Article:
Liao, Jinpeng, Zhang, Tianyu, Li, Chunhui et al. (2023) U-shaped fusion convolutional transformer based workflow for fast optical coherence tomography angiography generation in lips. Biomedical Optics Express. pp. 5583-5601. ISSN: 2156-7085

https://doi.org/10.1364/BOE.502085

Research Article | Vol. 14, No. 11 / 1 Nov 2023 / Biomedical Optics Express 5583
Biomedical Optics EXPRESS

U-shaped fusion convolutional transformer based workflow for fast optical coherence tomography angiography generation in lips

JINPENG LIAO, TIANYU ZHANG, CHUNHUI LI,* AND ZHIHONG HUANG

School of Science and Engineering, University of Dundee, DD1 4HN, Scotland, United Kingdom
*c.li@dundee.ac.uk

Abstract: Oral disorders, including oral cancer, pose substantial diagnostic challenges due to late-stage diagnosis, invasive biopsy procedures, and the limitations of existing non-invasive imaging techniques. Optical coherence tomography angiography (OCTA) shows potential in delivering non-invasive, real-time, high-resolution vasculature images. However, the quality of OCTA images are often compromised due to motion artifacts and noise, necessitating more robust and reliable image reconstruction approaches. To address these issues, we propose a novel model, a U-shaped fusion convolutional transformer (UFCT), for the reconstruction of high-quality, low-noise OCTA images from two-repeated OCT scans. UFCT integrates the strengths of convolutional neural networks (CNNs) and transformers, proficiently capturing both local and global image features. According to the qualitative and quantitative analysis in normal and pathological conditions, the performance of the proposed pipeline outperforms that of the traditional OCTA generation methods when only two repeated B-scans are performed. We further provide a comparative study with various CNN and transformer models and conduct ablation studies to validate the effectiveness of our proposed strategies. Based on the results, the UFCT model holds the potential to significantly enhance clinical workflow in oral medicine by facilitating early detection, reducing the need for invasive procedures, and improving overall patient outcomes.

1. Introduction

Oral cancer was highly attended by global public health, especially by dentists [1], which was a malignant tumor that appeared on lips and intra-oral. Lips, as a highly visible and cosmetically critical anatomical structure, can significantly affect an individual's appearance when altered. There exists an extensive range of mucocutaneous disorders affecting the lips, such as ulcers, salivary gland stones, and chronic nibbling of the lips [2–4]. Notably, lip ulcers are relatively common in oral medicine. Although most lip ulcers are benign and often resolve quickly without clinical intervention, some can signify potentially lethal systemic or neoplastic disorders stemming from infection, auto-immunity, trauma, or malignancy [1,5,6]. Given that the five-year survival rate for oral cancer is lower than 60%—primarily due to late-stage diagnosis [7] -early histopathological examination of lip ulcers is of vital importance. While biopsy remains the gold standard for diagnosing lip diseases, its invasive nature presents challenges, as it often results in bleeding, pain, and technical difficulties when applied for preventative health screenings [8,9]. Consequently, there is a pressing need for a rapid, non-invasive, and reproducible imaging modality to evaluate erosive or ulcerative areas. Several techniques, such as intraoral scanners [10], multispectral digital microscopes [11], and fluorescence visualization detection [12], have been explored for lip disease diagnosis. Nevertheless, these approaches cannot provide in-depth information on lip tissues, leading to potential misdiagnoses and reducing early oral disease detection chances.

#502085 https://doi.org/10.1364/BOE.502085
Journal © 2023 Received 1 Aug 2023; revised 15 Sep 2023; accepted 23 Sep 2023; published 5 Oct 2023

Biomedical Optics EXPRESS

Optical coherence tomography (OCT) offers a non-invasive imaging technique that is real-time, high-resolution (e.g., ~10 μm), and has a penetration depth of 1~2 mm [13]. The micro-anatomical information of oral tissues provided by OCT has proven to be of clinical importance in various oral diseases [9,14–17]. However, research indicates that OCT struggles to differentiate between oral cancer lesions based on structural images [18]. As an extended function of OCT, OCT-angiography (OCTA) can provide vasculature images at the capillary-level resolution, aiding in disease identification by examining vascular structure and perfusion in oral lip tissues, thereby contributing to the diagnosis of diverse lip conditions [19–22].

The generation of OCTA images hinges on suppressing static signals from tissue and extracting moving scattering signals from red blood cells in repeatedly scanned OCT signals taken from the same location at different time points [23,24]. Consequently, OCTA image quality is tightly linked to the number of repeated scans: more scans equate to better signal-to-noise ratio (SNR) and connectivity in the OCTA image. However, excessive scans can increase data acquisition time, particularly in lip OCTA imaging, where unintended motion artifacts due to patient and scanning probe movement are more likely, underscoring the need for an optimized OCTA scanning process that balances data acquisition speed and image quality.

Deep-learning-based strategies using convolution neural network (CNN) models have shown promise in reconstructing high-quality OCTA images based on two- or four-repeated OCT scans [25–32], even in dermatology and ophthalmology. However, these models need retraining to learn the specific vasculature features of lips and ulcers. More importantly, their data acquisition relies on a fixed sample lens, minimizing motion but adopting invasive collection methods [27–29]-an approach unsuitable for this study.

While CNNs excel in extracting local features based on local connections and sliding window mechanisms, they struggle to capture long-term information [33]. In contrast, transformers can readily capture global information by capturing distant features [34]. Notably, transformers are convolution-free neural networks that use a self-attention mechanism to extract intrinsic features, a feature that holds potential for broad application. For instance, Vision Transformer (ViT) [35] applies a pure transformer layer directly to image patches to denoise natural images. Additionally, transformer layers have shown impressive sensitivity and efficiency in image reconstruction, classification, and segmentation [36–40]. However, the training of transformers requires extensive datasets (e.g., ImageNet2K, JFT3M), and there is currently a shortage of datasets for high-quality lip OCTA image reconstruction.

To combine the strengths of CNNs and transformers, Wu et al. [41] introduced convolution operations into the transformer layer, enabling the model to extract spatial relationships between image patches. Wang et al. [42] proposed a fusion neural network design to augment feature extraction from two different models. Inspired by these works, we introduced a novel model named U-shaped Fusion Convolutional Transformer (UFCT) to reconstruct high-quality and low-noise OCTA images based on two-repeated OCT scans. This model integrates convolution operations in the sequence generation to better capture spatial information between image patches and employs a fusion network for a multi-scale feature representation strategy, capturing a more comprehensive set of features and enhancing generalization and performance.

Our contributions are as follows: (1) we proposed a deep-learning-based pipeline for lip OCTA to enhance the clinical workflow in oral medicine using only two repeated OCT scans; (2) we proposed a UFCT model combining vision transformer and convolution operation advantages, offering superior OCTA image reconstruction; (3) we conducted a comparative study of various CNN and transformer models for lip OCTA image reconstruction; (4) we carried out a series of ablation studies to verify our proposed training strategies, implementation details, and model architecture's effectiveness.

Biomedical Optics EXPRESS

2. System and data pre-processing

2.1. Introduction of the SSOCT system

Human lips OCT data were collected in-vivo using a laboratory-built swept-source optical coherence tomography (SSOCT) system (as illustrated in Fig. 1). More details of the system were described in [43]. The study was approved by the School of Science and Engineering Research Ethics Committee (SSEREC) of University of Dundee, which also conformed to the tenets of the Declaration of Helsinki. To develop a comprehensive assessment of the proposed model, both the normal and abnormal lips sites were included in this study. In the package of healthy lip sites, 22 health lip OCT data (#1 - #22) were taken from 22 healthy subjects aged between 20 and 30 years old, none of whom had any oral diseases. Three diseased subjects (#23 - #25) were collected data in this study, including one case (#23) with salivary gland stones and two cases (#24 - #25) with ulcers. Specifically, #23 consists of two OCT data collected in two continuous days, #24 consists of three OCT data collected in three continuous days, and #25 consists of five OCT data in five continuous days.

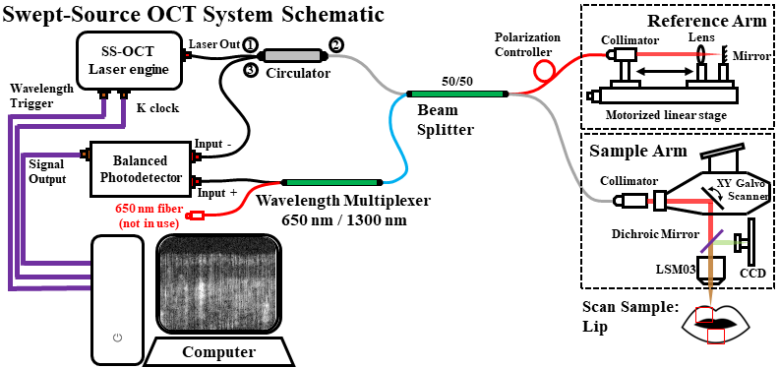

Fig. 1. System Schematic of the Swept-Source OCT System used in this study. The swept-source Laser used in this system has a wavelength of 1310 nm and bandwidth of 100 nm, and a 200 kHz swept rate (SL132120, Thorlabs Inc.). The lens used in the sample arm is LSM03 (Thorlabs Inc.) with a focal length of 35 mm. The theoretical axial resolution is 7.4 µm in air.

2.2. Data acquisition and pre-processing

OCTA scan from normal lips was used to build the input low-quality OCTA images (I_{LQ}) and counterpart ground-truth high-quality OCTA images (I_{HQ}) of training datasets due to their easy accessibility. As for the imaging protocol in normal lips, one OCTA scan consists of 600 B-scans (y-transverse axis), and each B-scan is composed of 600 A-scans (x-transverse axis). The penetration depth of our SSOCT system is approximately 1.5 mm (z-axial axis). The field of view (FOV) of OCTA imaging was set as 5.16 mm × 5.16 mm, and the spatial sampling interval is ~8.6 µm in this study. The number of repeated OCT scans for OCTA image generation is four at each fixed position. The scanning time for each completed OCTA scan is ~ 8 s. To suppress motion-induced artifacts during the OCT scan, an image registration toolbox Elastix [44] was performed in each repeated B-scan before vascular extraction. Eigen-decomposition (ED)-OCTA algorithm was utilized to generate OCTA images due to its good performance in suppressing static tissue while preserving vascular signals [24]. The flowchart for generating datasets for training and validation is shown in Fig. 2. The ground-truth high-quality OCTA images ($\boldsymbol{GT\_ED}_4$) were generated based on the four-repeated scans with ED-OCTA, and the input low-quality OCTA images ($\boldsymbol{M\_ED}_2$) were based on two-repeated scans with ED-OCTA.

Biomedical Optics EXPRESS

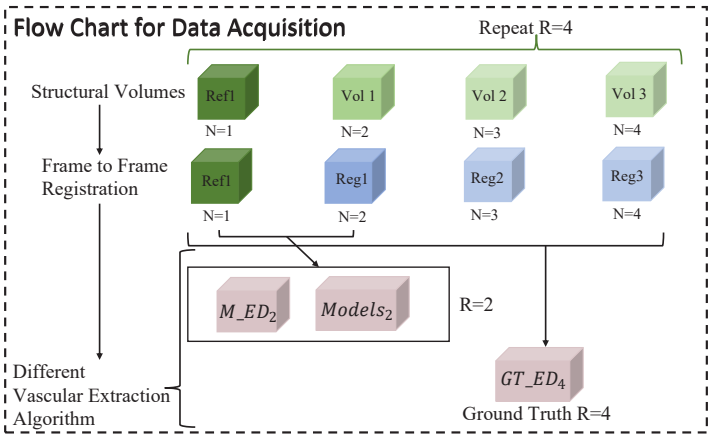

Fig. 2. Flow chart of the scanning and processing strategy to create high-quality ground-truth OCTA results using 4-repeated scan (GT_ED_4), and other strategies using 2-repeated scans (M_ED_2 and $Models_2$). M_ED_2 is used as the input low-quality OCTA images. $Models_2$ are reconstructed high-quality OCTA images from different neural networks.

The data package in normal lips consists of 22 OCT raw data (#1 - #22), including 13200 cross-section OCTA images (22 data $\times$ 600 B-scans). A select box with a size of 200×200 (x-transverse $\times$ z-axial) was used to crop the cross-section images due to the limited memory in the graphics card in this study. Finally, a total of 39600 images were generated for ground-truth, input, and reference data. Among them, 28800 pairs of images (#1 - #16 subjects) were selected as the input and ground truth of the training dataset. The rest of the 10800 pairs of images (#17 - #22) were used as a validation dataset on the health lip. Regarding the lip data with diseases, 10 OCT raw data (#23 $\times$ 2, #24 $\times$ 3, #25 $\times$ 5) are selected to test the feasibility of the proposed fast OCTA imaging pipeline. The scanning protocol and data pre-processing methods are the same as the normal lip data. In total, 18000 cross-section OCTA images were used as a validation dataset that represents disease lip.

3. Model Training

3.1. U-shaped fusion convolutional transformer

Figure 3 outlines the architecture of the proposed U-shaped Fusion Convolutional Transformer (UFCT). The UFCT encompasses an encoder, a decoder, a Simple Fusion Network (SFN), and a bottleneck. In the encoder block, features from the input low-quality OCTA image are progressively extracted through four downsample operations, including one patch extraction layer and three convolution layers (blue box in Fig. 3). The first downsample operation non-overlappingly splits the input image (of dimensions H x W) into patches of size 2×2, outputting image patches of dimensions H/2 x W/2 $\times$ 4. A position embedding layer with C hidden units is applied to these image patches, transforming their shape to H/2 x W/2 x C. The output from this layer is passed into the Modified Convolutional Transformer (MCT) layer, maintaining the shape of the feature maps.

To fuse features from the first layer of SFN, a concatenate layer is employed, concatenating the features from the MCT layer and SFN, yielding an output shape of H/2 x W/2 $\times$ 2C. To minimize computational cost while efficiently extracting features from the last MCT layer and SFN, a convolution layer with a filter size of C and a kernel size of 1 (depicted by the green box in Fig. 3) reduces the channels of the feature map. For the additional three downsample operations, the patch extraction layer and position embedding layer are replaced by a convolution layer with a

Biomedical Optics EXPRESS

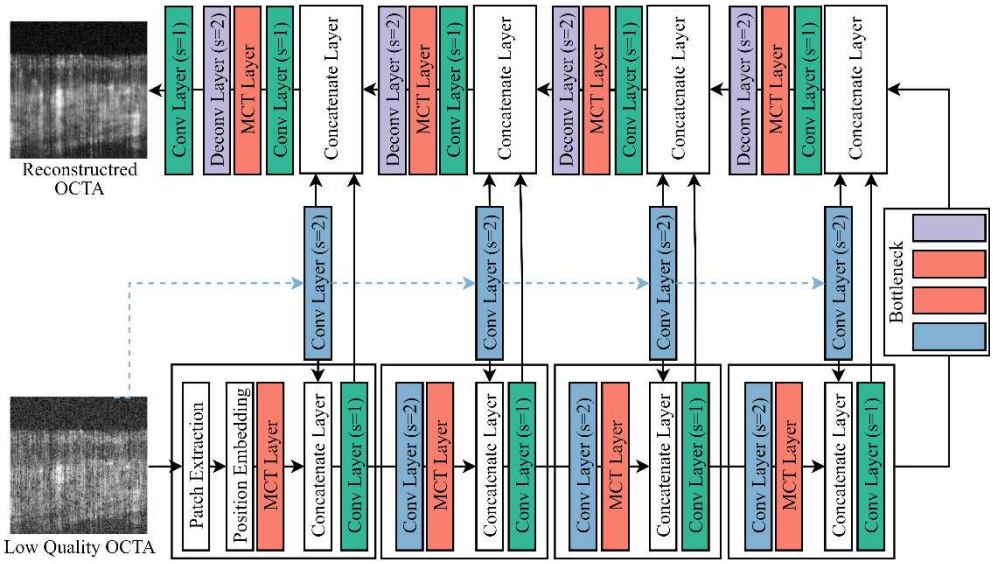

Fig. 3. The architecture of the proposed U-shaped Fusion Convolutional Transformer. The s is the strides of the convolution layer. All convolution layers in the blue box have a kernel size of 3, and all convolution layers in the green box have a kernel size of 1. Conv: convolution; Deconv: deconvolution or transpose convolution.

stride of 2, halving the spatial dimensions while doubling the channel depth of the input. The bottleneck comprises a downsample layer, two MCT layers, and a deconvolution (also known as transpose convolution) layer for feature map upsampling. In the upsample block, a concatenate layer first combines features from the preceding downsample and upsample blocks, and SFN. A convolution layer with a kernel size of 1 is used to reduce the channels of the feature map from the concatenate layer, which is fed into an MCT layer.

3.1.1. Modified convolutional transformer layer

In Fig. 4, a modified convolutional transformer (MCT) layer is introduced, and it contains a convolutional projection (CP) layer (Fig. 4 (A)), a multi-head attention (MHA) layer (Fig. 4 (B)), and a feed-forward network (FFN). The conventional convolutional transformer demands significant computational resources as it employs convolution layers to project the Query (Q), Key (K), and Value (V) sequences, as indicated by Wu et al. [41]. To reduce the computation cost of the convolutional transformer, we introduced a convolution layer with a stride of 2 into the MCT layer to reduce the size of the feature maps, thereby reducing the computation cost. A deconvolution layer with a stride of 2 is used to upsample the output feature maps to ensure the output shape is the same as the input shape. Taking X_1 as the input of the MCT layer, the processing of the MCT layer can be written as follows:

$$\hat{y}_1 = \mathrm{MHA}(\mathrm{CP}(\mathrm{LN}(\mathrm{Conv}_{s2}(X_1)))) + \mathrm{Conv}_{s2}(X_1) \tag{1}$$

$$Y = \mathrm{Deconv}_{s2}(\mathrm{FFN}(\mathrm{LN}(\hat{y}_1)) + \hat{y}_1) \tag{2}$$

where Conv_{s2} is a convolution layer with a stride of 2, and Deconv_{s2} is a deconvolution layer with a stride of 2. Y is the output of the MCT layer. FFN contains two dense layers with a GeLU activation layer. CP is the convolutional projection layer, MHA is the multi-head attention layer, and LN is layer normalization. Assume the shape of input (X_1) is H + W + C, the processing of $\mathrm{Conv}(X_1)$ will firstly reduce the shape of the X_1 from H + W + C to H/2 + W/2 + C. After

Biomedical Optics EXPRESS

processing by LN, a convolutional projection ($\mathrm{Conv_{CP}}$) is applied to obtain the Q, K, and V sequences for multi-heads attention as:

$$\mathrm{Sequences_{Q,K,V}} = \mathrm{Flatten}(\mathrm{Conv_{CP}}(\mathrm{Conv_{s2}}(\mathrm{X_1}))) \tag{3}$$

The filter size of $\mathrm{Conv_{CP}}$ is the same as the channel size (C) of the input $\mathrm{X_1}$, with a stride of 1 and a kernel size of 3. After the flatten processing, the shape of each sequence (Q, K, V) is HW/4 + C. A head split operation is then applied to each sequence to reshape the sequence from HW/4 + C to M + HW/4 + C/M, where M is the number of heads for multi-head attention. Then, a standard self-attention layer (Fig. 4 (C)) is operated on each separated head sequence, and the processing workflow to obtain an attention score can be written as:

$$\textit{Attention Score}\ (Q,\ K, V) = \mathrm{Softmax}\left(\frac{\mathrm{QK^T}}{\sqrt{d}}\right)\mathrm{V} \tag{4}$$

where d is a numerical value of $\sqrt{dimension\ of\ Q}$ and T is the transposing operation of sequence K. After the attention operation, a reshape layer is used to combine all separated heads into a matrix with a shape of HW/4 + C. Finally, a fully connected layer with a hidden unit of C is used to project the matrix as the output of multi-head attention layer.

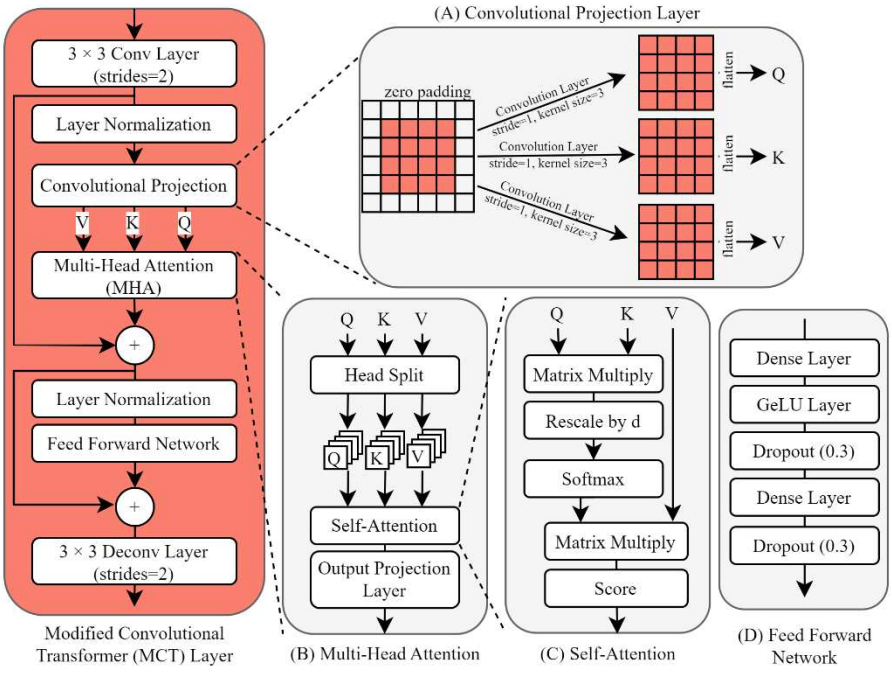

Fig. 4. The schematic of the modified convolutional transformer (MCT) layer. (A)-(D) are components of the MCT layer, where (A) is the convolutional projection layer, and (B) is the multi-head attention layer. (C) is the self-attention mechanism. (D) is the architecture of the feed-forward network. Conv: convolution; Deconv: deconvolution or transpose convolution.

3.1.2. Loss function

In neural network training, the combination of VGG19-based content loss and mean-square-error (MSE) loss function was utilized as a similarity metric between the high-quality ground-truth and reconstructed images by neural networks, as previous work proposed [32]. The MSE loss has been

Biomedical Optics EXPRESS

proven that has the capability to reduce noise and increase image contrast of the reconstructed OCTA images, according to [27]. Moreover, the utilization of the VGG19-based content loss has proven that can improve the image reconstruction performance of the networks in terms of the connectivity and texture details of the vasculature images [32]. The loss function of the MSE loss can be formulated as (5):

$$L_{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{5}$$

where y is the ground-truth OCTA image, and $\hat{y}$ is the reconstructed OCTA image from the network. N is the total number of pixels in y and $\hat{y}$. A VGG19-based content loss is then applied to improve the texture details of vasculature based on the calculation of the pixel-by-pixel difference between the feature maps from VGG19. And the formulation can be written as (6):

$$L_{content} = \frac{1}{N}\sum_{i=1}^{N}(G(y_i) - G(\hat{y}_i))^2 \tag{6}$$

where G is the ImageNet2K pretrained VGG19 network for feature map predictions [45]. The output layer of G is set as the fourth layer of block 5 in the VGG19 network since it can provide relatively stable training and the highest OCTA image reconstruction performance [32]. The combined loss for neural networks training is then shown as (7):

$$L_c = \alpha * L_{content} + \beta * L_{MSE} \tag{7}$$

where α and β are parameters to control the weights of the loss function.

3.2. Definition of deep-learning-based OCTA image reconstruction pipeline

The OCTA image reconstruction pipeline consists of training and testing stages, as illustrated in Fig. 5. In the training stage, a pre-processing strategy (blue box in Fig. 5, and details in Fig. 2) and an end-to-end OCTA image reconstruction process based on the model were performed. In the training stage (i.e., green box in Fig. 5), the low-quality OCTA images, which are used as

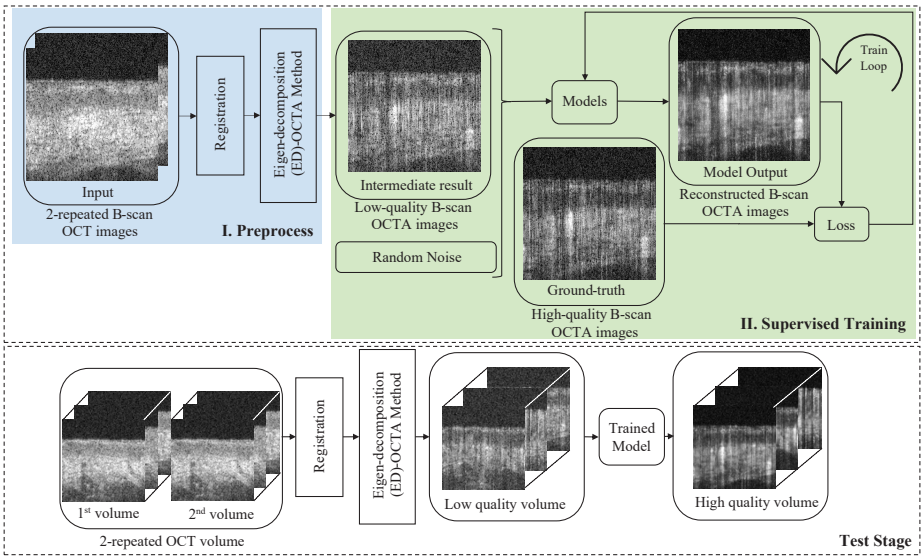

Fig. 5. The Pipeline of OCTA Image Reconstruction.

Biomedical Optics EXPRESS

the model input, are first generated based on the two-repeated OCT scan with the ED-OCTA algorithm. Random noise is utilized to simulate the shot noise from the photo-balance detector and system noise during the data acquisition. Supervised training is then applied to train the different models based on the calculated loss between the high-quality OCTA images and reconstructed OCTA images. In the test stage, the input is the two-repeated OCT volumes. After applying the volume registration and ED-OCTA method mentioned in Section 2.2, a series of OCTA images with low image quality in a cross-sectional view is obtained. A reconstructed OCTA image with high quality is acquired by the trained model. Finally, at the end-task stage, the en-face vasculature was generated by the maximum intensity projection (MIP) method for visual comparison in terms of vasculature texture reconstruction.

4. Evaluation

4.1. Evaluation metrics

To quantitatively compare the performance of different algorithms, including conventional OCTA algorithms (i.e., ED-OCTA) and deep-learning-based methods, two performance metrics are computed in the experiments: peak-signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM) [46]. We used the PSNR metric to gauge the model in terms of noise reduction. Since PSNR is described in decibels (dB) and is calculated as the square of the difference between the model output and the ground truth image:

$$\mathrm{PSNR} = 20 * \log_{20}\left(\frac{I_{max}}{\sqrt{MSE(y, \hat{y}_i)}}\right) \tag{8}$$

where MSE has the same formula as (5), I_{max} is the maximum value in image data y and $\hat{y}_i$. The SSIM evaluates image quality in terms of structural similarity. Higher SSIMs show a better structural similarity of model outputs to ground truth images.

$$SSIM = \frac{(2\mu_{GT}\mu_M + k1)(2\sigma_{cov} + k2)}{(\mu^2{}_{GT} + \mu^2{}_M + k1)(\sigma^2{}_{GT} + \sigma^2{}_M + k2)} \tag{9}$$

Here, μ_{GT} and (σ_{GT}) and μ_M and (σ_M) are the mean (variance) of the underlying truth and the output image using a different strategy, respectively; σ_{cov} shows the covariance between these two data. $k1$ and $k2$ are used to stabilize the division with a weak denominator.

4.2. Comparison with other neural networks

In order to verify the effectiveness of the proposed method in this study, we select six models for comparison, including 3 CNN models: DnCNN [47], SRResNet [48], and U-Net [49], and 3 Transformer-related models: TransUNet [40], Swin-UNet [38], and Lightweight U-shape Swin Transformer (LUSwin-T) [43]. The training details and training strategies of the compared-used models are the same as the proposed implementation details mentioned below (Section 5.1), to reduce the influence of the training details and graphics card utilization.

4.3. Ablation study

As mentioned above, our proposed UFCT consists of feature fusion and a series of MCT layers, and the neural networks are trained based on supervised learning with the proposed combined loss function (Eq. (7)). We conduct ablation experiments on feature fusion methods, reduced size of UFCT, and different training strategies to verify that the designed modules are effective in terms of improving the OCTA image reconstruction performance.

Effect of feature fusion method: The proposed feature fusion method in this study is to use the concatenate layer to fuse the features from the SFN with the encoder and decoder blocks.

Biomedical Optics EXPRESS

To explore the effects of the proposed feature fusion method, we conducted the experiments of UFCT with different feature fusion methods, including (1) using elementwise adding to replace the concatenate layer (marked as Exp.1); (2) using concatenate fusion methods in encoder only (marked as Exp.2); (3) do not use SFN (marked as Exp.3).

Effect of reducing the size of UFCT: In this experiment, the architecture of UFCT is constant as proposed in Fig. 3; however, different from the implementation details as proposed, the filter size of the MCT layer and downsample convolution layer is increased from 32, 64, 256, 256 in encoder block, while reversely decreased in decoder block. The head number of the MCT layer is increased from 2, 2, 4, and 4 in the encoder block, while reversely decreased in the decoder block. The MCT layer in the bottleneck has a filter size of 256 and a head number of 4. We define the name of the model as UFCT-tiny under this implementation details.

Effect of training strategy: Adversarial training is widely used in deep-learning-based image restoration tasks since it can provide an additional adversarial loss from the discriminator, and serval studies have shown that adversarial loss can improve the image restoration performance in terms of visualization and quantitative, according to [28,36,48,50]. Therefore, we explore the performance of different models that are trained by the adversarial loss [51], MSE loss only, and combined loss to find the best-match training strategy for the OCTA image reconstruction task. The definition and parameter setting of the adversarial loss in this ablation study is the same as the adversarial loss used in [48]. Equation (10) is the adversarial loss (L_D) for the discriminator model (D), and Eq. (11) is the adversarial loss (L_G) for the generator model (G) (e.g., UFCT in this study). The discriminator model has the same architecture as the VGG16 [45], designed to output a probability with a shape of 1. This probability discerns whether the input image to the discriminator is a ground-truth OCTA image (real) or a reconstructed OCTA image produced by the generator model (fake).

$$L_D(I_{HQ}, I_R) = E_{I_{HQ}\sim P_{data}(I_{HQ})}(\log(D(I_{HQ}))) + E_{I_R\sim P_{data}(I_R)}(\log(1 - D(I_R))) \tag{10}$$

$$L_G(I_{HQ}, I_R) = E_{I_R\sim P_{data}(I_R)}(1 - D(I_R)) \tag{11}$$

where I_{HQ} is the high-quality ground-truth OCTA images, and I_R is the reconstructed OCTA images by model G. $D(I_{HQ})$ and $D(I_R)$ are the probabilities produced by the discriminator model. $E_{I_{HQ}}$ and E_{I_R} are the expected value over the data distribution $P_{data}(I_{HQ})$ and $P_{data}(I_R)$. Finally, the adversarial loss for the model training in this ablation study can be written as (12):

$$L_{Adversarial}(I_{HQ}, I_R) = 0.001 \times L_G(I_{HQ}, I_R) + L_c(I_{HQ}, I_R) \tag{12}$$

5. Experiment and results

5.1. Implementation details

The neural networks used in this study were built and trained based on TensorFlow 2.9.0 [52]. The Adam optimizer was used in the experiment to optimize the networks, with an initial learning rate of 1e-4 and a momentum of 0.8 [53]. α=0.01 and β=1 were chosen as the weighting parameters for the combined loss function in Eq. (7). The number of training epochs was set to 600, and the batch size was set to 32. An NVIDIA GeForce 3090 graphics card with 24 GB memory was used to train the neural networks. An early stopping strategy was used to prevent overfitting when the validation loss is not decreased under 40 epochs.

Regarding the initialization of the proposed neural network, the filter size of the MCT layer and downsample convolution layer is increased from 64, 128, 256, and 512, while reversely decreased in the decoder block. The head number of the MCT layer also increased from 2, 4, 8, and 8. The filter size of the bottleneck is 512, and the head number of the MCT layer is 8. The head number of the MCT layer decreased from 8, 8, 4, and 2. The filter size of the last convolution layer is 1 which is the same as the channel of the input low-quality OCTA image. In

Biomedical Optics EXPRESS

terms of the convolution layer in SFN, the filter size setting is similar to downsample block (i.e., from 64 to 512), the stride is 2, and the kernel size is set as 3.

5.2. Quantitative comparison of different methods

Table 1 is the quantitative comparison between different models, and Table 2 represents the computational cost of each model in terms of floating point operations (FLOPs), parameters, and latency time for OCTA image reconstruction. It should be noticed that the quantitative comparisons are based on the cross-sectional images, and the details of data pre-processing are available in Section 2.2. Compared with the input low-quality OCTA images, all trainable model-based methods have achieved a better performance in PSNR and SSIM. In quantitative comparison results, the transformer-type models (i.e., TransUNet, LUSwin-T, Swin-UNet, and UFCT) have better SSIM and PSNR performances than the CNN models (i.e., DnCNN, SRResNet, and U-Net). Among the transformer-type models, the proposed UFCT has the best performance in terms of SSIM performance for the normal set (SSIM: 0.608), salivary gland stone set (SSIM: 0.406), lip ulcer 1 (SSIM: 0.320), and lip ulcer 2 (SSIM: 0.325). Regarding the PSNR performance, Swin-UNet has the best performance in normal set (PSNR: 20.71), TransUNet has the best performance in salivary gland stone set (PSNR: 18.75), UFCT has the best performance in lip ulcer 1 set (PSNR: 18.36), and LUSwin-T has the best performance in lip ulcer 2 set (PSNR: 18.44). Nevertheless, all CNN models have a lower latency time than all transformer-type models, according to Table 2. In Swin-based transformer models, the LUSwin-T (FLOPs: 3.93 G; Parameters: 11.92 M) and Swin-UNet (FLOPs: 16.12 G; Parameters: 50.28 M) have lower FLOPs and parameters but require more time for OCTA image reconstruction. We suggest that it is because of large amounts of reshaping operation in Swin-transformer operation, also according to [54].

Table 1. Quantitative Comparison (mean ± standard deviation) with State-of-the-art Methods for OCTA Image Reconstruction based on Two-Repeated OCT Scans[a]

Method	Normal Set (Cases #17-#22)		Salivary Gland Stone (Case #23)		Lip Ulcer 1 (Case #24)		Lip Ulcer 2 (Case #25)	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Input	16.85 ± 0.65	0.506 ± 0.039	16.28 ± 0.73	0.343 ± 0.034	15.82 ± 0.55	0.284 ± 0.025	15.88 ± 0.60	0.289 ± 0.023
DnCNN [47]	19.96 ± 0.87	0.575 ± 0.041	18.46 ± 0.72	0.384 ± 0.039	17.99 ± 0.60	0.308 ± 0.029	18.19 ± 0.65	0.319 ± 0.028
SRResNet [48]	20.11 ± 0.84	0.572 ± 0.041	18.58 ± 0.69	0.384 ± 0.039	18.13 ± 0.59	0.308 ± 0.029	18.39 ± 0.62	0.318 ± 0.028
UNet [49]	20.09 ± 0.94	0.566 ± 0.049	18.55 ± 0.63	0.376 ± 0.038	17.83 ± 0.58	0.298 ± 0.028	18.11 ± 0.59	0.308 ± 0.028
TransUNet [40]	20.45 ± 0.95	0.580 ± 0.050	**18.75 ± 0.75**	0.393 ± 0.040	18.14 ± 0.59	0.307 ± 0.032	18.39 ± 0.61	0.318 ± 0.031
LUSwin-T [43]	20.62 ± 0.96	0.597 ± 0.045	18.74 ± 0.92	0.397 ± 0.040	18.23 ± 0.52	0.314 ± 0.030	**18.44 ± 0.60**	0.322 ± 0.029
Swin-UNet [38]	**20.71 ± 1.01**	0.599 ± 0.046	18.71 ± 0.86	0.395 ± 0.039	18.10 ± 0.50	0.312 ± 0.030	18.32 ± 0.53	0.322 ± 0.028
UFCT (ours)	20.62 ± 0.74	**0.608 ± 0.038**	18.64 ± 0.99	**0.406 ± 0.042**	**18.36 ± 0.58**	**0.320 ± 0.032**	18.39 ± 0.64	**0.325 ± 0.030**

[a]*The results marked with bold in SSIM and PSNR mean the best mean performance among different methods.

5.3. Visual comparison of different methods

To better visualize and evaluate the vasculature details of OCTA reconstruction results from different methods, the maximum intensity en-face projection (MIP) method was used in this section, which was mentioned in Fig. 5. It should be noted that the quantitative comparison between the enface OCTA images that generated from different methods are calculated based on the enface OCTA images presents in figure. In appendix, Figs. 10 and 11 are additional visual results of two different lip ulcers cases.

Figure 6 is the visual comparison based on independent data from a normal set. Compared with input low-quality OCTA image (B), the reconstructed results from deep-learning-based models have better quantitative (i.e., PSNR and SSIM) and visual performances. Among

Biomedical Optics EXPRESS

Table 2. Computational Cost of Different Methods (Latency Time is evaluated on RTX3050)

Method	Type	FLOPs (G)	Parameters (M)	Latency Time
DnCNN	CNN	40.92	0.557	0.10 s/frame
SRResNet	CNN	41.68	0.567	0.10 s/frame
UNet	CNN	59.88	34.56	0.11 s/frame
TransUNet	Transformer	23.01	52.35	0.17 s/frame
LUSwin-T	Swin-Transformer	3.93	11.92	0.30 s/frame
Swin-UNet	Swin-Transformer	16.12	50.28	0.51 s/frame
UFCT	Conv-Transformer	25.15	82.73	0.29 s/frame

them, the proposed UFCT has achieved the best SSIM (0.722) and PSNR (21.03) performance. Furthermore, the reconstructed results by models (i.e., (C)-(I)) can provide a better vasculature connection and higher contrast (e.g., red and blue zoom in box) than the ground-truth image (A) and input low-quality image (B). In addition, the results from deep-learning-based methods efficiently reduce the motion artifacts and reconstruct the hidden vasculature textures, compared with the input low-quality image (B).

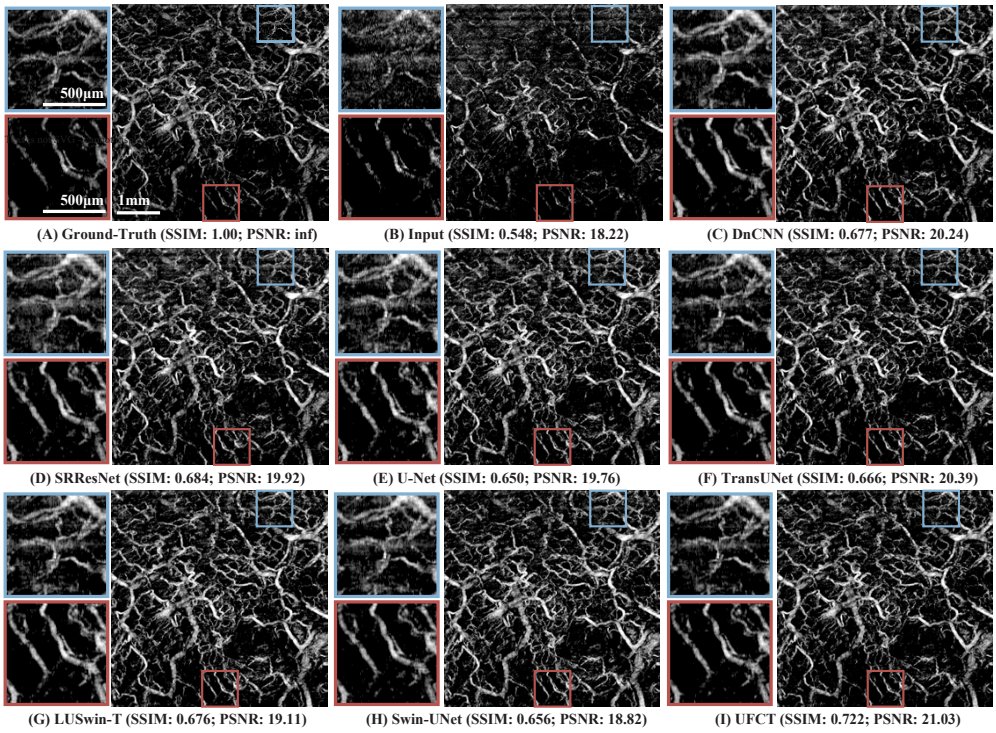

Fig. 6. Visual comparison of the OCTA image reconstruction based on normal lip OCT data. (A) is ground-truth image obtained by ED-OCTA with four-repeated OCT scan. (B) is input low-quality OCTA image obtained by ED-OCTA with two-repeated OCT scans. (C)-(I) are reconstructed OCTA images by different trained models. The white scale bar is used in (A).

Figure 7 is a visual comparison result based on a disease subject (salivary gland stone in lip). Compared with the input low-quality image (B), the results from the model have higher SSIM and

Biomedical Optics EXPRESS

PSNR performance and visual performance. The proposed UFCT has the highest PSNR (18.31) and SSIM (0.674). Moreover, the reconstrued results by models (i.e., (C)-(I)) can provide better vasculature texture details and connection than input low-quality OCTA images (B). However, in terms of visual comparison, all results from neural networks have lower contrast and vasculature details than the ground-truth OCTA image (A), which is different from the visual comparison in Fig. 6.

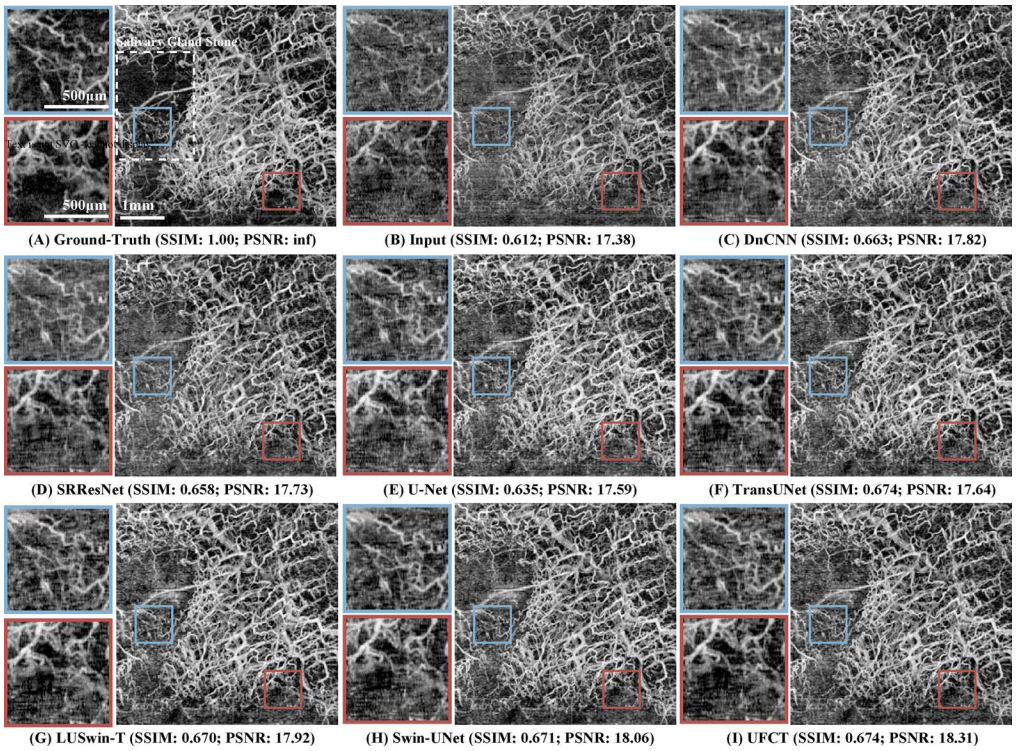

Fig. 7. Visual comparison of the OCTA image reconstruction based on lip OCT data with salivary gland stone. (A) is a ground-truth image obtained by the ED-OCTA algorithm with a four-repeated OCT scan. (B) is input low-quality OCTA image obtained by ED-OCTA algorithm with two-repeated OCT scan. (C)-(I) are reconstructed OCTA images by different deep-learning-based models, including (C) DnCNN, (D) SRResNet, (E) U-Net, (F) TransUNet, (G) LUSwin-T, (H) Swin-UNet, and (I) UFCT. The white scale bar is used in (A).

5.4. Comparison in ablation study

5.4.1. Effect of the feature fusion method

Figure 8 is the error bar of the UFCT under different feature fusion methods, and the details of the ablation study setup are mentioned in Section 4.3. Among them, the proposed UFCT (i.e., control) architecture can provide the best SSIM and PSNR results in four various validation datasets. Table **3** is the computational cost of the different UFCTs under different feature fusion methods. Although the proposed UFCT has a similar latency time as other feature fusion methods set up, it has the highest FLOPs (25.15) and parameters (82.73 M).

Biomedical Optics EXPRESS

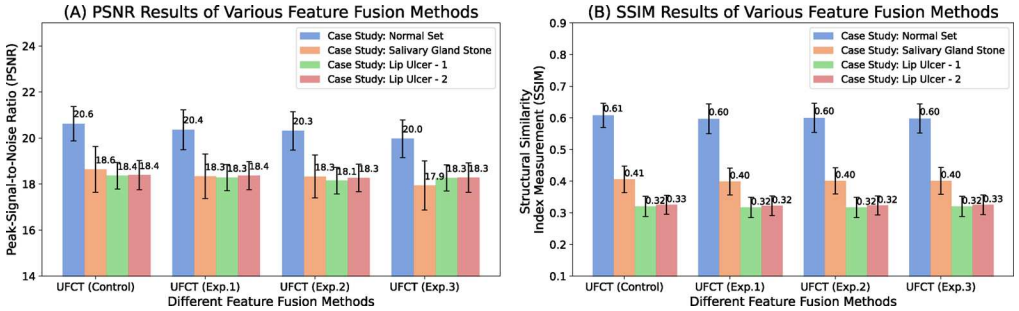

Fig. 8. Error bar of the results between different feature fusion methods (from exp.1 to exp.3). The results are shown in mean ± standard deviation. Control is the same as the proposed method.

Table 3. Computational Cost of Different Methods in Ablation Study (Latency Time is evaluated on RTX3050)

Method	FLOPs (G)	Parameters (M)	Latency Time
UFCT (Exp.1)	18.81	75.41	0.29 s/frame
UFCT (Exp.2)	19.71	76.46	0.29 s/frame
UFCT (Exp.3)	17.77	73.85	0.29 s/frame
UFCT (Control)	25.15	82.73	0.29 s/frame
UFCT-tiny	7.189	20.70	0.23 s/frame

5.4.2. Effect of the size of the UFCT

Table 3 shows the computational cost between the UFCT and UFCT-tiny. Table 4 shows the quantitative comparison between the UFCT and the UFCT-tiny. Compared with the UFCT, the UFCT-tiny model has lower FLOPs (7.189 G < 25.15 G) and parameters (20.70 M < 82.73 M); however, the performance of UFCT-tiny is lower than the proposed UFCT. We therefore suggest that utilize the UFCT with proposed implementation details as mentioned above.

Table 4. Quantitative Comparison in Ablation Study (Size of UFCT)

Size of UFCT	Normal Set (Cases #17-#22)		Salivary Gland Stone (Case #23)		Lip Ulcer 1 (Case #24)		Lip Ulcer 2 (Case #25)	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
UFCT (Control)	20.62 ± 0.74	0.608 ± 0.038	18.64 ± 0.99	0.406 ± 0.042	18.36 ± 0.58	0.320 ± 0.032	18.39 ± 0.63	0.325 ± 0.030
UFCT-tiny	20.45 ± 0.90	0.586 ± 0.049	18.52 ± 0.92	0.391 ± 0.040	18.23 ± 0.55	0.313 ± 0.031	18.34 ± 0.58	0.318 ± 0.029

5.4.3. Effect of training strategy

Figure 9 represents how the training loss function can influence the performance of different models. In normal sets (A, B), the models trained with the proposed combined loss can provide better PSNR and SSIM performances, except for the U-Net. In disease subjects (i.e., (D), (F), and (H)) SSIM results, most of the models trained with combined loss function (Eq. (7)) can provide better performance than the models trained with MSE-loss and adversarial loss. Moreover, we found that the models trained with adversarial loss will seriously degrade the

Biomedical Optics EXPRESS

PSNR performance in PSNR performance. In normal set (A, B), the comparison between MSE loss and combined loss shows that all transformer-type model trained with the combined loss (Eq. (7)) have better PSNR and SSIM results than the models trained with MSE loss. In case #23 (C, D), all transformer-type models trained with the combined loss can provide a higher PSNR performance than the models trained with the MSE loss. However, LUSwin-T and Swin-UNet will have a slight decrease on SSIM performance when trained with the combined loss, compared to the MSE loss. While UFCT can obtain a better SSIM performance when trained with the combined loss, compared to MSE loss. In case #24 and #25 (E-H), the UFCT trained with combined loss can provide a higher PSNR performance than the model trained with MSE loss. However, the SSIM performances are similar in terms of SSIM results.

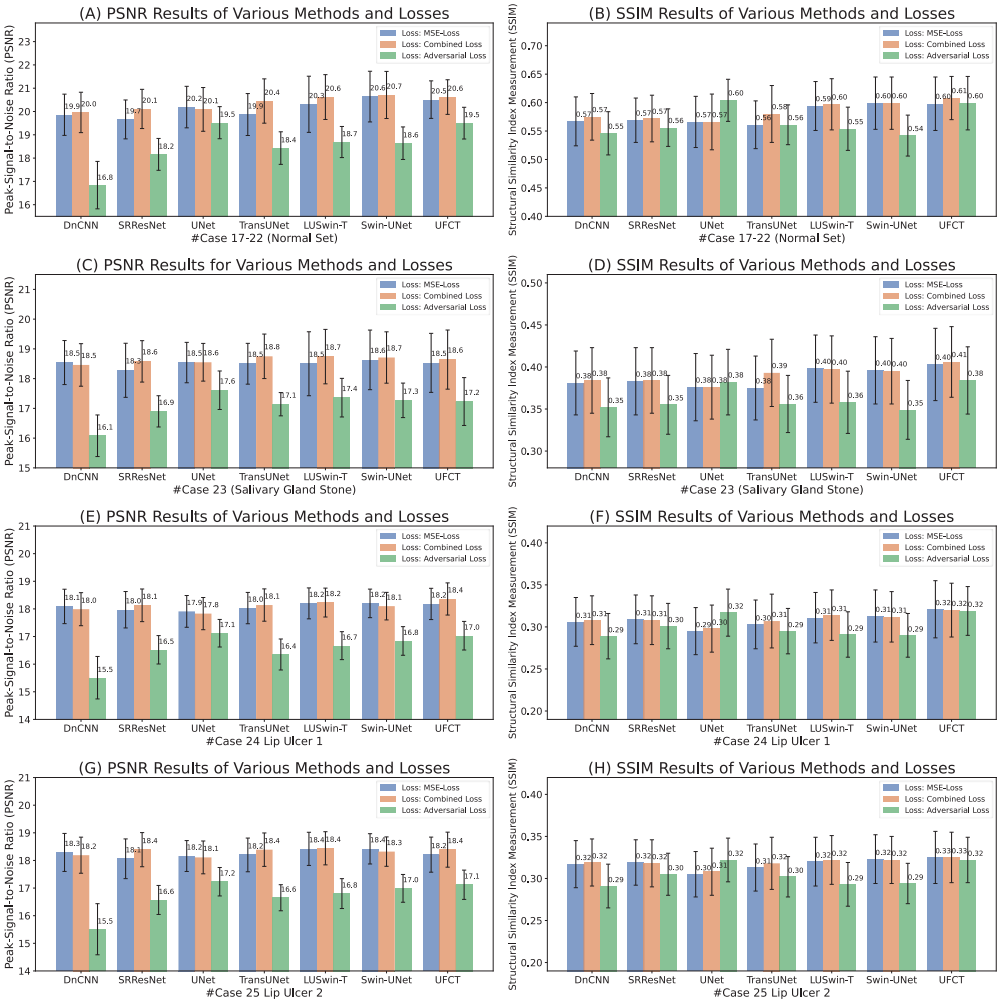

Fig. 9. Error bar of the results between different training strategies.

Biomedical Optics EXPRESS

6. Discussion and conclusions

In this study, we proposed a U-shaped fusion convolutional transformer (UFCT) model-based workflow to generate high-quality lip OCTA images based on a fast two-repeated OCT scan. The traditional vascular extraction method, the ED-OCTA algorithm, is not able to give micro-vasculature images with acceptable quality using two repeated scans. With the UFCT model, the proposed workflow can reconstruct an OCTA image with more vasculature texture details and higher contrast based on a fast two-repeated scan (~3.5 s), while presenting a similar visual quality as a four-repeated high-quality OCTA image in normal lip OCT data (Fig. 6). By combining the advantages of convolution operation and self-attention mechanism, a feature fusion method and convolutional transformer are introduced into the UFCT to enhance the capabilities of OCTA image reconstruction. Finally, we also provide the proposed feature fusion method and the optimal setup of the UFCT to obtain the best OCTA image reconstruction performance.

Due to having more accessibility than the pathological data, we built the training database in normal lip data from various subjects. To verify the performance of the proposed UFCT, we conducted a full comparison with a series of state-of-the-art models in Table **1**, including CNN-type, transformer-type, and Swin-transformer-type. The experiment results show that our UFCT has the best SSIM and visual performances among the four independent validation sets (i.e., one normal set, two ulcer sets, and one salivary gland stone). Based on visual comparison in Fig. 6 and Fig. 7, the proposed UFCT model has the best performance to adapt to other lip conditions (e.g., salivary gland stone) with training samples based on the normal lip.

Furthermore, we conduct an ablation study (Fig. 8) to show that the proposed fusion feature method can mostly increase the SSIM and PSNR performance of the proposed UFCT model, compared with the UFCT model that does not consist of a simple fusion network (SFN). While the latency time increased by SFN slightly (Table **3**). In terms of the training strategy, our results (Fig. 9) show that the proposed combined loss can have a higher PSNR and SSIM than the models trained with adversarial loss and MSE loss. We attribute the observed performance degradation of models using the adversarial loss function, as shown in Fig. 9, to the inherent instability of adversarial training. Such instability often culminates in suboptimal OCTA image reconstruction, as supported by Refs. [55,56]. Moreover, our study discerned a tendency for the discriminator model to develop premature confidence during adversarial training, further contributing to reduced OCTA image reconstruction performance.

It is worth noting that the human pathological datasets used in the current study were from patients with ulcers and salivary gland stones. Besides, the contrast of the reconstructed OCTA images in salivary gland stones (Fig. 7 (C-I)) is lower than ground-truth image (Fig. 7 (A)) in visual observation. We postulate this contrast discrepancy arises due to our models being primarily trained on normal lip datasets. This potentially hinders the models from accurately capturing the distinct features inherent to diseased lip data. Therefore, we will collect more lip disease OCT datasets in the future for the model training, ensuring that the proposed OCTA image reconstruction pipeline can perform better in lip disease subjects. Additionally, further work is required to examine the performance of the proposed method in cases of other lip disorders, which may have lower capillary flow speed and cause lower contrast of vascular images. Although we have achieved excellent results in the lip data, they are obtained by the same scanners, system, and scan mode. Further validation work should include performance in different commercial OCTA scanners.

This study has opened up a number of directions for future possible studies. For instance, the fast OCTA imaging workflow proposed in this study can be used for skin cancer diagnosis, providing in-depth information on the skin tissue non-invasively and rapidly. Furthermore, we observed that few operators looked at soft tissue such as oral tissue. It would be interesting to acquire OCTA images utilizing a lower-repeated scan and use the methodology reported in this paper to assist in understanding the standardization of the first oral OCTA scanner. The further

Biomedical Optics EXPRESS

direction of the study might utilize the current knowledge obtained from this study to provide the justification for, and subsequent evaluation of assistive tools for deep-learning-based image reconstruction for other domains such as photoacoustic, and 3D ultrasound.

In conclusion, we have demonstrated the ability of a vascular enhancement UFCT-based workflow to generate high-quality OCTA images in the lip using two repeated B-scans. Further follow-up studies will include expanding the disease incidence and demonstrating the external validity of the model using images from other institutions.

Appendix

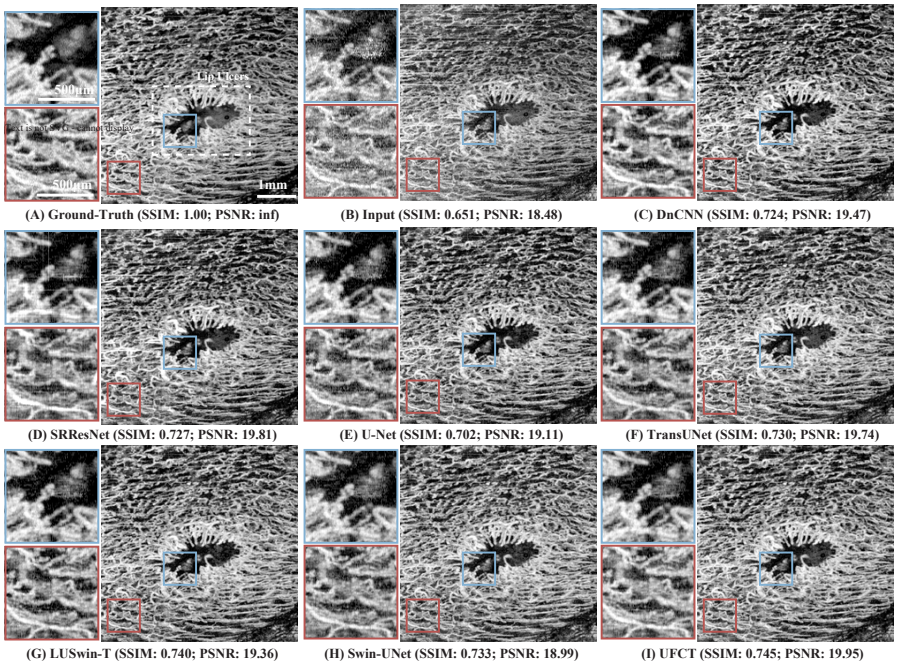

Fig. 10. Visual comparison of the OCTA image reconstruction based on lip OCT data with ulcers. (A) is a ground-truth image obtained by the ED-OCTA algorithm with a four-repeated OCT scan. (B) is input low-quality OCTA image obtained by ED-OCTA algorithm with two-repeated OCT scan. (C)-(I) are reconstructed OCTA images by different deep-learning-based models, including (C) DnCNN, (D) SRResNet, (E) U-Net, (F) TransUNet, (G) LUSwin-T, (H) Swin-UNet, and (I) UFCT. The white scale bar is used in (A).

Biomedical Optics EXPRESS

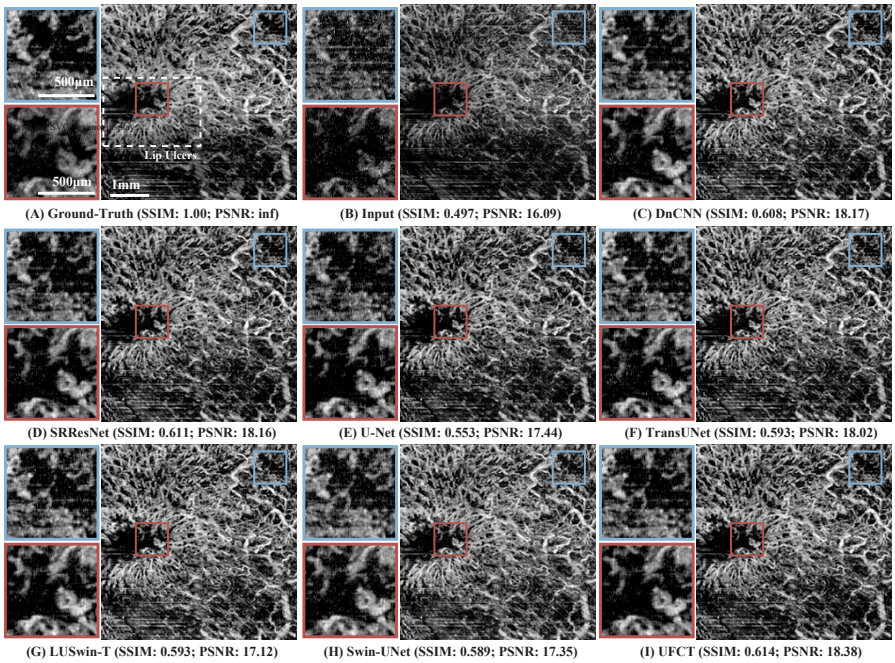

Fig. 11. Visual comparison of the OCTA image reconstruction based on lip OCT data with ulcers. (A) is a ground-truth image obtained by the ED-OCTA algorithm with a four-repeated OCT scan. (B) is input low-quality OCTA image obtained by ED-OCTA algorithm with two-repeated OCT scan. (C)-(I) are reconstructed OCTA images by different deep-learning-based models, including (C) DnCNN, (D) SRResNet, (E) U-Net, (F) TransUNet, (G) LUSwin-T, (H) Swin-UNet, and (I) UFCT. The white scale bar is used in (A).

Acknowledgments. Jinpeng would like to provide his greatest thanks to Prof. Zhihong Huang and Dr. Chunhui Li for their support and invaluable advice.

Disclosures. The authors declare no conflicts of interest.

Data availability. Data underlying the results presented in this paper are not publicly available at this time but may be obtained from the authors upon reasonable request.

References

1. C. Rivera, "Essentials of oral cancer," Int. J. Clin. Exp. Pathol. **8**, 11884 (2015).
2. S. A. Greenberg, B. J. Schlosser, and G. W. Mirowski, "Diseases of the lips," Clin. Dermatol. **35**(5), e1–e14 (2017).
3. S. Kraaij, K. H. Karagozoglu, T. Forouzanfar, E. C. I. Veerman, and H. S. Brand, "Salivary stones: symptoms, aetiology, biochemical composition and treatment," Br. Dent. J. **217**(11), E23 (2014).
4. S. Yogarajah and J. Setterfield, "Mouth ulcers and diseases of the oral cavity," Medicine **49**(7), 407–413 (2021).
5. J. W. Mays, M. Sarmadi, and N. M. Moutsopoulos, "Oral Manifestations of Systemic Autoimmune and Inflammatory Diseases: Diagnosis and Clinical Management," Journal of Evidence Based Dental Practice **12**(3), 265–282 (2012).
6. S. R. Porter and J. C. Leao, "Oral ulcers and its relevance to systemic disorders," Aliment. Pharmacol. Ther. **21**(4), 295–306 (2005).
7. R. Saini, N. V Lee, K. Y. P. Liu, and C. F. Poh, "Prospects in the application of photodynamic therapy in oral cancer and premalignant lesions," Cancers **8**(9), 83 (2016).
8. R. Mehrotra and D. K. Gupta, "Exciting new advances in oral cancer diagnosis: avenues to early detection," Head Neck Oncol. **3**(1), 1–9 (2011).
9. A. Gambino, A. Cafaro, R. Broccoletti, L. Turotti, D. Karimi, G. El Haddad, C. Hopper, S. R. Porter, L. Chiusa, and P. G. Arduino, "In vivo evaluation of traumatic and malignant oral ulcers with optical coherence tomography: A comparison between histopathological and ultrastructural findings," Photodiagn. Photodyn. Ther. **39**, 103019 (2022).
10. M. Zimmermann, A. Mehl, W. H. Mörmann, and S. Reich, "Intraoral scanning systems-a current overview," Int. J. Comput. Dent. **18**, 101–129 (2015).

Biomedical Optics EXPRESS

11. D. M. Roblyer, R. R. Richards-Kortum, K. v Sokolov, A. K. El-Naggar, M. D. Williams, C. Kurachi, and A. Gillenwater, "Multispectral optical imaging device for in vivo detection of oral neoplasia," J. Biomed. Opt. **13**(2), 024019 (2008).
12. C. F. Poh, L. Zhang, D. W. Anderson, J. S. Durham, P. M. Williams, R. W. Priddy, K. W. Berean, S. Ng, O. L. Tseng, and C. MacAulay, "Fluorescence visualization detection of field alterations in tumor margins of oral cancer patients," Clin. Cancer Res. **12**(22), 6716–6722 (2006).
13. W. Fujimoto and J. G. Drexler, "Introduction to OCT," in *Optical Coherence Tomography: Technology and Applications*, J. G. Drexler and W. Fujimoto, eds. (Springer International Publishing, 2015), pp. 3–64.
14. A. Gambino, M. Cabras, A. Cafaro, R. Broccoletti, S. Carossa, C. Hopper, D. Conrotto, S. R. Porter, and P. G. Arduino, "Preliminary evaluation of the utility of optical coherence tomography in detecting structural changes during photobiomodulation treatment in patients with atrophic-erosive oral lichen planus," Photodiagn. Photodyn. Ther. **34**, 102255 (2021).
15. A. Gambino, M. Cabras, A. Cafaro, R. Broccoletti, S. Carossa, C. Hopper, L. Chiusa, G. El Haddad, S. R. Porter, and P. G. Arduino, "In-vivo usefulness of optical coherence tomography in atrophic-erosive oral lichen planus: Comparison between histopathological and ultrastructural findings," J. Photochem. Photobiol., B **211**, 112009 (2020).
16. W. Jerjes, Z. Hamdoon, A. A. Yousif, N. H. Al-Rawi, and C. Hopper, "Epithelial tissue thickness improves optical coherence tomography's ability in detecting oral cancer," Photodiagn. Photodyn. Ther. **28**, 69–74 (2019).
17. D. Di Stasio, D. Lauritano, F. Loffredo, E. Gentile, F. Della Vella, M. Petruzzi, and A. Lucchese, "Optical coherence tomography imaging of oral mucosa bullous diseases: A preliminary study," Dentomaxillofacial Radiology **49**(2), 20190071 (2020).
18. W. Jerjes, T. Upile, B. Conn, Z. Hamdoon, C. S. Betz, G. McKenzie, H. Radhi, M. Vourvachis, M. el Maaytah, and A. Sandison, "In vitro examination of suspicious oral lesions using optical coherence tomography," British Journal of Oral and Maxillofacial Surgery **48**(1), 18–25 (2010).
19. B. Zabihian, Z. Chen, E. Rank, C. Sinz, M. Bonesi, H. Sattmann, J. R. Ensher, M. P. Minneman, E. E. Hoover, and J. Weingast, "Comprehensive vascular imaging using optical coherence tomography-based angiography and photoacoustic tomography," J. Biomed. Opt. **21**(09), 1 (2016).
20. W. Wei, W. J. Choi, S. Men, S. Song, and R. K. Wang, "wide-field and long-ranging-depth optical coherence tomography microangiography of human oral mucosa (Conference Presentation)," in *Proc. SPIE* (2018), Vol. 10473, p. 104730 H.
21. W. Wei, W. J. Choi, and R. K. Wang, "Microvascular imaging and monitoring of human oral cavity lesions in vivo by swept-source OCT-based angiography," Lasers Med. Sci. **33**(1), 123–134 (2018).
22. W. J. Choi and R. K. Wang, "In vivo imaging of functional microvasculature within tissue beds of oral and nasal cavities by swept-source optical coherence tomography with a forward/side-viewing probe," Biomed. Opt. Express **5**(8), 2620–2634 (2014).
23. A. Mariampillai, B. A. Standish, E. H. Moriyama, M. Khurana, N. R. Munce, M. K. K. Leung, J. Jiang, A. Cable, B. C. Wilson, and I. A. Vitkin, "Speckle variance detection of microvasculature using swept-source optical coherence tomography," Opt. Lett. **33**(13), 1530–1532 (2008).
24. S. Yousefi, Z. Zhi, and R. K. Wang, "Eigendecomposition-based clutter filtering technique for optical microangiography," IEEE Trans. Biomed. Eng. **58**(8), 2316–2323 (2011).
25. A. Tavakkoli, S. A. Kamran, K. F. Hossain, and S. L. Zuckerbrod, "A novel deep learning conditional generative adversarial network for producing angiography images from retinal fundus photographs," Sci. Rep. **10**(1), 21580 (2020).
26. M. Gao, Y. Guo, T. T. Hormel, J. Sun, T. S. Hwang, and Y. Jia, "Reconstruction of high-resolution 6× 6-mm OCT angiograms using deep learning," Biomed. Opt. Express **11**(7), 3585–3600 (2020).
27. X. Liu, Z. Huang, Z. Wang, C. Wen, Z. Jiang, Z. Yu, J. Liu, G. Liu, X. Huang, A. Maier, Qiushu Ren, and Yanye Lu, "A deep learning based pipeline for optical coherence tomography angiography," J. Biophotonics **12**(10), e201900008 (2019).
28. Z. Jiang, Z. Huang, B. Qiu, X. Meng, Y. You, X. Liu, M. Geng, G. Liu, C. Zhou, and K. Yang, "Weakly supervised deep learning-based optical coherence tomography angiography," IEEE Trans. Med. Imaging **40**(2), 688–698 (2021).
29. Z. Jiang, Z. Huang, B. Qiu, X. Meng, Y. You, X. Liu, G. Liu, C. Zhou, K. Yang, and A. Maier, "Comparative study of deep learning models for optical coherence tomography angiography," Biomed. Opt. Express **11**(3), 1580–1597 (2020).
30. C. S. Lee, A. J. Tyring, Y. Wu, S. Xiao, A. S. Rokem, N. P. DeRuyter, Q. Zhang, A. Tufail, R. K. Wang, and A. Y. Lee, "Generating retinal flow maps from structural optical coherence tomography with artificial intelligence," Sci. Rep. **9**(1), 1–11 (2019).
31. P. L. Li, C. O'Neil, S. Saberi, K. Sinder, K. Wang, B. Tan, Z. Hosseinaee, K. Bizhevat, and V. Lakshminarayanan, "Deep learning algorithm for generating optical coherence tomography angiography (OCTA) maps of the retinal vasculature," in *Applications of Machine Learning 2020* (International Society for Optics and Photonics, 2020), Vol. 11511, p. 1151109.
32. J. Liao, S. Yang, T. Zhang, C. Li, and Z. Huang, "A Fast Optical Coherence Tomography Angiography Image Acquisition and Reconstruction Pipeline for Skin Application," Biomed. Opt. Express **14**(8), 3899–38913 (2023).
33. Y. Song, J. Y.-C. Teoh, K.-S. Choi, and J. Qin, "Dynamic Loss Weighting for Multiorgan Segmentation in Medical Images," IEEE Trans Neural Netw Learn Syst (2023).

Biomedical Optics EXPRESS

34. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, and S. Gelly, "An image is worth 16×16 words: Transformers for image recognition at scale," arXiv, arXiv:2010.11929 (2020).
35. H. Chen, Y. Wang, T. Guo, C. Xu, Y. Deng, Z. Liu, S. Ma, C. Xu, C. Xu, and W. Gao, "Pre-trained image processing transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 12299–12310.
36. J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "Swinir: Image restoration using swin transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 1833–1844.
37. S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Restormer: Efficient transformer for high-resolution image restoration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 5728–5739.
38. H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-unet: Unet-like pure transformer for medical image segmentation," arXiv, arXiv:2105.05537 (2021).
39. Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 10012–10022.
40. J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," arXiv, arXiv:2102.04306 (2021).
41. H. Wu, B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan, and L. Zhang, "Cvt: Introducing convolutions to vision transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 22–31.
42. J. Wang, Z. Wei, T. Zhang, and W. Zeng, "Deeply-fused nets," arXiv, arXiv:1605.07716 (2016).
43. J. Liao, C. Li, and Z. Huang, "A Lightweight Swin Transformer-Based Pipeline for Optical Coherence Tomography Image Denoising in Skin Application," in *Photonics* (MDPI, 2023), Vol. 10, p. 468.
44. S. Klein, M. Staring, K. Murphy, M. A. Viergever, and J. P. W. Pluim, "Elastix: a toolbox for intensity-based medical image registration," IEEE Trans. Med. Imaging **29**(1), 196–205 (2010).
45. K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," arXiv, arXiv:1409.1556 (2014).
46. Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," IEEE Trans. on Image Process. **13**(4), 600–612 (2004).
47. K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising," IEEE Trans. on Image Process. **26**(7), 3142–3155 (2017).
48. C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, and Z. Wang, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017), pp. 4681–4690.
49. O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Springer, 2015), pp. 234–241.
50. X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. Change Loy, "Esrgan: Enhanced super-resolution generative adversarial networks," in *Proceedings of the European Conference on Computer Vision (ECCV)* (2018), p. 0.
51. I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Nets," in *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, eds. (Curran Associates, Inc., 2014), Vol. 27.
52. M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, and M. Isard, "Tensorflow: A system for large-scale machine learning," in *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)* (2016), pp. 265–283.
53. D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv, arXiv:1412.6980 (2014).
54. H. Chen, Y. Wang, J. Guo, and D. Tao, "VanillaNet: the Power of Minimalism in Deep Learning," arXiv, arXiv:2305.12972 (2023).
55. M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *International Conference on Machine Learning* (PMLR, 2017), pp. 214–223.
56. M. Mirza and S. Osindero, "Conditional generative adversarial nets," arXiv, arXiv:1411.1784 (2014).