green in vessels, where a green vessel is usually adjacent to non-green vessels, and (iv) The co-occurrence of red and green regions with blue elements.
- Overall aesthetic: The visual appeal of the image as a whole was also considered.
- Artifact exclusion: Any artifacts that do not impact the justification of overall image content should be overlooked.

Additionally, to substantiate the reliability of selected metrics(FID768, FID2048, and KID) for evaluating OCT to confocal image translations against MOS, Spearman's rank-order correlation coefficient (SROCC) and linear correlation coefficient (LCC) were applied to both selected DB metrics and a range of no-reference (NR) metrics, including FID64, FID192, FID768, FID2048, KID, NIQE[75], NIQE_M (a modified NIQE version trained specifically with parameters from the original confocal image dataset), and BRISQUE[76]. Both SROCC and LCC range from −1 to +1, where +1 indicates a perfect positive correlation, 0 denotes no correlation, and −1 signifies a perfect negative correlation. These analyses correlate the metrics with the MOS to assess the consistency and predictive accuracy of FID and KID in reflecting subjective image quality assessments.

From Table 1, the negative correlation of FID and KID metrics with MOS, as indicated by their SROCC values, aligns with the expectation for lower-the-better metrics. Notably, KID demonstrates the strongest negative correlation (−0.8271), closely followed by FID2048 (−0.7823) and FID768 (−0.7666), suggesting their effectiveness in reflecting perceived image quality. Conversely, NIQE's positive correlation (0.6346) contradicts this principle, questioning its suitability, while the modified NIQE_M shows some improvement with a negative correlation (−0.5416). BRISQUE's low positive correlation (0.032) indicates a nearly negligible relationship with MOS. LCC results reinforce these findings, particularly highlighting KID's superior correlation (−0.8099). These analyses collectively suggest that KID, FID768, and FID2048 are relatively the most reliable metrics for evaluating the quality of translated Confocal images in this context, while the results for NIQE and BRISQUE imply limited applicability.

## 6. Results and analysis

In this section, we analyze our proposed model through ablation studies and compare it with baseline methods both quantitatively and qualitatively. For clearer visualization, results are displayed as fundus-like 2D projections from the translated 3D volume.

The ablation study investigates the impact of different generator architectures, hyperparameters of loss functions, and the number of input slices on our model's performance. This study is essential for

https://doi.org/10.1017/S2633903X24000163 Published online by Cambridge University Press

understanding how each component contributes to the efficacy of the 3D CycleGAN framework in translating OCT to confocal images.

We compare our model against the UNSB diffusion model[43] and the conventional 2D CycleGAN[33], underscoring the effectiveness of the GAN architecture and 3D network. As the UNSB and 2D CycleGAN are 2D models, 3D OCT and confocal images are processed into 2D slices along the z-direction, which are then individually translated and subsequently reassembled back into a 3D volume. This process allows us to directly compare the efficacy of 2D translation techniques on 3D data reconstruction. We evaluated against 3D CycleGAN variants: one with 2 downsampling layers (3D CycleGAN-2) without Gradient Loss, and another with the same layers but including Gradient Loss (3D CycleGAN-GL). Our final model, 3D CycleGAN-3 with 3 downsampling layers and gradient loss is also included in these comparisons. Each model is retrained on the same datasets and configurations for consistency.

### *6.1. Ablation study*

#### *6.1.1. Generator architecture*

In our experiments, the structure of the generator is found to have the most significant impact on the generated results, overshadowing other factors such as the hyperparameters of gradient loss and identity loss. From Table 2, the ResNet 9 configuration emerges as the most effective structure, outperforming both U-Net and WGAN-GP models. The ResNet 9's lower FID scores suggest a superior ability to produce high-quality images that more closely resemble the target confocal domain. While WGAN-GP attains the lowest KID score, visual assessment in Figure 4 shows that it still produces significant artifacts, underscoring the limitation of WGAN-GP and the limitation of FID and KID metrics assessing image quality in medical imaging contexts. On the other hand, the U-Net architecture, although commonly used for medical image segmentation, falls short in this generative task, particularly in preserving the definition and complex anatomical structures such as blood vessels and positions of the optic disc (where blood vessels converge), as shown in the second column of Figure 4. Meanwhile, the ResNet 9 maintains spatial consistency and detail fidelity, ensuring that synthesized images better preserve critical anatomical features, which is paramount in medical diagnostics.

#### *6.1.2. Impact of gradient and Identity loss hyperparameters*

In our evaluation of the impact of identity loss ($\lambda_2$) and gradient loss ($\lambda_3$), we explore a range of values: $\lambda_2$ at 0, 0.1, 0.5, and 1.5, and $\lambda_3$ at 0, 0.1, 0.3, and 1.0. The line graphs in Figure 5 illustrate how these values affect the FID and KID scores, with the optimal balance achieved at 0.1 for both parameters, where the fidelity, the textural, and edge details from the original OCT domain the target confocal domain are balanced.

We observe that the absence of identity loss ($\lambda_2 = 0$), as visualized in Figure 6, sometimes results in color misrepresentation in the translated images, such as pervasive blue or absent green hues,

***Table 2.*** *Comparative results of a different generator architecture in 3D CycleGAN-3. The table presents FID768, FID2048, and KID scores for U-Net, WGAN-GP, and ResNet 9 generators. Lower scores indicate better performance, with the best result colored in red*

| Generator | FID768 ↓ | FID2048 ↓ | KID ↓ |
|---|---|---|---|
| U-Net | 1.135 | 178.445 | 0.182 |
| WGAN-GP | 1.202 | 173.142 | 0.129 |
| ResNet 9 | 0.785 | 151.302 | 0.143 |

*Note:* FID768 and FID2048 refer to the Fréchet Inception Distance computed with 768 and 2048 features, respectively. KID refers to the Kernel Inception Distance. Both FID and KID indicate better performance with lower scores.

https://doi.org/10.1017/S2633903X24000163 Published online by Cambridge University Press

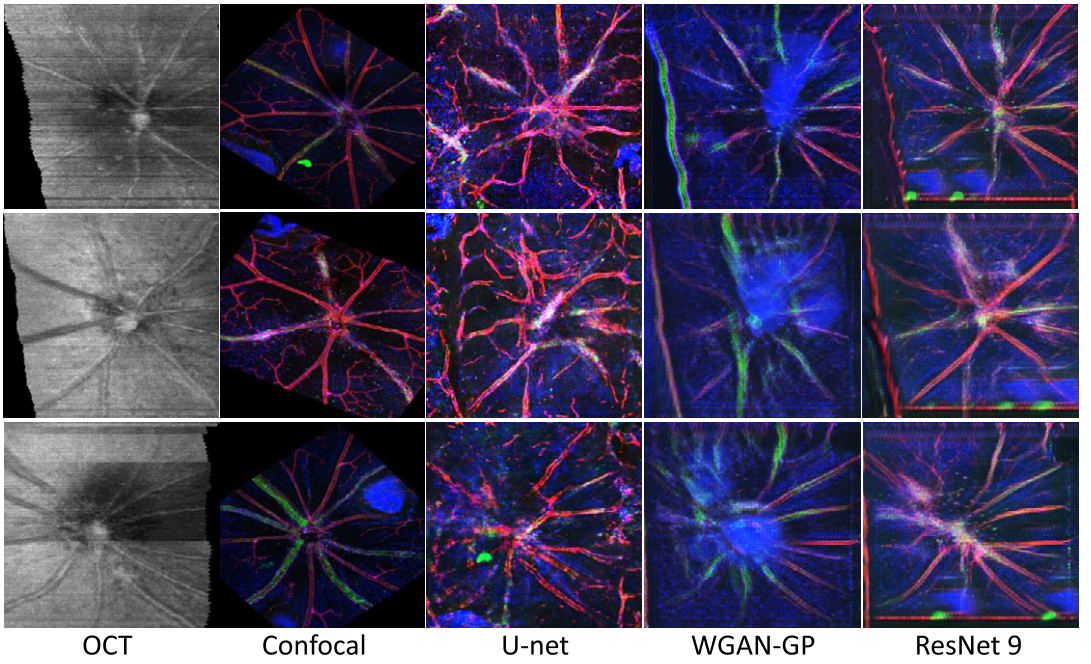
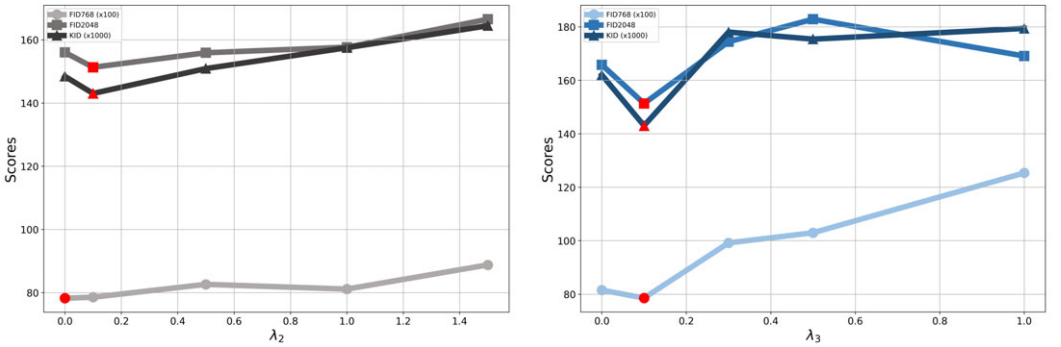



***Figure 4.*** *Visual comparison of translated images using different generator architectures. This figure displays the translated confocal images using U-Net, WGAN-GP, and ResNet 9 architectures.*




(a) Impact of Hyperparameter $\lambda_2$ of Identity Loss on FID and KID Scores

(b) Impact of Hyperparameter $\lambda_3$ of Gradient Loss on FID and KID Scores

***Figure 5.*** *Impact of Gradient and Identity Loss Hyperparameters $\lambda_2$ and $\lambda_3$ on FID and KID. The lowest (optimal) score is highlighted in red.*

underscoring its role in maintaining accurate color distribution. In contrast, overemphasizing identity loss ($\lambda_2 = 1.0$) could lead to the over-representation of specific colors, raising the likelihood of artifacts.

Similarly, without the gradient loss ($\lambda_3 = 0$), as shown in Figure 6, some images exhibit a loss of detail, particularly blurring the delineation of cellular and vascular boundaries. Conversely, an excessive gradient loss ($\lambda_3 = 1.0$) overemphasizes minor vessels in the background and over-sharpens structures, occasionally distorting primary vessels.

In conclusion, the identity loss and the gradient loss are 2 losses with small but significant weights that help the model to focus on important essential features without causing an overemphasis that could detract from the overall image quality for OCT-to-confocal image translation.

https://doi.org/10.1017/S2633903X24000163 Published online by Cambridge University Press

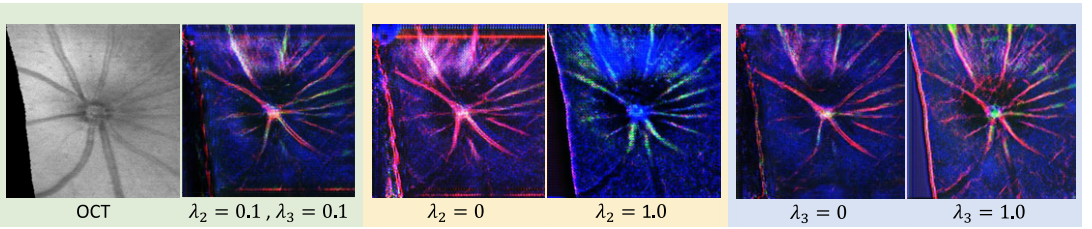



***Figure 6.*** *Visual comparison of translated confocal images with different $\lambda_2$ and $\lambda_3$ values against the optimized setting.*

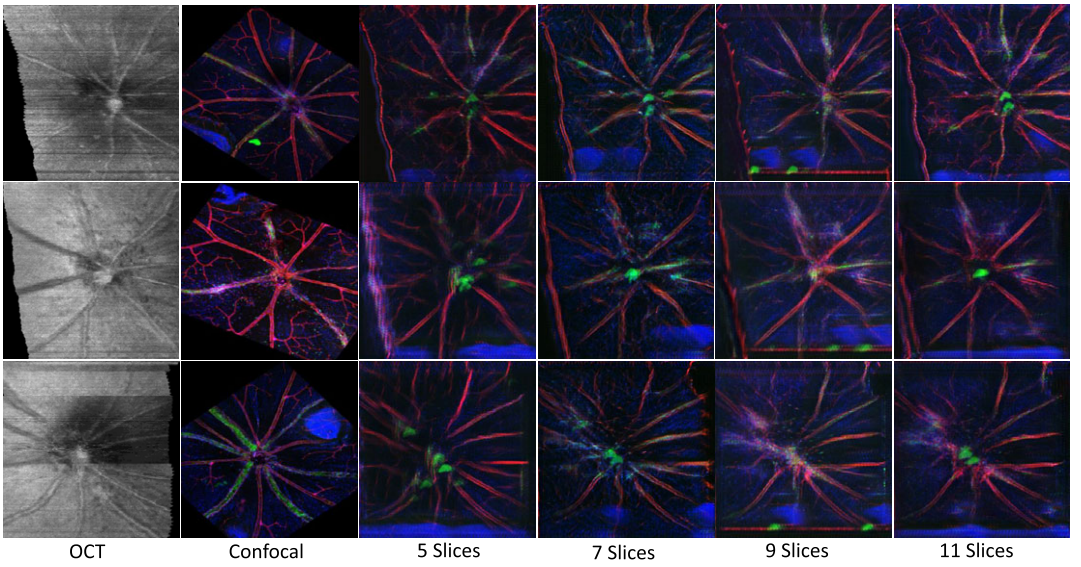



***Figure 7.*** *Visual comparison of translated images with varying input slice depths (5, 7, 9, 11 slices). This figure demonstrates the impact of different slice depths on the quality of image translation by the 3D CycleGAN-3 model.*

### *6.1.3. Impact of the input number of slices of OCT and confocal images*

In our assessment of the 3D CycleGAN model's performance with different numbers of input slices (depth) for OCT and confocal images, we experimented with 5, 7, 9, and 11 slices. Due to the limited correlation between FID and KID metrics with visual quality across different slice counts, we primarily relied on visual assessments, as detailed in Figure 7.

Our findings indicate that at a depth of 5 slices, the model frequently exhibited repetitive artifacts and blocky textures, struggling to accurately map the color distribution from confocal to OCT images, which resulted in spatial inconsistencies and shadowing effects on blood vessels. Increasing the slice count to 7 improved color code preservation, yet issues with background shadowing remained, likely due to persisting spatial discrepancies. The optimal outcome is achieved with 9 slices, which effectively represented cell color distribution and maintained edge details, with minimal artifacts confined to less critical areas such as image borders. Although 11 slices theoretically should provide further improvements, it did not significantly outperform the 9 slice input and sometimes introduced central image artifacts. Considering computational efficiency and image quality, an input depth of 9 slices is selected as the standard for our model.

https://doi.org/10.1017/S2633903X24000163 Published online by Cambridge University Press

### *6.2. Quantitative evaluation*

In Table 3, we present both the DB image quality assessment results and subjective scores of the 13 selected OCT images set used in the subjective test. Across all DB metrics, the 3D CycleGAN-3 model outperformed other methods, achieving the lowest FID and KID scores in all scenarios (with reference, without reference, and total dataset). These results suggest that this model is most effective in aligning the statistical distribution of generated images with those of real images, indicating higher image fidelity and better perceptual quality. The 3D CycleGAN-2 model follows as the second best, performing notably well in the with-reference scenario. This suggests that the additional complexity of a third downsampling layer in 3D CycleGAN-3 does confer an advantage. An inference is that an extra downsampling layer in a 3D convolutional network improves feature extraction by broadening the receptive field, enabling the model to better discern and synthesize the key structural elements within volumetric medical images. Overall, the 3D CycleGAN models outperform the UNSB diffusion model and the 2D CycleGAN, demonstrating the inadequacy of the diffusion model-based UNSB for translating OCT to confocal images and illustrating the limitations of 2D models when dealing with volumetric data.

### *6.3. Qualitative evaluation*

The 3D CycleGAN-3 model, as shown in Table 3, scored the highest in the MOS across all three scenarios as determined by the expert panel's rankings. This reflects the model's superior performance in terms of authenticity, detail preservation, and overall aesthetic quality. Notably, it also minimizes the presence of non-impactful artifacts, which is critical for the utility of translated images in clinical settings.

*Subjective test.* Analysis based on MOS and visual observations from Figure 8 and Figure 9 indicates that all 3D CycleGAN models effectively preserve blood vessel clarity, shape, and color code. The 3D CycleGAN-3 model, which received the highest MOS ratings in all scenarios, is reported to reflect the capacity for retaining more background detail and overall authenticity. Particularly in translating lower-quality *in vivo* OCT images (e.g., Set 6 in Figure 9), the 3D CycleGAN-3 model demonstrates superior performance, highlighting its effectiveness in capturing the complex relationships between OCT and confocal domains.

In contrast, the 2D models (2D CycleGAN and UNSB) sometimes introduce random colors, disregard edges, and inaccurately replicate retinal vessel color patterns. The UNSB, as a diffusion model, theoretically can generate diverse outputs. However, as indicated by its lower MOS scores and observed in Figures 8 and 9, it struggles significantly with preserving accurate color codes and structural details, leading to reduced visual quality and clinical usability in OCT to confocal translation. Conversely, CycleGAN-based models employ adversarial training to directly learn the transformation of input images into the target domain. This method is better at maintaining continuity in image quality and structure, providing visual advantages over the UNSB model. However, when compared to 3D models, these advantages diminish.

Specifically, when compared to the 3D CycleGAN-3, reconstructions from the 2D CycleGAN exhibit significant issues: assembling 2D-processed images back into 3D often results in discontinuities in blood vessels and features across slices ($z$-direction). It manifests as repeated artifacts and features at various locations across different slices (xy-plane) and duplicated structures in 2D projections. As observed in Figure 8 (Set 2) and Figure 9 (Sets 6 and 11), the 2D CycleGAN results in more visible hallucinations than 3D CycleGAN-3 in the 2D projection images, where green artifacts occur at the optic disc (the convergence point for blood vessels). Moreover, numerous fine, hallucinated vascular structures appear in the background beyond the main vascular structures, which are absent in both the original OCT and confocal images, underscoring the limitations of 2D CycleGAN in handling the complexity of 3D data structures and maintaining spatial consistency.

Figure 10 presents a boxplot of MOS for the five evaluated methods, where 3D CycleGAN models outperform 2D models in translating OCT to confocal images. Specifically, the 3D CycleGAN-3 exhibits a more concentrated distribution of scores in MOS, indicating a consensus among experts on the quality of the generated confocal images by this model, underlining its proficiency in producing consistent and

https://doi.org/10.1017/S2633903X24000163 Published online by Cambridge University Press

***Table 3.*** *The performance of models was evaluated by DB metrics FID scores and KID scores, alongside the subjective MOS rating. The results are referred to categories with reference ('W Ref'), without reference ('W/O Ref'), and total ('Total') image sets. For each column, the best result is colored in red and the second best is colored in blue*

| | W Ref | | | | W/O Ref | | | | Total | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | FID768↓ | FID2048↓ | KID↓ | MOS↑ | FID768↓ | FID2048↓ | KID↓ | MOS↑ | FID768↓ | FID2048↓ | KID↓ | MOS↑ |
| UNSB | 1.659 | 313.189 | 0.597 | 29.300 | 1.611 | 301.666 | 0.655 | 25.360 | 1.622 | 304.325 | 0.641 | 26.269 |
| 2D CycleGAN | 1.547 | 225.302 | 0.300 | 36.667 | 1.420 | 231.048 | 0.326 | 41.630 | 1.449 | 229.722 | 0.320 | 40.485 |
| 3D CycleGAN-GL | 1.281 | 202.795 | 0.267 | 50.400 | 1.170 | 169.556 | 0.215 | 49.550 | 1.195 | 177.227 | 0.227 | 49.746 |
| 3D CycleGAN–2 | 0.852 | 149.486 | 0.144 | 53.967 | 0.890 | 166.473 | 0.160 | 52.860 | 0.881 | 162.553 | 0.156 | 53.115 |
| 3D CycleGAN–3 | 0.766 | 154.756 | 0.153 | 56.867 | 0.780 | 155.188 | 0.156 | 56.350 | 0.777 | 155.089 | 0.155 | 56.469 |

*Note:* FID768 and FID2048 refer to the Fréchet Inception Distance computed with 768 and 2048 features, respectively. KID refers to the Kernel Inception Distance. Both FID and KID indicate better performance with lower scores. Mean Opinion Score (MOS) rates the subjective quality of images with higher scores reflecting better quality.

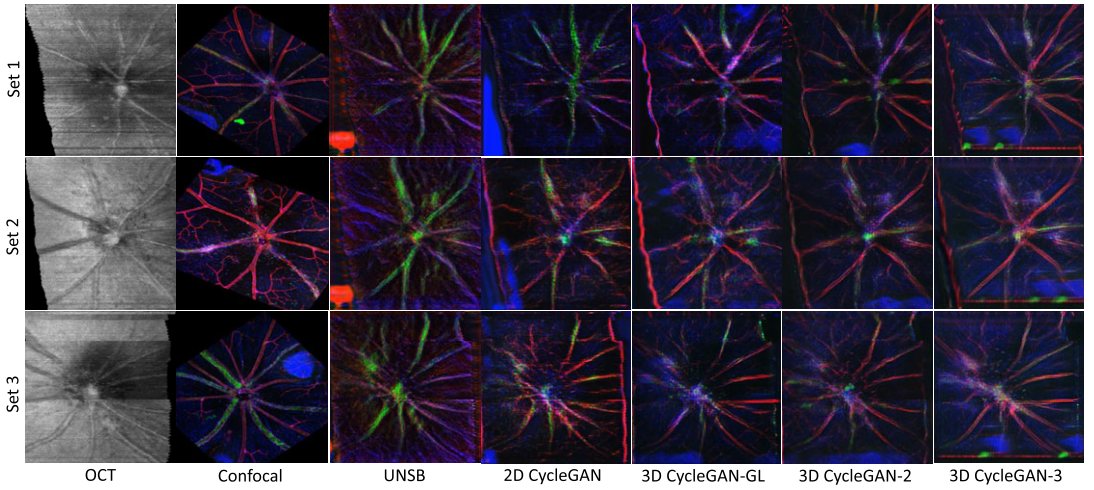



***Figure 8.*** *Visual comparative translation results with reference.*

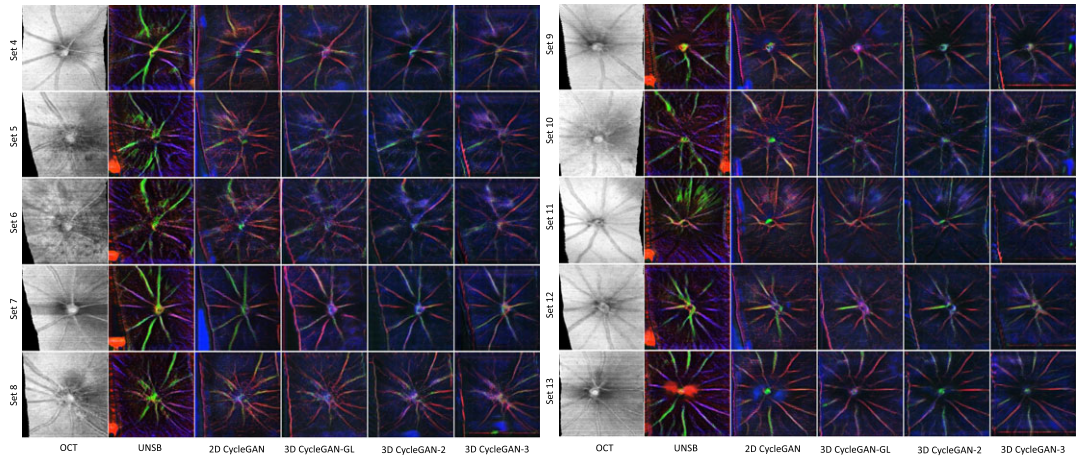



***Figure 9.*** *Visual comparative translation results without reference.*

reliable translations. The statistical analysis conducted via Kruskal–Wallis tests across each scenario confirms significant differences among the methods ($p < 0.001$). Subsequent pairwise Mann–Whitney $U$ tests with Bonferroni adjustments clearly demonstrate that the 3D CycleGAN-3 model significantly outperforms both the 2D CycleGAN and UNSB models in all scenarios evaluated. For more detailed qualitative and quantitative results, please refer to Appendix A, where Table 5 presents FID, KID, and MOS scores for each set evaluated in the subjective test.

*Ophthalmologist feedback.* In the subjective evaluations, ophthalmologists primarily assessed the clarity and shape of blood vessels, with the majority acknowledging that the 3D CycleGAN-3 model preserved blood vessel clarity and shape effectively, as well as the edges. The next aspect they considered was color code preservation, particularly the representation of the green channel, which is crucial for biological interpretation. Attention was also given to background detail, overall quality, aesthetics, and the correct distribution of colors, a critical factor for the biological accuracy of the images. For example, in scoring Set 2 of Figure 8, some experts preferred the 3D CycleGAN-3 for its accuracy in the green channel, compared to the 3D CycleGAN-GL, which displayed slightly more background vessels but less accuracy.

https://doi.org/10.1017/S2633903X24000163 Published online by Cambridge University Press

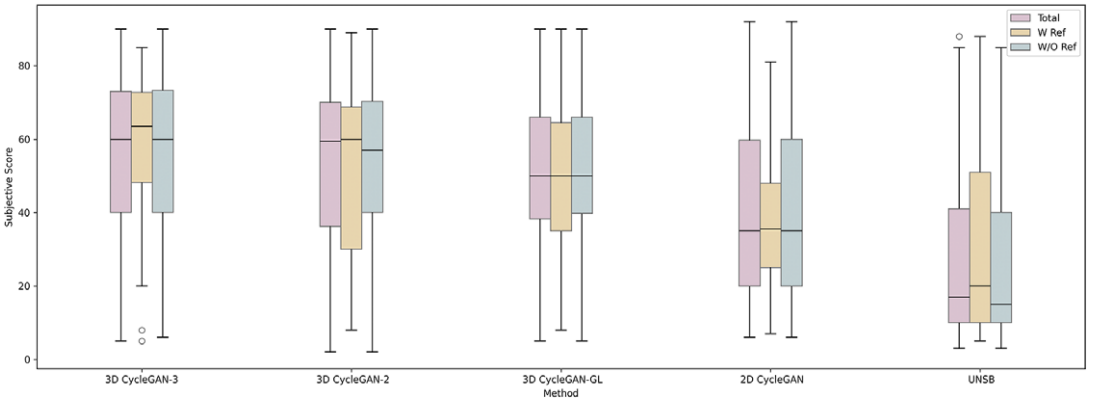



***Figure 10.*** *Boxplot of subjective evaluation scores for comparison across scenarios with reference ('W Ref'), without reference ('W/O Ref'), and the combined total ('Total'). The circles indicate outliers in the data.*

The UNSB model, however, received criticism for incorrect color code preservation. Set 6 of Figure 9 was noted for instances where the 2D CycleGAN and UNSB models ignored edges, and in Set 7 of the same figure, the 2D CycleGAN was criticized for exhibiting too much random color, missing the green staining seen in the reference, and unclear imaging.

The feedback from ophthalmologists suggests that the 3D CycleGAN-3 model not only effectively achieved a stylistic modal transfer but also more importantly preserved the biological content of the medical images, which is vital for clinical interpretation and diagnosis.

*Analysis of hallucinations.* Following feedback from ophthalmological evaluations, we now focus on analyzing how accurately our models avoid introducing hallucinations – false features not actually present in the true anatomy.

For subjectively evaluating the model hallucinations, we focus on two key biological features relevant to retinal imaging: vessel structures in the red channel and immune cell markers, CD4+ T cells, in the green channel. These features are crucial for assessing vascular structure and immune responses within the retina for uveitis, respectively.

As shown in Figure 11, our analysis in Set 2 reveals notable differences in the clarity and fidelity of vascular structures among the models. The 3D CycleGAN-3 model generally outperforms both the 2D models and other 3D variations in preserving the integrity of major blood vessels with minimal distortions. Specifically, areas highlighted in the red channel exhibit fewer hallucinated vessels, which are incorrectly generated features not aligned with the underlying anatomical structure of the retina.

Similarly, in the green channel, which focuses on the distribution of CD4+ T cells indicative of immune activity, the 3D CycleGAN-3 model shows a stronger correlation with the original confocal images in terms of brightness which indicates the immune response areas. However, artifacts around the ONH in the center are present across all models, with our proposed 3D CycleGAN-3 model demonstrating the least severity in artifact generation.

### *6.4. Computational efficiency*

To assess the computational demands of each model, we analyzed the number of parameters (#Params), number of Floating Point Operations (#FLOPs), and RunTime (RT) for each translation process. The #Params and #FLOPs represent the total number of trainable parameters and floating-point operations required to generate an image, respectively. #Params measures the model complexity and memory usage. #FLOPs provides an estimate of computational intensity, crucial for understanding the processing power required and potential latency in real-time applications. The RT is measured during inference, indicating

https://doi.org/10.1017/S2633903X24000163 Published online by Cambridge University Press

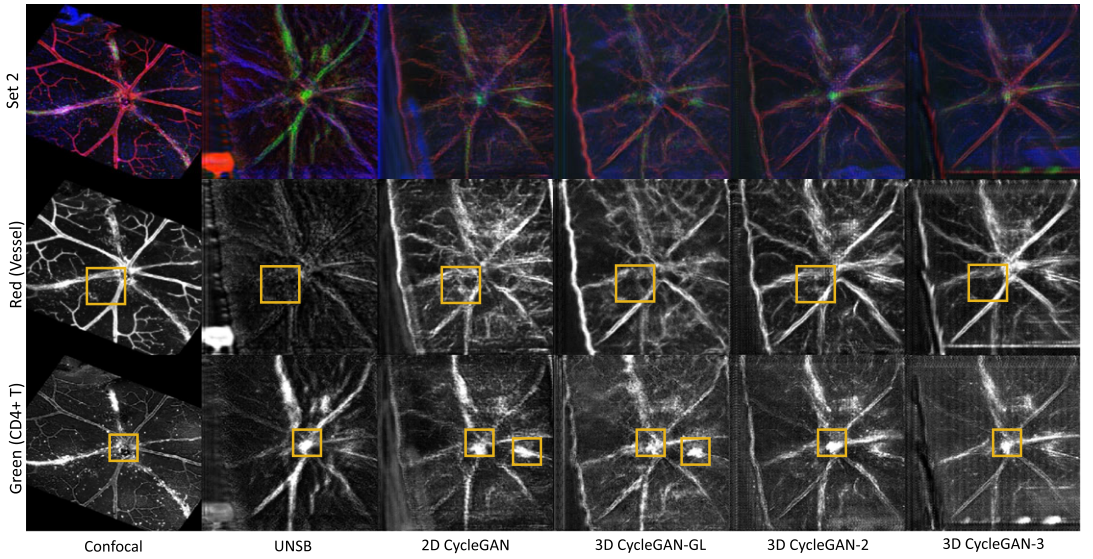



***Figure 11.*** *Example of model hallucination analysis. Focusing on the red channel for vascular structures and the green channel for CD4+ T cells. Areas highlighted (yellow boxes) show where each model introduces inaccuracies in the representation of vascular and immune cell distributions.*

***Table 4.*** *Comparative computational of different models. For each column, the red indicates the most computationally efficient values for each metric*

| Model | #Params (M) | #FLOPs (G) | RT (s) | MOS ↑ |
|---|---|---|---|---|
| UNSB | 14.684 | 253.829 | 9.891 | 26.269 |
| 2D CycleGAN | 11.378 | 631.029 | 0.945 | 40.485 |
| 3D CycleGAN-GL | 47.793 | 1323.729 | 32.140 | 49.746 |
| 3D CycleGAN–2 | 47.793 | 1323.729 | 35.250 | 53.115 |
| 3D CycleGAN–3 | 191.126 | 1585.332 | 94.451 | 56.469 |

*Note:* '#Params' denotes the total number of trainable parameters in millions (M), '#FLOPs' represents the computational complexity in billions (G) of floating-point operations, and 'RT' indicates the average execution time in seconds (s) per image. Lower values in each metric indicate more efficient computational performance. MOS (Mean Opinion Score) rates the subjective quality of images with higher scores reflecting better quality.

the practical deployment efficiency of each model, reflecting the time taken to process an image. For 2D models, the RT is calculated by summing the times required to process nine of 512 × 512 images, simulating the workflow for generating a complete 3D volume from 3D models.

Table 4 illustrates the tradeoff between computational efficiency and image quality. While the 2D models (UNSB and 2D CycleGAN) demonstrate quicker processing times, their lower MOS scores suggest a compromise in image quality. In contrast, the extended runtimes associated with 3D models, although potentially limiting for real-time applications, result in higher-quality images that are more clinically valuable, as reflected by their higher MOS scores and positive feedback from ophthalmologists.

The 2D CycleGAN, with the lowest #Params (11.378M) and moderate #FLOPs (631.029G) among all models, offers rapid inference times at 0.945s for processing 512 × 512 × 9 3D data. This indicates that 2D CycleGAN is more suitable for applications requiring quick image processing such as real-time. However, as revealed by its MOS and the feedback from quality evaluations, the lower complexity cannot adequately capture the spatial relationships and structural complexity inherent in 3D data.

Despite having a higher parameter count than a 2D model, the UNSB model exhibits relatively fewer #FLOPs (253.829G), which may be attributed to its diffusion-based generative process. Although this

https://doi.org/10.1017/S2633903X24000163 Published online by Cambridge University Press

process involves numerous iterations, each iteration consists of simpler operations, thus accumulating a lower total computational load (#FLOPs). However, the need for multiple iterations to refine image quality leads to significantly longer runtimes—up to ten times longer than the 2D CycleGAN—illustrating its inefficiency in time-sensitive scenarios.

The 3D CycleGAN models involve substantially higher #Params and #FLOPs. Particularly the 3D CycleGAN-3, with 191.126M #Params and 1585.332G #FLOPs and the longest RT of 94.451s. However, this investment in computational resources facilitates a more accurate rendering of complex 3D structures, as evidenced by its highest MOS of 56.469, suggesting superior image quality and detail retention. However, the increased computational demands of 3D models present a challenge for real-time applications, where quick processing is essential. Therefore, future efforts will focus on optimizing the computational efficiency of 3D models without compromising their ability to deliver high-quality 3D image translations to enable time-sensitive applications.

## 7. Conclusion and future work

In this paper, we present the 3D CycleGAN framework as an effective tool for translating information inherent in OCT images to the confocal domain, thereby effectively bridging *in vivo* and *ex vivo* imaging modalities. Although limited by dataset size, our quantitative and qualitative experiments showcased the 3D model's superiority over 2D models in maintaining critical image characteristics, such as blood vessel clarity and color code preservation. Our method demonstrates significant potential in providing non-invasive access to retinal confocal microscopy, which could be revolutionary for observing pathological changes, early disease detection, and studying drug responses in biomedical research. Results from our uveitis dataset could help retinal vein occlusion or retinal inflammation observation, as detailed visualization of inflammatory cell distribution (the color distribution) in the retina can provide insights into the inflammatory processes. While the current translation results require further refinement for clinical application, the potential to identify different immune cell types such as lymphocytes and monocytes and layer changes in high-resolution translated retinal images could notably enhance the assessment of immune responses and pathologic conditions in retinal diseases like AMD and DR directly from OCT scans. Thus, future efforts will focus on expanding the dataset for more accurate and higher resolution outputs and optimizing the 3D framework for computational efficiency, aiming to advance preclinical study, early disease detection, and diagnostics. In line with these enhancements, we intend to explore the integration of 2D projections with sampled 3D data for a 3D reconstruction-based OCT to confocal translation. This approach is designed to maintain the 3D spatial information while reducing computational demands. Further development will include adapting the model for human OCT to confocal translation and applying the translated results for early disease detection and enhanced diagnostic practice.

**Data availability statement.** Data are available at the University of Bristol data repository, https://data.bris.ac.uk/data/, at https://doi.org/10.5523/bris.1hvd8aw0g6l6g28fnub18hgse4. Code for training and for using the pre-trained model for translation is available on GitHub at https://github.com/xintian-99/OCT2Confocal_3DCycleGAN.

**Acknowledgements.** We would like to thank all the people from Bristol VI-Lab for their positive input and fruitful discussions during this project. We thank researchers from the Autoimmune Inflammation Research (AIR) group at the University of Bristol for their expert support. Additionally, we acknowledge that this study utilised images originally generated by Oliver H. Bell and Colin J. Chu, whose contributions have been invaluable to this work.

**Author contribution.** Conceptualization: X.T., N.N., A.A., and L.N.; Data acquisition: L.N.; Methodology: X.T., N.N., and A.A.; Model coding and training: X.T.; Writing original draft: X.T.; Writing revisions: X.T., N.N., A.A., and L.N. All authors approved the final submitted draft.

**Funding statement.** X.T. was supported by grants from the China Scholarship Council (CSC).

**Competing interest.** The authors declare no competing interests exist.

**Ethical standard.** All mice experiments were approved by the local Animal Welfare and Ethical Review Board (Bristol AWERB), and were conducted under a Home Office Project Licence.

## References


1. Abràmoff M, Garvin M and Sonka M (2010) Retinal imaging and image analysis. *IEEE Reviews in Biomedical Engineering* **3**, 169–208.
2. Kang Y, Yeung L, Lee Y, *et al.* (2021) A multimodal imaging–based deep learning model for detecting treatment-requiring retinal vascular diseases: model development and validation study. *JMIR Medical Informatics* **9**(5), e28868.
3. Meleppat R, Ronning K, Karlen S, Burns M, Pugh E and Zawadzki R (2021) In vivo multimodal retinal imaging of disease-related pigmentary changes in retinal pigment epithelium. *Scientific Reports* **2**(2), 16252.
4. Morano J, Hervella Á, Barreira N, Novo J and Rouco J (2020) Multimodal transfer learning-based approaches for retinal vascular segmentation. *arXiv preprint arXiv:2012.10160*
5. Hu Z, Niemeijer M, Abràmoff M and Garvin M (2012) Multimodal retinal vessel segmentation from spectral-domain optical coherence tomography and fundus photography. *IEEE Transactions on Medical Imaging* **31**(10), 1900–1911.
6. Vidal P, Moura J, Novo J, Penedo M and Ortega M (2012) Image-to-image translation with generative adversarial networks via retinal masks for realistic optical coherence tomography imaging of diabetic macular edema disorders. *Biomedical Signal Processing and Control* **79**, 104098.
7. Abdelmotaal H, Sharaf M, Soliman W, Wasfi E and Kedwany S (2022) Bridging the resources gap: deep learning for fluorescein angiography and optical coherence tomography macular thickness map image translation. *BMC Ophthalmology* **22**(1), 355.
8. El-Ateif S and Idri A (2024) Multimodality fusion strategies in eye disease diagnosis, *Journal of Imaging Informatics in Medicine*, **37**(5), 2524–2558.
9. Tian X, Zheng R, Chu C, Bell O, Nicholson L and Achim A (2019) Multimodal retinal image registration and fusion based on sparse regularization via a generalized minimax-concave penalty. In *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. pp. 1010–1014.
10. Paula K, Balaratnasingam C, Cringle S, McAllister I, Provis J and Yu D (2010) Microstructure and network organization of the microvasculature in the human macula, *Investigative Ophthalmology and Visual Science* **51**(12), 6735–6743.
11. Parrozzani R, D Lazzarini, A Dario, Midena E (2011) In vivo confocal microscopy of ocular surface squamous neoplasia, *Eye* **25**(4), 455–460.
12. Kojima T, Ishida R, Sato E, Kawakita T, Ibrahim O, Matsumoto Y, Kaido M, Dogru M and Tsubota K (2011) In vivo evaluation of ocular demodicosis using laser scanning confocal microscopy, *Investigative Ophthalmology and Visual Science* **52**(1), 565–569.
13. Al-Aqaba M, Alomar T, Miri A, Fares U, Otri A and Dua H (2010) Ex vivo confocal microscopy of human corneal nerves. *British Journal of Ophthalmology* **94**, 1251–1257.
14. Bhattacharya P, Edwards K and Schmid K (2022) Segmentation methods and morphometry of confocal microscopy imaged corneal epithelial cells, *Contact Lens and Anterior Eye* **45**(6), 1367–0484.
15. Yu P, Balaratnasingam C, Morgan W, Cringle S, McAllister I and Yu D (2010) The structural relationship between the microvasculature, neurons, and glia in the human retina, *Investigative Ophthalmology and Visual Science* **51**(1), 447–458.
16. Yu P, Tan PE, Morgan W, Cringle S, McAllister I and Yu D (2012) Age-related changes in venous endothelial phenotype at human retinal artery–vein crossing points, *Investigative Ophthalmology and Visual Science* **53**(3), 1108–1116.
17. Tan PE, Yu P, Balaratnasingam C, Cringle S, Morgan W, McAllister I and Yu D (2012) Quantitative confocal imaging of the retinal microvasculature in the human retina, *Investigative Ophthalmology and Visual Science* **53**(9), 5728–5736.
18. Ramos D, Navarro M, Mendes-Jorge L, Carretero A, López-Luppo M, Nacher V, Rodríguez-Baeza A and Ruberte J (2013) The use of confocal laser microscopy to analyze mouse retinal blood vessels. In *Confocal Laser Microscopy-Principles and Applications in Medicine, Biology, and the Food Sciences*. pp. 19–37.
19. Leandro I, Lorenzo B, Aleksandar M, Rosa G, Agostino A and Daniele T (2023) OCT-based deep-learning models for the identification of retinal key signs. *Scientific Reports.* **13**(1), 14628.
20. Anantrasirichai N, Achim A, Morgan J, Erchova I and Nicholson L (2013) SVM-based texture classification in optical coherence tomography. In 2013 *IEEE 10th International Symposium on Biomedical Imaging (ISBI)*. pp. 1332–1335.
21. Chen J, Chen S, Wee L, Dekker A and Bermejo I (2023) Deep learning based unpaired image-to-image translation applications for medical physics: a systematic review. *Physics in Medicine & Biology.* **68**(2), 05TR01.
22. Wang J, Zhao Y, Noble J and Dawant B (2018) Conditional generative adversarial networks for metal artifact reduction in CT images of the ear: a literature review. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference*. pp. 3–11.
23. Liao H, Huo Z, Sehnert W, Zhou S and Luo J (2018) Adversarial sparse-view CBCT artifact reduction. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference*. pp. 154–162.
24. Zhao Y, Liao S, Guo Y, Zhao L, Yan Z, Hong S, Hermosillo G, Liu T, Zhou X and Zhan Y (2018) Towards MR-only radiotherapy treatment planning: synthetic CT generation using multi-view deep convolutional neural networks. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference*. pp. 286–294.
25. Mahapatra D, Bozorgtabar B, Thiran J and Reyes M (2018) Efficient active learning for image classification and segmentation using a sample selection and conditional generative adversarial network. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. pp. 580–588.


https://doi.org/10.1017/S2633903X24000163 Published online by Cambridge University Press

26. Mahapatra D, Bozorgtabar B, Hewavitharanage S and Garnavi R (2017) 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*. pp. 382–390.
27. Kalloniatis M, Bui B and Phu J (2024) Glaucoma: Challenges and opportunities, *Clinical and Experimental Optometry* **107** (2), 107–109.
28. El-Ateif S and Idri A (2023) Eye diseases diagnosis using deep learning and multimodal medical eye imaging. *Multimedia Tools and Applications*, **83**(10), 30773–30818.
29. Hasan M, Phu J, Sowmya A, Meijering E and Kalloniatis M (2024) Artificial intelligence in the diagnosis of glaucoma and neurodegenerative diseases. *Clinical and Experimental Optometry* **107**(2), 130–146.
30. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A and Bengio Y (2014) Generative adversarial nets. *Advances in Neural Information Processing Systems (NeurIPS)* **2**, 2678–2680.
31. Shi D, Zhang W, He S, *et al.* (2023) Translation of color fundus photography into fluorescein angiography using deep learning for enhanced diabetic retinopathy screening. *Ophthalmology Science* **3**(4), 100401.
32. Isola P, Zhu J, Zhou T and Efros A (2017) Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 1125–1134.
33. Zhu J, Park T, Isola P and Efros A (2017) Unpaired image-to-image translation using cycle-consistent adversarial networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*. pp. 2242–2251.
34. Xia Y, Monica J, Chao W, Hariharan B, Weinberger K and Campbell M (2022) Image-to-image translation for autonomous driving from coarsely-aligned image pairs. *arXiv preprint* arXiv:2209.11673.
35. Zhang Z, Yang L and Zheng Y (2018) Translating and segmenting multimodal medical volumes with cycle- and shape-consistency generative adversarial network. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 9242–9251.
36. Boulanger M, Nunes Jean-Claude, Chourak H, Largent A, Tahri S, Acosta O, De Crevoisier R, Lafond C and Barateau A (2021) Deep learning methods to generate synthetic CT from MRI in radiotherapy: A literature review, *Physica Medica* **89**, 265–281.
37. Gu X, Zhang Y, Zeng W, *et al.* (2023) Cross-modality image translation: CT image synthesis of MR brain images using multi generative network with perceptual supervision. *Computer Methods and Programs in Biomedicine* **237**, 107571.
38. Welander P, Karlsson S and Eklund A (2018) Generative adversarial networks for image-to-image translation on multi-contrast MR images-a comparison of CycleGAN and UNIT. *arXiv preprint* arXiv:1806.07777.
39. Tian X, Anantrasirichai N, Nicholson L and Achim A (2023) OCT2Confocal: 3D CycleGAN based translation of retinal OCT images to confocal microscopy. *arXiv preprint* arXiv:2311.10902.
40. Baraheem S, Le T and Nguyen T (2023) Image synthesis: A review of methods, datasets, evaluation metrics, and future outlook. *Artificial Intelligence Review* **56**(10), 10813–10865.
41. Doersch C (2016) Tutorial on variational autoencoders. *arXiv preprint* arXiv:1606.05908.
42. Ho J, Jain A and Abbeel P (2020) Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems (NeurIPS)* **33**, 6840–6851.
43. Kim B, Kwon G, Kim K and Ye J (2023) Unpaired image-to-image translation via neural Schrödinger bridge. *arXiv preprint* arXiv:2305.15086.
44. Zhao S, Song J and Ermon S (2017) Towards deeper understanding of variational autoencoding models. *arXiv preprint* arXiv:1702.08658.
45. Dalmaz O, Saglam B, Elmas G, Mirza M and Çukur T (2023) Denoising diffusion adversarial models for unconditional medical image generation. In *2023 31st Signal Processing and Communications Applications Conference (SIU)*. pp. 1–5.
46. Brock A, Donahue J and Simonyan K (2018) Large scale GAN training for high fidelity natural image synthesis. *arXiv preprint* arXiv:1809.11096.
47. Karras T, Laine S, Aittala M, Hellsten J, Lehtinen J and Aila T (2020) Analyzing and improving the image quality of StyleGAN. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*. pp. 8110–8119.
48. Armanious K, Jiang C, Fischer M, *et al.* (2020) MedGAN: Medical image translation using GANs. *Computerized Medical Imaging and Graphics* **79**, 101684.
49. Wang R, Butt D, Cross S, Verkade P and Achim A (2023) Bright-field to fluorescence microscopy image translation for cell nuclei health quantification. *Biological Imaging* **3**, e12. https://doi.org/10.1017/S2633903X23000120.
50. Li W, Kong W, Chen Y, Wang J, He Y, Shi G and Deng G (2020) Generating fundus fluorescence angiography images from structure fundus images using generative adversarial networks. *arXiv preprint* arXiv:2006.10216.
51. Karras T, Laine S and Aila T (2019) A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 4401–4410.
52. Zhou T, Krahenbuhl P, Aubry M, Huang Q and Efros A (2016) Learning dense correspondence via 3D-guided cycle consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 117–126.
53. Bourou A, Daupin K, Dubreuil V, De Thonel A, Mezger-Lallemand V and Genovesio A (2023) Unpaired image-to-image translation with limited data to reveal subtle phenotypes. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*. pp. 1–5.

54. Wang C, Yang G, Papanastasiou G, *et al.* (2021) DiCyc: GAN-based deformation invariant cross-domain information fusion for medical image synthesis. *Information Fusion* **67**, 147–160.
55. Hervella A, Rouco J, Novo J and Ortega M (2019) Deep multimodal reconstruction of retinal images using paired or unpaired data. In *2019 International Joint Conference on Neural Networks (IJCNN)*. pp. 1–8.
56. Peng Y, Meng Z and Yang L (2023) Image-to-image translation for data augmentation on multimodal medical images, *IEICE Transactions on Information and Systems* **106**(5), 686–696.
57. Sun H, Fan R, Li C, *et al.* (2021) Imaging study of pseudo-CT synthesized from cone-beam CT based on 3D CycleGAN in radiotherapy. *Frontiers in Oncology* **11**, 603844.
58. He K, Zhang X, Ren S and Sun J (2016) Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 770–778.
59. Ronneberger O, Fischer P and Brox T (2015) 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, Lecture Notes in Computer Science 9351*, 234–241.
60. Arjovsky M, Chintala S and Bottou L (2017) Wasserstein generative adversarial networks. In *International Conference on Machine Learning (ICML)* **70** , 214–223.
61. Taigman Y, Polyak A and Wolf L (2016) Unsupervised cross-domain image generation. *arXiv preprint arXiv:1611.02200*.
62. Ward A, Bell O.H, Chu C.J, Nicholson L, Tian X, Anantrasirichai N and Achim A (2024) OCT2Confocal. https://doi.org/10.5523/bris.1hvd8aw0g6l6g28fnub18hgse4.
63. Boldison J, Khera T, Copland D, Stimpson M, Crawford G, Dick A and Nicholson L (2015) A novel pathogenic RBP-3 peptide reveals epitope spreading in persistent experimental autoimmune uveoretinitis. *Immunology* **146**(2), 301–311.
64. Caspi R (2010) A look at autoimmunity and inflammation in the eye, *The Journal of Clinical Investigation* **120**(9), 3073–3083.
65. Anantrasirichai N, Nicholson L, Morgan J, Erchova I, Mortlock K, North R, Albon J and Achim A (2014) Adaptive-weighted bilateral filtering and other pre-processing techniques for optical coherence tomography, *Computerized Medical Imaging and Graphics* **38** (6), 526–539.
66. Mellak Y, Achim A, Ward A, Nicholson L and Descombes X (2023) A machine learning framework for the quantification of experimental uveitis in murine oct, *Biomedical Optics Express.* **14**(7), 3413–3432.
67. Tian X, Anantrasirichai N, Nicholson L and Achim A (2022) Optimal transport-based graph matching for 3D retinal OCT image registration. In *2022 IEEE International Conference on Image Processing (ICIP)*. pp. 2791–2795.
68. Goceri E (2023) Medical image data augmentation: techniques, comparisons and interpretations. *Artificial Intelligence Review* **56**, 12561–12605.
69. University of Bristol (2017) *BluePebble Supercomputer*. Advanced Computing Research Centre, University of Bristol. https://www.bristol.ac.uk/acrc/high-performance-computing
70. Kingma DP and Ba J (2014) Adam: A method for stochastic optimization. *arXiv preprint* arXiv:1412.6980.
71. Rodrigues R, Lévêque L and Gutiérrez J (2022) Objective quality assessment of medical images and videos: Review and challenges. *arXiv preprint arXiv:2212.07396*.
72. Heusel M, Ramsauer H, Unterthiner T, Nessler B and Hochreiter S (2017) GANs trained by a two time-scale update rule converge to a local Nash equilibrium. *Advances in Neural Information Processing Systems (NeurIPS)* **30**, 6629–6640.
73. Bińkowski M, Sutherland DJ, Arbel M and Gretton A (2018) Demystifying MMD GANs. *arXiv preprint arXiv:1801.01401*.
74. Szegedy C, Vanhoucke V, Ioffe S, Shlens J and Wojna Z (2016) Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 2818–2826.
75. Mittal A, Soundararajan R and Bovik A (2012) Making a "completely blind" image quality analyzer, *IEEE Signal Processing Letters* **20**(3), 209–212.
76. Mittal A, Moorthy AK and Bovik A (2012) No-reference image quality assessment in the spatial domain, *IEEE Transactions on Image Processing* **21**(12), 4695–4708.


## A Appendix

**Cite this article:** Tian X, Anantrasirichai N, Nicholson L & Achim A (2024). The quest for early detection of retinal disease: 3D CycleGAN-based translation of optical coherence tomography into confocal microscopy. *Biological Imaging*, 4: e15. doi:https://doi.org/10.1017/S2633903X24000163

https://doi.org/10.1017/S2633903X24000163 Published online by Cambridge University Press

***Table 5.*** The performance of models was evaluated by DB metrics FID scores and KID scores, alongside the subjective MOS rating of each individual set. The best result is colored in red and the second best is colored in blue

| | | W Ref | | | W/O Ref | | | | | | | | | | Total | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | Metric | A2L_ D24 | A2R_ D24 | B3R_ D24 | A2R_D10 | A2R_ D14 | A2R_ D17 | B3L_ D24 | B3R_ D17 | TX12E2L_ D14 | TX12E2R_ D14 | TX12E3L_ D14 | TX13B1L_ D14 | TX13B3R_ D14 | W Ref | W/O Ref | All |
| UNSB | FID768 ↓ | 1.768 | 1.699 | 1.510 | 1.975 | 1.638 | 1.567 | 1.643 | 1.620 | 1.372 | 1.454 | 1.594 | 1.641 | 1.603 | 1.659 | 1.611 | 1.622 |
| | FID2048 ↓ | 335.571 | 336.422 | 267.572 | 434.316 | 318.284 | 324.705 | 317.714 | 271.564 | 318.818 | 274.218 | 312.400 | 243.367 | 201.276 | 313.189 | 301.666 | 304.325 |
| | KID ↓ | 0.663 | 0.643 | 0.485 | 0.968 | 0.645 | 0.707 | 0.750 | 0.585 | 0.715 | 0.537 | 0.685 | 0.516 | 0.439 | 0.597 | 0.655 | 0.641 |
| | MOS ↑ | 31.700 | 22.900 | 33.300 | 26.100 | 19.400 | 25.800 | 26.600 | 30.000 | 33.900 | 24.200 | 22.200 | 26.300 | 19.100 | 29.300 | 25.360 | 26.269 |
| 2D CycleGAN | FID768 ↓ | 1.668 | 1.509 | 1.465 | 1.553 | 1.364 | 1.640 | 1.415 | 1.349 | 1.490 | 1.250 | 1.437 | 1.564 | 1.140 | 1.547 | 1.420 | 1.449 |
| | FID2048 ↓ | 215.451 | 240.731 | 219.724 | 260.986 | 250.141 | 247.595 | 217.812 | 214.397 | 254.606 | 216.557 | 254.916 | 215.287 | 178.178 | 225.302 | 231.048 | 229.722 |
| | KID ↓ | 0.284 | 0.327 | 0.288 | 0.383 | 0.353 | 0.351 | 0.339 | 0.293 | 0.366 | 0.283 | 0.364 | 0.300 | 0.226 | 0.300 | 0.326 | 0.320 |
| | MOS ↑ | 37.000 | 42.300 | 30.700 | 31.400 | 44.200 | 42.700 | 40.400 | 48.000 | 35.900 | 42.200 | 35.400 | 49.000 | 47.100 | 36.667 | 41.630 | 40.485 |
| 3D CycleGAN-GL | FID768 ↓ | 1.633 | 1.128 | 1.081 | 1.287 | 1.247 | 1.283 | 1.075 | 1.089 | 1.304 | 1.283 | 1.162 | 1.077 | 0.892 | 1.281 | 1.170 | 1.195 |
| | FID2048 ↓ | 281.033 | 165.763 | 161.590 | 173.151 | 185.445 | 193.396 | 166.591 | 154.707 | 183.646 | 195.985 | 171.409 | 152.085 | 119.148 | 202.795 | 169.556 | 177.227 |
| | KID ↓ | 0.401 | 0.206 | 0.193 | 0.212 | 0.246 | 0.294 | 0.207 | 0.195 | 0.224 | 0.255 | 0.228 | 0.164 | 0.124 | 0.267 | 0.215 | 0.227 |
| | MOS ↑ | 46.900 | 42.100 | 62.200 | 63.300 | 39.000 | 46.300 | 50.400 | 42.800 | 46.100 | 47.100 | 47.400 | 50.800 | 62.300 | 50.400 | 49.550 | 49.746 |
| 3D CycleGAN–2 | FID768 ↓ | 0.994 | 0.840 | 0.723 | 0.969 | 0.927 | 0.953 | 0.794 | 0.793 | 0.949 | 1.023 | 0.859 | 0.880 | 0.749 | 0.852 | 0.890 | 0.881 |
| | FID2048 ↓ | 176.433 | 140.350 | 131.675 | 202.865 | 174.571 | 165.633 | 140.414 | 136.360 | 197.414 | 187.966 | 171.102 | 153.010 | 135.396 | 149.486 | 166.473 | 162.553 |
| | KID ↓ | 0.190 | 0.136 | 0.107 | 0.198 | 0.161 | 0.193 | 0.121 | 0.125 | 0.176 | 0.238 | 0.159 | 0.110 | 0.115 | 0.144 | 0.160 | 0.156 |
| | MOS ↑ | 56.900 | 65.300 | 39.700 | 37.500 | 62.900 | 45.700 | 55.900 | 57.600 | 55.400 | 48.200 | 52.900 | 62.300 | 50.200 | 53.967 | 52.860 | 53.115 |
| 3D CycleGAN–3 | FID768 ↓ | 0.821 | 0.730 | 0.747 | 0.883 | 0.793 | 0.763 | 0.701 | 0.715 | 0.822 | 0.835 | 0.825 | 0.775 | 0.685 | 0.766 | 0.780 | 0.777 |
| | FID2048 ↓ | 177.323 | 139.419 | 147.526 | 172.697 | 161.605 | 134.269 | 133.621 | 165.109 | 166.120 | 169.130 | 164.375 | 150.130 | 134.827 | 154.756 | 155.188 | 155.089 |
| | KID ↓ | 0.187 | 0.127 | 0.144 | 0.179 | 0.167 | 0.122 | 0.126 | 0.172 | 0.180 | 0.185 | 0.182 | 0.136 | 0.114 | 0.153 | 0.156 | 0.155 |
| | MOS ↑ | 54.500 | 65.900 | 50.200 | 57.700 | 56.800 | 58.900 | 60.000 | 50.600 | 56.600 | 56.100 | 40.300 | 55.500 | 71.000 | 56.867 | 56.350 | 56.469 |

*Note:* FID768 and FID2048 refer to the Fréchet Inception Distance computed with 768 and 2048 features, respectively. KID refers to the Kernel Inception Distance. Both FID and KID indicate better performance with lower scores. Mean Opinion Score (MOS) rates the subjective quality of images with higher scores reflecting better quality.