VET: Vasculature Extraction Transformer for Single-Scan Optical Coherence Tomography Angiography

Jinpeng Liao, Tianyu Zhang, Yilong Zhang, Chunhui Li*, and Zhihong Huang

***Abstract*—Optical coherence tomography angiography (OCTA) is a non-invasive imaging modality for analyzing skin microvasculature, enabling non-invasive diagnosis and treatment monitoring. Traditional OCTA algorithms necessitate at least two-repeated scans to generate microvasculature images, while image quality is highly dependent on the repetitions of scans (e.g., 4-8). Nevertheless, a higher repetition count increases data acquisition time, causing patient discomfort and more unpredictable motion artifacts, which can result in potential misdiagnosis. To address these limitations, we proposed a vasculature extraction pipeline based on the novelty vasculature extraction transformer (VET) to generate OCTA images by using a single OCT scan. Distinct from the vision Transformer, VET utilizes convolutional projection to better learn the spatial relationships between image patches. This study recruited 15 healthy participants. The OCT scans were performed in five various skin sites, i.e., palm, arm, face, neck, and lip. Our results show that in comparison to OCTA images obtained by the speckle variance OCTA (peak-signal-to-noise ratio (PSNR): 16.13) and eigen-decomposition OCTA (PSNR: 17.08) using four repeated OCT scans, OCTA images extracted by the proposed pipeline exhibit a better PSNR (18.03) performance while reducing the data acquisition time by 75%. Visual comparisons show that the proposed pipeline outperformed traditional OCTA algorithms, particularly in the imaging of lip and face areas, where artifacts are commonly encountered. This study is the first to demonstrate that the VET can efficiently extract high-quality vasculature images from a single, rapid OCT scan. This capability significantly enhances diagnostic accuracy for patients and streamlines the imaging process.**

***Index Terms*—Image reconstruction, Optical coherence tomography angiography, Deep-learning.**

Manuscript received xx, xx, 2023; revised xx xx, 2023; accepted xx xx, 202x. Date of publication xx xx, 20xx; date of current version xx xx, 202x. (Jinpeng Liao is the first author.) (Corresponding author: Chunhui Li.).

Jinpeng Liao, Tianyu Zhang, Yilong Zhang, Chunhui Li and Zhihong Huang are with the School of Science and Engineering, University of Dundee, DD1 4HN, UK. (e-mail: jyliao@dundee.ac.uk, t.x.zhang@dundee.ac.uk, y.y.zhang@dundee.ac.uk, c.li@dundee.ac.uk, z.y.huang@dundee.ac.uk).

I. Introduction

SKIN microvascular mapping can not only help identify skin pathological conditions but also provide insights into systemic diseases [1]–[3]. For instance, decreased microvascular density has been associated with cardiovascular and metabolic diseases, such as hypertension, diabetes, obesity, and metabolic syndrome, as well as an increased risk of coronary artery disease [4]–[7]. Optical coherence tomography angiography (OCTA) is an extension function based on OCT, providing a microvascular image by extracting the moving red blood cell signals from the surrounding relatively static biological tissue signal [8]–[10]. In scholarly research, OCTA has been substantiated as a method for identifying skin disease by assessing the distribution of vasculature [11]; in particular, it has emerged as a valuable tool for analyzing skin microvasculature, allowing for non-invasive diagnosis and monitoring of treatment in skin diseases and cancer [12]–[14]. As the clinical application of OCTA increases, enhancing its imaging speed and quality will drive precise diagnostics and treatment plans, while also widening its potential clinical applications, such as in oral and endoscopic procedures.

Among the conventional OCTA algorithms that utilize the differentiation of information (e.g., phase and complex information) present in OCT signals, speckle variance (SV)-OCTA [15] and eigen-decomposition (ED)-OCTA [16] are highly efficient methods for extracting vasculature images [9]. However, *in vivo* skin OCTA imaging faces challenges that can significantly compromise the quality of vascular signals. These include the speckle noise inherent to the OCT system, bulk tissue motion-induced artifacts, and light wave scattering due to the complex structure of skin tissues. Moreover, *in vivo* skin OCTA scan requires a flexible scanning probe to approach various sites of sun-exposed skin (e.g., face, hand, and arm areas with higher skin cancer risk [17]). This method introduces additional motion artifacts due to patient and probe movement.

Implementing a higher repetition of OCT scans (e.g., 6 times) can improve the quality of OCTA images generated through conventional algorithms [18]. Nevertheless, this method also lengthens the data acquisition time, thereby introducing a greater likelihood of unpredictable motion artifacts from both the scanning probe and the patient. Such artifacts can adversely affect the overall image quality. An

additional option to reduce motion artifacts and scanning time is to increase the swept rate of the swept laser. However, this solution comes with its own drawbacks. Upgrading the laser system can be expensive, and a higher swept-rate laser requires a higher-performance gravo-mirror for effective adaptation.

A series of convolution neural network (CNN)-based methods were proposed to reconstruct the high-quality OCTA images by using two- or four-repeated OCT scans [19]–[21]. Those approaches have been successful in reconstructing high-quality OCTA images with low repetition of OCT scans, but they have largely been applied to the study of mice brains through invasive OCTA scans. Instead of exclusively focusing on OCTA image reconstruction, it is critical for the models to relearn the distinct characteristics of skin vasculatures in the field of dermatology. In addition, these require a minimum of two repeated OCT scans for high-quality vasculature imaging.

In this study, we propose a vasculature extraction transformer (VET)-based pipeline designed to facilitate *in vivo* skin OCTA scan speed and mitigate motion artifacts associated with the use of a flexible scanning probe. Contrary to previous deep-learning methods that require at least two repeated scans, our proposed pipeline targets the extraction of skin micro-vasculature images from a single OCT scan. Our pipeline enables real-time OCTA imaging as it eliminates the need for an offline process to extract vasculature signals. This allows for OCTA images to be directly procured from single-scan-based structural OCT images. As for the deep neural network applied in OCTA image restoration, VET employs the strengths of convolutional projection [22] and Transformer for vasculature feature extraction. Varied from the linear projection used in Transformer, convolutional projection uses a convolution operation to obtain the key, value, and query sequences, providing spatial relationships between the image patches.

Consequently, our study has the following contributions: (1) To the best of our knowledge, we are the first to introduce a single-scan-based OCTA imaging pipeline that significantly reduces the data acquisition time by up to 75%, while improve the peak-signal-to-noise ratio performance as compared to four-repeated OCTA images produced by ED-OCTA and SV-OCTA algorithms. (2) We proposed a novel VET model that uses convolutional projection to help the model learn the spatial relationships between the image patches. (3) As far as we are aware, this is the first competitive study of neural networks in skin OCTA imaging to extract vasculature images based on a single OCT scan. (4) We evaluate the performance of the proposed pipeline with a flexible scanning probe for five different scan positions.

A. Related Work

CNN-based approaches have demonstrated their capability to reconstruct high-quality skin OCTA images using only two repeated scans. These include techniques utilizing denoising deep convolution neural network (DnCNN) [19], residual deep neural networks [21], [23], residual densely deep neural network [24], and U-shape deep neural network [25]. However, current CNN models fall short of the requisite capabilities needed for high-quality reconstruction of skin OCTA images in this study. Since the CNN-based methods are difficult to learn the global and long-term information [26], [27], and they also have a high dependency on the locality convolution operation.

Recently, vision transformer (ViT) has gained attention as an alternative to CNNs for image classification tasks due to their scalability, flexibility, and ability to handle long-range dependencies [28]. In Liu et al. work [29], a hierarchical shift window (Swin)-transformer was proposed and achieved state-of-the-art results in image classification. Based on the Swin-transformer, SwinIR [26] was proposed to reconstruct the high-quality nature images from the counterpart degraded images, and Swin-UNet [27] for medical image segmentation, and both of them achieved better competitive results than the CNN models. ViT and Swin-Transformer architectures use a linear projection layer (also referred to as a fully connected layer) to generate query, key, and value sequences for multi-head self-attention. However, this can result in a significant increase in the number of parameters, which can affect the efficiency and practicality of these models. Besides, the limitation of the linear projection layer is that it does not take into account the spatial relationships between the patches, which can be important for OCTA image reconstruction in this study.

II. Vasculature Extraction Methods

A. Conventional OCTA Algorithms

Speckle variance (SV) algorithm based on consecutive B-scans is performed to obtain motion-contrast information, which can be formulated as the (1):

$$Flow_{SV}(x,z) = \frac{1}{NR}\sum_{i=1}^{N} |(A_{i+1}(x,z) - A_i(x,z)| \qquad (1)$$

where NR is the number of repeated scans at the same location. $A_i(x,z)$ indicates the amplitude signal in i-th B-scans at lateral location x and depth position z.

Eigen decomposition (ED) algorithm is following the principle of orthogonality. Orthogonality gave the idea that an autocorrelation matrix, containing noise subspace eigenvectors is orthogonal to the signal eigenvectors. By suppressing the eigenvectors with a large numerical value that represents the static tissues, the clarity vascular image is extracted, according to [8]. The procedure is in (2), (3):

$$E \wedge E^H = \sum_{i=1}^{N} \lambda_B(i) e_B(i) e_B^H(i) \qquad (2)$$

where $E = [e_B(1), e_B(2), \dots, e_B(N)]$ is the $N \times N$ unitary matrix of eigenvectors, $\wedge = [\lambda_B(1), \lambda_B(2), \dots, \lambda_B(N)]$ is the $N \times N$ diagonal matrix of eigenvalues, and H is the Hermitian transpose. The eigenvalues $\wedge$ are sorted in descending order. By subtracting the first k^{th} eigenvectors which mainly are tissue signals, the extraction of the vessel signals X_v under K-repeat scans OCT signal X can be written as (3):

$$X_v = \left[I - \sum_{i=1}^{K} e_B(i) e_B^H(i)\right] X \qquad (3)$$

where I is the identity matrix. $e_B(i)$ is the $1 \times N$ unitary matrix of eigenvectors.

B. Single-Scan Vasculature Extraction Pipeline

A schematic diagram of the single-scan vasculature extraction pipeline and neural network training pipeline is

shown in Fig. 1. In the training stage, the input of neural networks is generated based on the first repeat of multi-repeated OCT signals. The high-quality vascular signal for neural network loss calculation is extracted by the all-repeated OCT signal with the ED-OCTA algorithm. In the test stage, the trained network utilizes the structural image generated based on the single-scan OCT signal and outputs the predicted vascular signal. The data preprocessing for neural network training, validation, and testing will be described in the following paragraph.

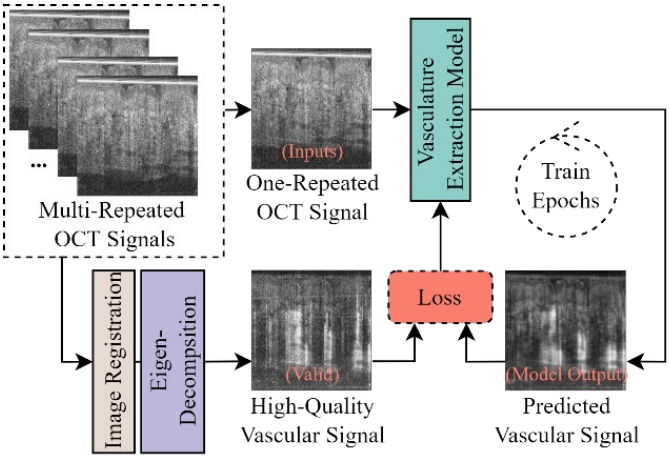

Fig. 1. The vasculature extraction pipeline for single-scan OCT image, including the model training pipeline. In the training stage, the predicted vascular signal from the model is used to calculate the loss for the vasculature extraction model's trainable weights updating.

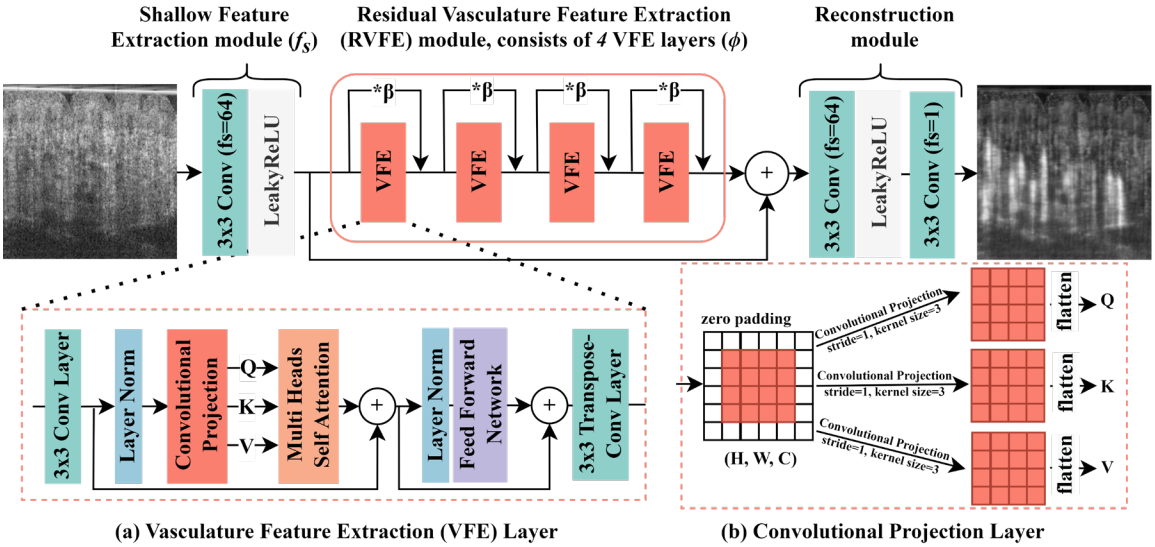

Fig. 2. The architecture of the proposed vasculature extraction transformer.

C. *Vasculature Extraction Transformer*

Vasculature Extraction Transformer (VET) consists of three modules: shallow feature extraction, residual vasculature feature extraction (RVFE), and feature combination and output block, as shown in Fig. 2.

Shallow feature extraction. The shallow feature extraction layer (f_s) is composed of a 3 × 3 convolution layer (64 filters and strides 1) with a LeakyReLU activation layer. Given a single scan OCT signal (i.e., structural image) input I_{stru} with shape H × W × C, where H, W, and C are image height, width, and channel, respectively, and the processing of the shallow feature extraction layer can be written as:

$$F_s = LeakyReLU\left(f_s(I_{stru})\right) \quad (4)$$

where F_s is the obtained shallow feature of the input structural image. According to [30], incorporating an early convolution layer in a transformer architecture model for visual processing can improve optimization stability and lead to improved results.

Residual vasculature feature extraction. The residual vasculature feature extraction (RVFE) consists of four VFE layers (ϕ) and leverages a residual scaling parameter (β) to establish an identity connection between VFE layers and the reconstruction module, allowing the aggregation of different levels of features. The forward processing of a VFE layer and a residual connection in RVFE can be written as:

$$F_{out} = F_{in} * \beta + \phi(F_{in}) \quad (5)$$

where F_{in} is the input feature from the previous layer, and F_{out} is the output feature, residual scaling parameter β is set as 0.4. The architecture of the VFE layer is illustrated in Fig. 2 (a), while Fig. 2 (b) depicts the convolutional projection layer, inspired by [22]. To mitigate the computing cost of multi-head self-attention, in the VFE layer, we employ a 3 × 3 convolution layer (f_{c1}) with a stride of 2 that downsample the input feature (F_{input}) shape from H × W × C to H/2 × W/2 × C.

$$F_{c1} = f_{c1}(F_{input}) \quad (6)$$

where F_{c1} is the output downsampled features with shape H/2 × W/2 × C, and F_{c1} is then used as the input of the

convolutional projection layer. To ensure both training effectiveness and stability, we opt for a different approach than the squeezed convolutional projection layer used in [22]. Instead, we implement a 3 × 3 convolution projection layer (f_{CP}) to obtain query (Q), key (K), and value (V) sequences. This processing procedure (Fig. 2 (b)) can be formulated as:

$$Q, K, V = Flatten\left(f_{CP}(LN(F_{c1}))\right) \quad (7)$$

where LN is the layer normalization layer, and output $Q, K,$ and V are then used as the input for multi-head self-attention (MSA). After Flatten processing, the shape of Q, K, and V sequences is (HW/4) × C, and each sequence is split with multi-head by reshaping from (HW/4) × C to M × (HW/4) × C/M, where M is the number of heads. The attention score of each head (M) is then computed using the self-attention mechanism as (8). We perform the attention function in parallel M times and concatenate the resulting scores to achieve multi-head self-attention.

$$F_{score} = Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d}}\right) * V \quad (8)$$

where d is a rescale parameter with a numerical value of $1/\sqrt{dims\ of\ Q}$. After the multi-head self-attention operation, the shape of the feature map is (HW/4) × C. Then, a feed-forward network (FFN) that consists of two fully-connected layers with a GELU non-linearity activation layer between them is used for feature transformations. A 2D reshape layer is used to reshape the output of FFN from (HW/4) × C to H/2 × W/2 × C. Finally, a 3 × 3 transpose convolution layer (f_{tc1}) with a stride of 2 is used to upscale the shape of the feature map from H/2 × W/2 × C to H × W × C. Generally, the whole process of a VFE layer is formulated as (9) and (10):

$$Y = MSA(f_{CP}(LN(f_{c1}(X))) + f_{c1}(X) \quad (9)$$

$$Output = f_{tc1}(FFN(LN(Y)) + Y) \quad (10)$$

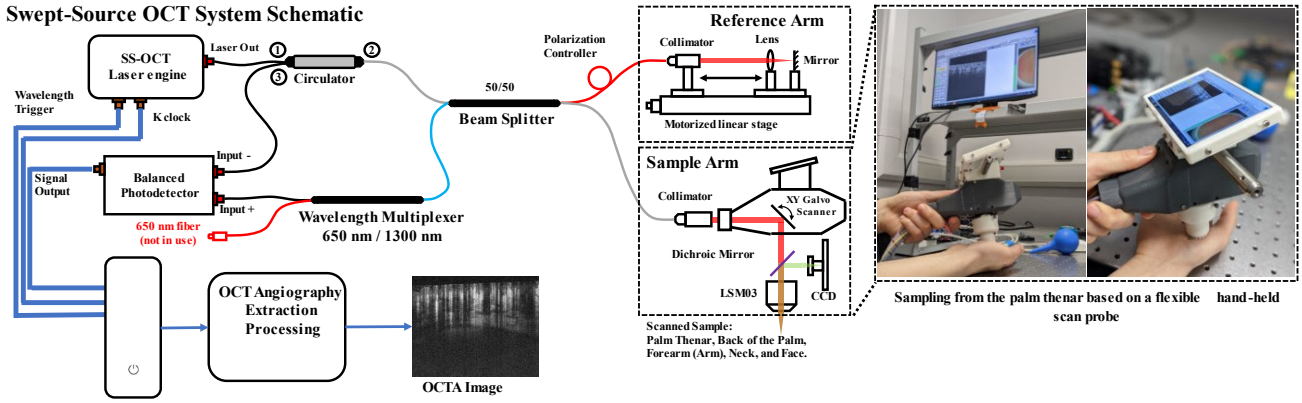

Fig. 3. The schematic of the lab-built swept-source optical coherence tomography system. The Laser wavelength is 1310 nm with 100nm bandwidth. The A-scan swept rate is 200 kHz. The flexible hand-held scan probe (sample lens) is demonstrated in the right figure.

Reconstruction module. We reconstruct the vascular signal by aggregating the shallow features (F_s) from shallow feature extraction module and deep features (F_{RVEF}) from residual vasculature feature extraction module:

$$I_V = H_R(F_s + F_{RVEF}) \quad (11)$$

where I_V is the reconstructed vascular signal, and H_R is the reconstruction module as depicted in Fig. 2. Shallow features primarily contain low-frequency details, whereas deep features concentrate on recovering lost high-frequency vascular signals. To enhance the ability of feature integration and increase the non-linearity of VET, a convolution layer with a ReLU activation layer is used to combine low-frequency and high-frequency details extracted from the shallow feature extraction module and RVFE module, respectively. Besides, a convolution layer with a filter size of 1 is used as the output layer of VET to output the reconstructed vascular signal. Notably, VET utilizes a global skip connection from the shallow feature extraction module to transmit low-frequency information directly to the reconstruction module. This enables the deep feature extraction module to focus on high-frequency information and stabilize training [26].

III. Experiment Setup

A. Data Acquisition and Pre-Processing

A lab-built 200 kHz swept rate swept-source (SS)OCT scan system was utilized to non-invasively collect the OCT data with a hand-held probe, as demonstrated in Fig. 3. More details of the SSOCT system were demonstrated in [31]. The data collection of the volunteers was approved by the School of Science and Engineering Research Ethics Committee of University of Dundee, which also conformed to the tenets of the Declaration of Helsinki. All participants had to give their informed consent before entering the lab for the data collection, and the data collected in this article obtained the informed consent of the participants. To develop a comprehensive assessment of the proposed VET, the scan positions were palm and arm (representative 'thick' skin), and face, lip, and neck (representative 'thin' skin) taken from 26 subjects ages between 20 and 35 years old. Among them, 22 subjects are healthy and none of them have any disease condition, 3 subjects have lip ulcers, and 1 subject has face acne.

In terms of imaging protocol for data acquisition, one OCTA scan can acquire data with a pixel size of NR × 600 × 600 × 300 (NR × x × y × z). Here, NR refers to the number of

repeated scans, while x and y represent the transverse axis, and z represents the axial axis. During the OCTA data acquisition for healthy subjects, 12 repeated scans were performed for the palm and arm area, and 6 for the face, neck, and lip areas. Regarding the subjects with lip ulcers or face acne, the repetition of the OCT scan is set as 8. Each repeated scan took approximately 1.8 seconds. The spatial interval in the transverse axis is ~8.6 μm/pixel and ~ 7.4 μm/pixel in the axial axis. After manually removing the low-quality and high-motion artifacts data, we finally collected a total of 42 OCT raw data (14 palm, 5 face, 4 neck, 2 arm, 13 lip, and 3 lip with ulcers, and 1 face with acne). 21 raw data (9 palm, 2 face, 2 neck, 1 arm, and 7 lip) were randomly selected to generate train datasets. 6 raw data (2 palm, 1 face, 1 neck, and 2 lip) were for validation. In terms of the test set, the remaining 11 raw data from healthy subjects (3 palm, 2 face, 1 neck, 1 arm, and 4 lip) were used to generate a healthy test set. 3 raw data with lip ulcers and 1 raw data with face acne were used as a disease test set. The training dataset was used to train models in this study, while validation sets were used to monitor the model training and prevent overfitting. The separated test set was used to evaluate the performance of the trained models, preventing data leakage.

The flowchart for dataset pre-processing is shown in Fig. 4. To better describe the data pre-processing, we define that one OCT raw data consists of NR volumes, and each volume has a size of 1 × 600 × 600 × 300 (1 × x × y × z), where NR is the number of repeated OCT scans. Firstly, all NR volumes are processed by frame-to-frame registration based on the fast Fourier transform (FFT), and then an FFT-based per A-lines alignment is used to reduce the motion artifacts [32], [33]. The ground-truth high-quality OCTA images are generated by using all NR volumes with ED-OCTA algorithms mentioned in (3). Since the ED-OCTA has an outstanding performance in suppressing static tissue while preserving vascular signals [34]. The input skin structural images for neural networks are then generated by using only one OCT volume. The baseline OCTA images are obtained by SV-OCTA and ED-OCTA algorithms with the first four OCT volumes. Since the four-repeated OCTA scans are most frequently used in clinical setups, based on the consideration of imaging acquisition efficiency.

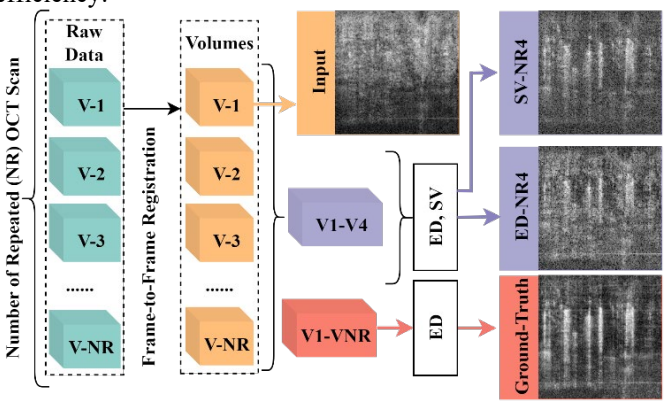

Fig. 4. Flow chart of the scanning and processing strategy to create ground-truth OCTA results using twelve-repeated scans, baseline OCTA results using four-repeated scans, and the strategy to obtain input structural images based on a single OCT scan. The frame-to-frame registration consists of fast Fourier transfer (FFT) to obtain structural volume and FFT-based per A-lines alignment (V-1 is used as reference) to reduce the motion artifacts.

After the data pre-processing for all 42 OCT raw data (21 for training, 6 for validation, and 15 for testing), 25200 B-frames were extracted (42 × 600 frames/data). An image crop box with a size of 192 × 192 is then used to extract image patches from each B-frame image. Finally, a total of 75600 pairs of images are generated for the ground-truth, baseline, and input datasets. Among them, 37800 images (from 21 raw data) are used as training datasets for neural network training, and 10800 images (from 6 raw data) are used as the validation set. The remaining 19800 images (from 11 health raw data) are selected as the healthy test set, and 7200 images (from 3 lip ulcers raw data and 1 face acne raw data) are selected as the disease test set.

B. Implementation Details

The VET is trained based on TensorFlow 2.9.0. To enhance data diversity during the training phase, data augmentation techniques such as flipping and rotations were employed, contributing to the improved generalization of the trained model and mitigating overfitting. The filter size for all convolution layers in the VET is set to 64, with the exception of the final output convolution layer. Within the feed-forward network, the first fully-connected layer comprises 256 hidden units, while the second fully-connected layer contains 64 units. All other aspects of the VET implementation remain consistent with the methodology described in the corresponding section.

The VET model was optimized using an Adam optimizer [35] (with a 0.0001 learning rate, 0.8 for beta1, and 0.999 for beta2) on an Nvidia A100 with 40GB memory. The training process utilized a batch size of 4 and ran for 200 epochs, using mean-square-error (MSE) as the loss function because it can provide a better performance and training stability over the mean-absolute-error loss function in this study.

C. Comparison with the Networks

To assess the performance of our proposed VET model for vasculature extraction, we conducted a comparative analysis of the image quality between OCTA images extracted using various neural networks, including DnCNN [36], U-Net [37], SRGAN [38], ESRGAN [39], TransUNet [40], SwinIR [26], Swin-UNet [27], UFormer [41], Restormer [42], and Lightweight U-shape Swin-Transformer (LUSwin-T) [43]. The image quality evaluation of the OCTA images was performed both quantitatively and qualitatively. Additionally, we provide the total number of parameters and floating-point operations (based on a 192 × 192 image size).

Notably, SRGAN and ESRGAN were originally designed for natural image super-resolution; therefore, we removed the upsample layers from these two networks. To minimize the influence of network training specifics, we maintained the implementation details for DnCNN, SRGAN, ESRGAN, SwinIR, UFormer, Restormer and LUSwin-T as per the published sources. As for U-Net, TransUNet, and Swin-UNet, which were initially developed for image segmentation, we utilized the mean squared error (MSE) loss function with supervised training (i.e., the same as the VET implementation details). Regarding the optimizer, epochs, batch size, and data

augmentation, all compared networks follow the same configuration as described in Section III.B.

D. Ablation Study Setup

To investigate the reconstruction performance of the proposed VET under different settings of the model, we further performed an ablation study in terms of head number, the number of VFE layers in RVFE, and the filter size of all convolution layers. The details setup of different parameters are depicted in **Table 1**, and the setups with the underline are the control group, which has the same implementation details as this study proposed.

Table 1. Experiment setup for the ablation study

Study	Parameter Setup			
VFE Layer	2	<u>4</u>	6	8
Heads	2	<u>4</u>	8	16
Filter Size	32	48	<u>64</u>	80

E. Evaluation Metrics

To conduct a quantitative performance comparison of various methods, including SV-OCTA, ED-OCTA, and deep-learning-based methods, this study utilized peak-signal-to-noise ratio (PSNR), structural similarity (SSIM) [44] and multi-scale (MS)-SSIM [45] as objective evaluation metrics.

$$PSNR = 20log_{10}(\frac{I_{max}}{\sqrt{MSE}}) \quad (12)$$

The mean-square-error, also called MSE, is defined as below:

$$\text{MSE} = \frac{1}{MN}\sum_{m=0}^{M-1}\sum_{n=0}^{N-1}(I_{GT}(m,n) - I_R(m,n))^2 \quad (13)$$

where I_{GT} and I_R are the ground truth and the reconstructed OCTA images, respectively. The term I_{max} is set as 1 in this evaluation, which refers to the maximum value in the image. The SSIM evaluates image quality in terms of structural similarity. A Higher SSIM shows a better structural similarity of model outputs to ground-based real-world data.

$$SSIM = \frac{(2\mu_{GT}\mu_R + k1)(2\sigma_{cov} + k2)}{(\mu^2{}_{GT} + \mu^2{}_R + k1)(\sigma^2{}_{GT} + \sigma^2{}_R + k2)} \quad (14)$$

Here, μ_{GT} and (σ_{GT}) and μ_R and (σ_R) 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.

Additionally, to offer a more comprehensive analysis of the vasculature extraction performance, we utilized enface OCTA images generated using the maximum intensity projection (MIP) for visual comparison. These enface OCTA images were compared against a baseline image (Fig. 4 purple blocks) to assess the performance of the methods in terms of vascular connectivity and vasculature extraction. This visual evaluation approach provided an additional perspective to complement the quantitative analysis, allowing for a more nuanced and accurate assessment of the extraction methods.

IV. Results

After training all of the networks including the proposed VET model and compared-used networks, we then applied them to extract vascular signals from a set of test data. The quantitative comparison is based on the cross-sectional images from the healthy test set and the disease test set, and the visual comparison is based on the enface OCTA images generated by different methods. In this section, we discuss the advantages of using neural networks for single-scan OCTA image generation.

A. Quantitative Comparison between Various Methods

Table 2 demonstrates a quantitative comparison of different methods, with all methods improving the image quality of single-repeated structural OCT images in terms of PSNR, SSIM, and MS-SSIM performance. The ED-OCTA method with four repetitions achieves the best performance in terms of SSIM (0.465) and MS-SSIM (0.702) in the healthy test set. Among the results from the various methods, VET has the highest PSNR performance in the healthy test set (18.03) and disease test set (17.42), while the FLOPs is the fourth smallest (27.57G). In terms of the SSIM performance, Restormer has the best performance in the healthy test set (0.340), and Swin-UNet has the highest results in the disease test set (0.220). Regarding the MS-SSIM, Restormer has the best performance in the health test set (0.592) and SwinIR is the best (0.525) in the disease test set. In the disease test set, the PSNR performance between the VET (17.42) and SwinIR (17.40) is similar while VET has relatively lower FLOPs (27.57G < 103.5G). In the comparison between Restormer and VET, the Restormer has a better SSIM (0.340 > 0.328) than the VET in the healthy test set; however, the VET can provide a better SSIM (0.212 > 0.208) performance than Restormer in the disease test set, while the FLOPs is approximately 5 times smaller (27.57G < 142.7G). Besides, the SSIM and MS-SSIM performance among the Swin-UNet (0.220, 0.516), SwinIR (0.211, 0.525), Restormer (0.208, 0.516), UFormer (0.209, 0.520), and VET (0.212, 0.519) are similar, and the difference between them is slight, while the Swin-UNet has the lowest FLOPs (16.12G) and VET is the second-lowest FLOPs (27.57G).

B. Visual Comparison Result

Visual results of vasculature extraction by the different methods in various positions, including palm, face (with acne), and lip (health and with ulcer), are demonstrated in this section. The visual comparison and quantitative comparison between the different methods are based on enface images generated by the maximum intensity projection method.

Fig. 5 demonstrates the visual results based on the skin palm area. The result generated by ED-OCTA (C) has fewer micro-vasculature details, while the SV-OCTA (D) presents more micro-vasculature details. In terms of network performance, results from DnCNN (E), SRGAN (F), U-Net (H), and Trans-UNet (I) have good vascular connectivity, but none of them have an SSIM performance higher than 0.25, and PSNR higher than 13.0. The result from ESRGAN (G) has a poor vasculature extraction performance, based on the visual observations and quantitative results (i.e., lowest SSIM: 0.217, and lowest PSNR: 11.671). Among them, the results from Swin-UNet (K), Swin-IR (M), UFormer (L), Restormer (N), and VET (O) have a relatively better vascular connectivity and vessel contrast, based on visual observation. The results from the VET (O) have the best quantitative results (PSNR: 0.342; SSIM: 15.132).

Fig. 6 is a visual comparison based on the face area with acne. The acne area is marked with a red-dot-line box in the structural image (A), thereby the acne area contains fewer vasculature signals in the ground truth (B). In Fig. 6, the result generated by ED-OCTA (C) has fewer vasculature details than SV-OCTA (D). In the comparison between the neural network results, the SRGAN (F), ESRGAN (G), U-Net (H), and Swin-UNet (K) have relatively poor vascular connectivity and vasculature details; furthermore, the boundary between the acne area and the nearby normal area is hard to classify based on visual observation. Among them, the results from DnCNN (E), Swin-IR (M), Restormer (N), and VET (O) have a clearer and relatively better vasculature extraction result in terms of visual observation. The VET (O) has the highest SSIM (0.251) and PSNR (12.399) in this face acne comparison group.

Fig. 7 demonstrates the vasculature extraction results based on a normal lip subject, and Fig. 8 is a visual comparison of results based on lip ulcer. In Fig. 7, the motion artifact of ground-truth (B) is relatively higher than the results in Fig. 5. Those artifacts are presented as bright light artifacts in the results (e.g., red allows in Fig. 7 (B)). In this stage, the results from neural networks (i.e., (E)-(O)) perform a better vasculature extraction than ED-OCTA (i.e., (C)) while reducing the bright line artifacts, based on visual observations. Among them, the DnCNN (E), SRGAN (F), SwinIR (M), Restormer (N), and VET (O) can provide relatively more vasculature details and clearer enface OCTA images. The VET (O) has the highest SSIM (0.388) and the second-highest PSNR (13.228).

In Fig. 8, the ulcer area is marked with a red dot box in the structural image (A), and the ground-truth (B) results show the site with fewer vasculature details. Expect the ESRGAN (G), the edge of the lip ulcer is clear in all neural network results. The results by SRGAN (F), ESRGAN (G), U-Net (H), LUSwin -T (J), and Swin-UNet (K) show relatively fewer vasculature details in the right bottom of the enface OCTA image. The results from SwinIR (M), Restormer (N), and VET (O) perform less noise, better vascular connectivity, and more vasculature details. The VET (O) has the highest SSIM (0.469), and SwinIR (M) has the highest PSNR (14.485).

Table 2. Quantitative Comparison of the vasculature images (Mean ± Standard Deviation) Extracted by Different Methods (Bold means the highest numerical value among the neural network results. Params means the parameters of each network. R stands for repetitions of scan)

Methods	Params (M)	FLOPs (G)	R	Healthy Test Set			Disease Test Set		
				PSNR	SSIM	MS-SSIM	PSNR	SSIM	MS-SSIM
Inputs	N/A	N/A	1	10.11±0.89	0.106±0.114	0.141±0.162	11.58±0.58	0.102±0.031	0.255±0.047
SV-OCTA [15]	N/A	N/A	4	16.13±0.68	0.278±0.030	0.591±0.056	15.63±0.68	0.257±0.024	0.611±0.038
ED-OCTA [16]	N/A	N/A	4	17.08±1.50	0.465±0.108	0.702±0.102	15.81±0.91	0.384±0.029	0.682±0.038
DnCNN [36]	0.557	40.92	1	17.05±1.36	0.318±0.063	0.502±0.071	16.20±0.85	0.180±0.057	0.449±0.072
SRGAN [38]	0.567	41.68	1	17.10±1.40	0.319±0.068	0.524±0.075	16.31±0.79	0.170±0.056	0.457±0.081
ESRGAN [39]	3.506	258.5	1	17.45±0.87	0.318±0.062	0.505±0.065	16.78±0.73	0.179±0.055	0.451±0.072
U-Net [37]	34.56	59.88	1	17.30±1.06	0.302±0.069	0.550±0.108	16.98±0.72	0.196±0.059	0.494±0.076
TransUNet [40]	52.35	23.01	1	16.86±0.95	0.292±0.059	0.525±0.082	16.59±0.72	0.199±0.058	0.491±0.082
LUSwin-T [43]	11.92	3.930	1	17.28±1.00	0.288±0.062	0.548±0.081	17.32±0.78	0.208±0.067	0.513±0.088
Swin-UNet [27]	50.28	16.12	1	16.87±1.03	0.268±0.055	0.519±0.086	17.26±0.79	**0.220±0.067**	0.516±0.082
UFormer [41]	24.38	38.89	1	17.64±1.12	0.321±0.063	0.576±0.075	17.10±0.79	0.209±0.064	0.520±0.084
SwinIR [26]	1.739	103.5	1	17.83±1.05	0.328±0.058	0.587±0.079	17.40±0.73	0.211±0.062	**0.525±0.078**
Restormer [42]	16.24	142.7	1	17.79±1.08	**0.340±0.059**	**0.592±0.084**	17.12±0.74	0.208±0.059	0.516±0.070
VET (ours)	0.929	27.57	1	**18.03±1.07**	0.328±0.059	0.576±0.077	**17.42±0.65**	0.212±0.061	0.519±0.074

Table 3. Quantitative Comparison of the Different VET Parameters Setup. (Bold means the highest numerical value among the neural network results. Params means the parameters of each network).

Ablation Study	Filter Size	Head	VFE Layer	Params (M)	FLOPs (G)	Healthy Test Set			Disease Test Set		
						PSNR	SSIM	MS-SSIM	PSNR	SSIM	MS-SSIM
VFE Layer	64	4	2	0.485	15.28	17.46±1.13	0.314±0.057	0.518±0.091	17.07±0.63	0.186±0.059	0.478±0.076
			4	0.929	27.57	18.03±1.07	0.328±0.059	0.576±0.077	17.42±0.65	0.212±0.061	0.519±0.074
			6	1.374	39.85	18.06±1.14	0.333±0.045	0.584±0.066	17.37±0.67	0.199±0.061	0.510±0.073
			8	1.818	52.13	**18.11±1.43**	**0.339±0.062**	**0.594±0.085**	17.25±0.68	0.194±0.060	0.515±0.073
Heads	64	2	4	0.929	23.49	18.01±1.31	0.336±0.062	0.569±0.101	17.12±0.63	0.191±0.059	0.492±0.074
		4		0.929	27.57	18.03±1.07	0.328±0.059	0.576±0.077	17.42±0.65	0.212±0.061	0.519±0.074
		8		0.929	35.72	17.90±1.15	0.303±0.049	0.572±0.083	17.35±0.64	0.196±0.059	0.503±0.073
		16		0.929	52.03	17.76±1.06	0.289±0.055	0.537±0.072	16.83±0.87	0.197±0.061	0.489±0.076
Filter Size	32	4	4	0.234	19.15	17.54±1.38	0.312±0.056	0.541±0.086	17.22±0.67	0.197±0.060	0.507±0.073
	48			0.524	22.23	17.65±1.07	0.317±0.057	0.553±0.098	17.35±0.66	0.186±0.060	0.493±0.074
	64			0.929	27.57	18.03±1.07	0.328±0.059	0.576±0.077	**17.42±0.65**	**0.212±0.061**	**0.519±0.074**
	80			1.450	44.00	18.04±1.18	0.329±0.056	0.576±0.075	17.38±0.69	0.201±0.062	0.513±0.074

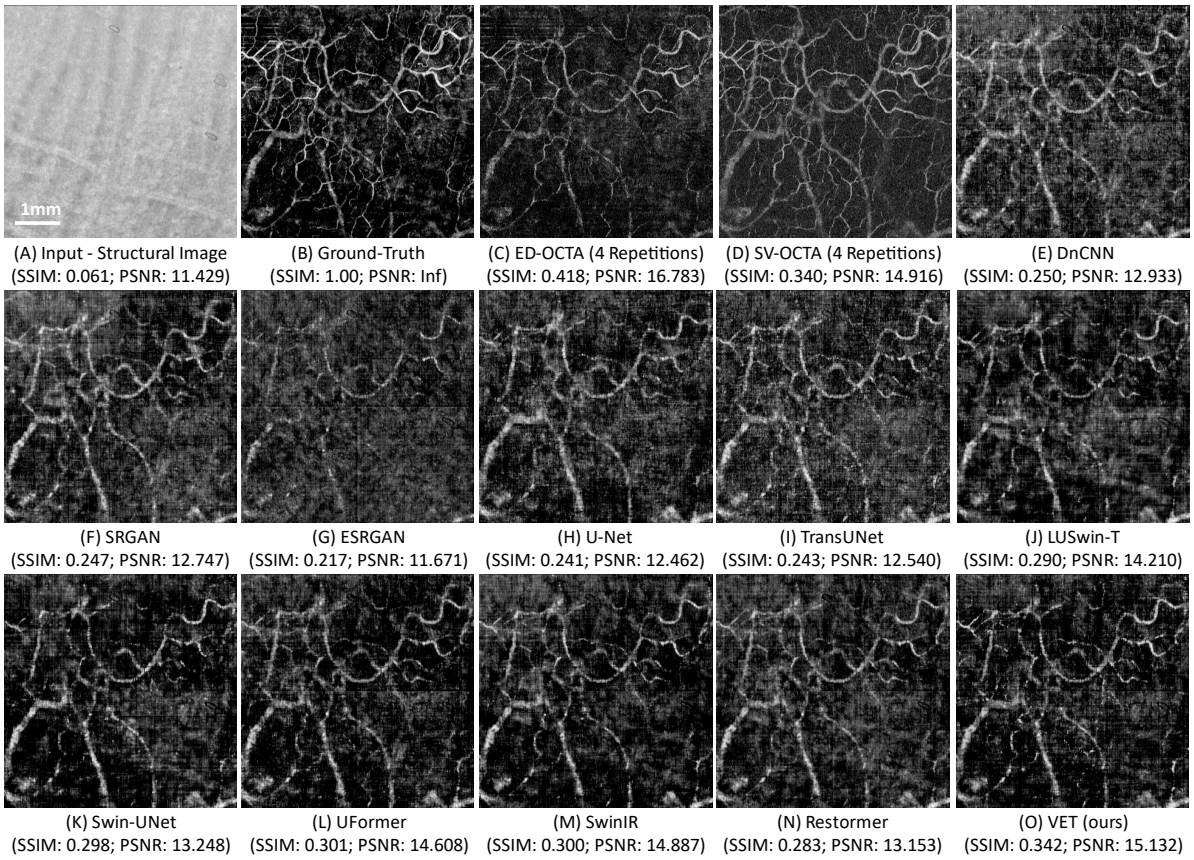

Fig. 5. Visual comparison of the hand-held skin palm area. (A) to (O) are enface OCTA images of Input structural image (A), Ground-truth (B), ED-OCTA with four-repeated scan (C), SV-OCTA with four-repeated scan (D), DnCNN (E), SRGAN (F), ESRGAN (G), UNet (H), TransUNet (I), LUSwin-T (J), Swin-UNet (K), UFormer (L), SwinIR (M), Restormer (N), and VET (O). The white scale bar is 1 mm.

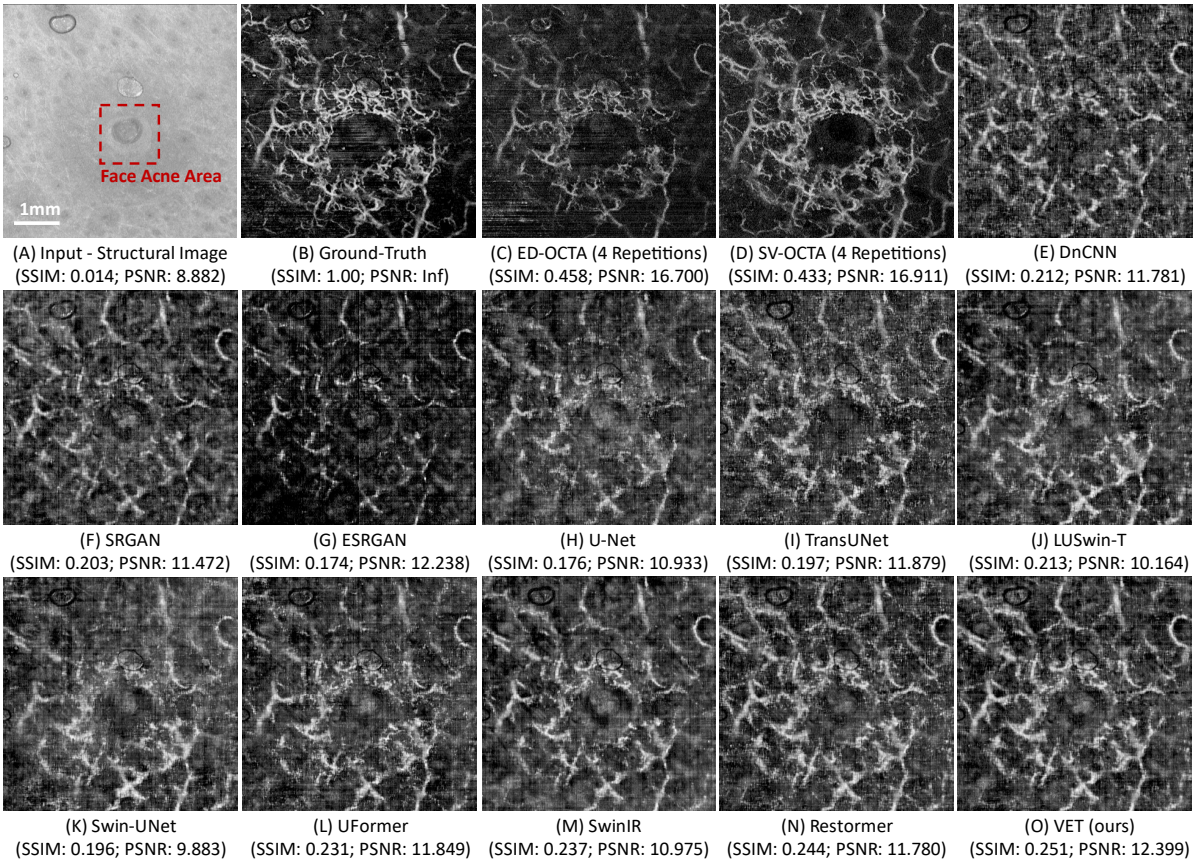

Fig. 6. Visual comparison of the hand-held skin face area with acne. (A) to (O) are enface OCTA images of Input structural image (A), Ground-truth (B), ED-OCTA with four-repeated scan (C), SV-OCTA with four-repeated scan (D), DnCNN (E), SRGAN (F), ESRGAN (G), UNet (H), TransUNet (I), LUSwin-T (J), Swin-UNet (K), UFormer (L), SwinIR (M), Restormer (N), and VET (O). The white scale bar is 1 mm.

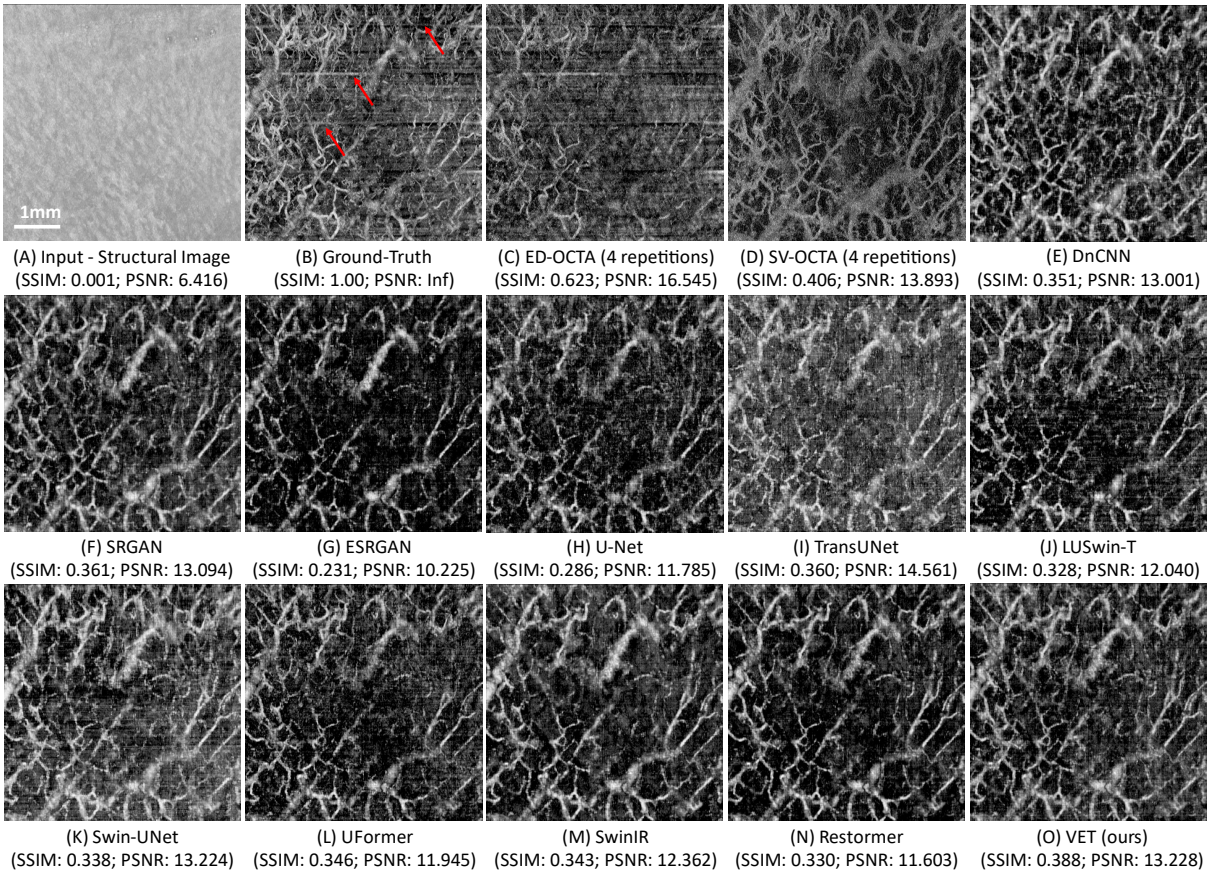

Fig. 7. Visual Comparison of the healthy lip area. (A) to (O) are enface OCTA images of Input structural image (A), Ground-truth (B), ED-OCTA with four-repeated scan (C), SV-OCTA with four-repeated scan (D), DnCNN (E), SRGAN (F), ESRGAN (G), UNet (H), TransUNet (I), LUSwin-T (J), Swin-UNet (K), UFormer (L), SwinIR (M), Restormer (N), and VET (O). The white scale bar is 1 mm. Red allows are used to point out the bright light artifact area.

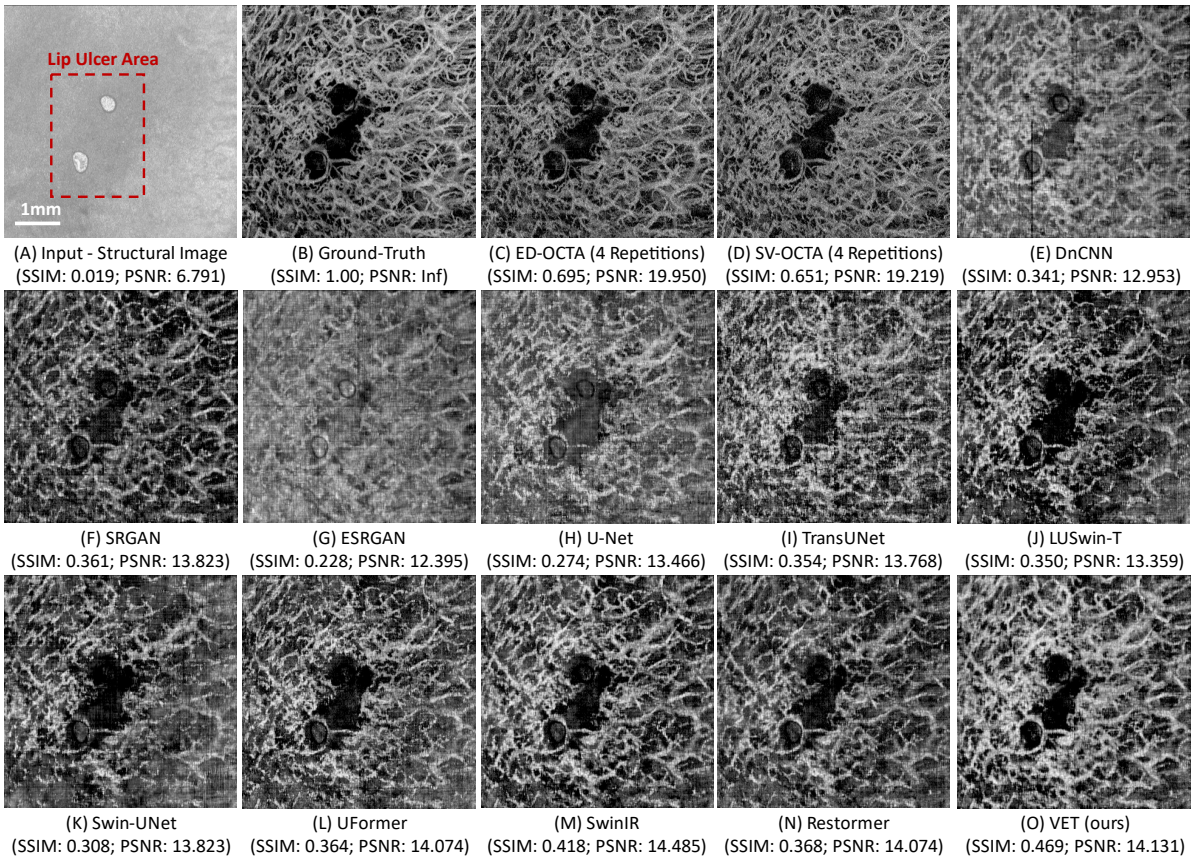

Fig. 8. Visual Comparison of the lip with ulcer area. (A) to (O) are enface OCTA images of Input structural image (A), Ground-truth (B), ED-OCTA with four-repeated scan (C), SV-OCTA with four-repeated scan (D), DnCNN (E), SRGAN (F), ESRGAN (G), UNet (H), TransUNet (I), LUSwin-T (J), Swin-UNet (K), UFormer (L), SwinIR (M), Restormer (N), and VET (O). The white scale bar is 1 mm.

C. *Ablation Study*

Table 3 shows the quantitative results of the ablation study based on the VET. In the healthy test set, the study shows that the more VFE layers used in the VET model, the higher the performance of the VET model (PSNR from 17.46 to 18.11, and SSIM from 0.314 to 0.339). Furthermore, the performance of VET is proportional to the filter sizes (PSNR from 17.54 to 18.04, and SSIM from 0.312 to 0.329). Regarding the influence of head number, compared with the control group that has a head number of 4, the smaller or higher the numerical value of the head will decrease the performance of VET. Distinct from the results in the healthy test, in the disease test set, the increase of the VFE layer and filter size will decrease the performance in terms of PSNR (VFE layer: from 17.42 to 17.25; Filter Size: from 17.42 to 17.38) and SSIM (VFE layer: from 0.212 to 0.194; Filter Size: from 0.0.212 to 0.201). The proposed control group has the best PSNR (17.42) and SSIM (0.212) performance in the disease test set while maintaining relatively low FLOPs (27.57G).

V. Discussion

In this study, we present a fast end-to-end vasculature extraction pipeline based on a single *in-vivo* skin OCT scan that requires only ~2 s for data acquisition. Our pipeline employs a novel vasculature extraction transformer (VET), which provides moderate quality OCTA images with a single OCT scan, as opposed to conventional OCTA algorithms like ED-OCTA and SV-OCTA that necessitate at least two-repeated scans. With the proposed single-scan OCTA pipeline, real-time OCTA imaging is available by utilizing the trained VET to extract vascular signals from each single-scan OCT structure signal. Furthermore, the single-scan OCTA pipeline is faster than the ED-OCTA algorithm since it does not require eigen decomposition calculations and can be directly processed on the graphics processing unit (GPU). Notably, the VET utilizes convolutional projection to generate query, key, and value sequences for multi-head self-attention computations, preserving spatial relationships between image patches better than fully connected layers used in Trans-UNet, SwinIR, and Swin-UNet. The results exhibit that our proposed pipeline has significant potential for clinical applications, as it reduces motion artifacts and accelerates imaging speed by reducing the repeated scan of OCTA imaging.

Table 2 demonstrates a quantitative comparison of different methods. Compared with ED-OCTA with four-repeated scans, the VET model can provide a higher PSNR performance (18.03 > 17.08), but the SSIM (0.328 < 0.465) and MS-SSIM (0.576 < 0.702) results are lower than ED-OCTA. Among deep learning-based approaches, our VET model strikes a balance between the number of network parameters (0.929M), FLOPs (26.57G), and performance metrics (PSNR: 18.03 in healthy test set, and 17.42 in disease test set). Although Restormer has the highest SSIM (0.340) and the highest MS-SSIM (0.592), Restormer has high FLOPs (142.7G), compared with the models used in this study. In the disease test set, mostly transformer-based models (i.e., LUSwin-T, Swin-UNet, UFormer, SwinIR, Restormer, and VET) outperform

CNN-based models (i.e., DnCNN, SRGAN, ESRGAN, U-Net) in terms of PSNR, SSIM, and MS-SSIM metrics. Regarding network architecture, end-to-end architectures (e.g., SwinIR, VET) achieve relativity higher PSNR performance than encoder-decoder architectures (e.g., Swin-UNet, UFormer). Nevertheless, transformer-type models with encoder-decoder architectures offer smaller FLOPs.

Visual inspection of Fig. 5 reveals that most transformer-type models (i.e., Swin-UNet, UFormer, SwinIR, Restormer, and VET) can correctly extract the vasculature signals from the single-scan OCT signals, and provide good vasculature details and vascular connectivity. Compared to ground truth which was generated by twelve repetitions of scan, VET has the highest SSIM (0.342) and PSNR (15.132) performance. Fig. 6 presents vasculature extraction performance based on an abnormal face acne area. In the results generated by SwinIR, Restormer, and VET, the boundary between the acne and the normal area is clear to classify based on visual observation. In this stage, the trained model is proof that can classify abnormal areas.

Fig. 7 and Fig. 8 present vasculature extraction results based on normal lip and lip ulcer areas. In Fig. 7, results from the networks exhibit fewer motion artifacts than the ground-truth, which uses six-repeated OCT scans, as motion artifacts due to the scanning probe and participants lead to low-quality OCTA images. Moreover, the results from SwinIR, UFormer, and VET show better vasculature extraction and connectivity than the ground truth, based on visual performance. In Fig. 8, except for the ESRGAN, all results from neural networks provide a clear boundary of the lip ulcer area. Among the neural networks, UFormer, SwinIR, Restormer, and VET can provide relatively more vasculature details (e.g., in the right-bottom of enface OCTA images). Besides, VET has the highest SSIM (0.469) and PSNR (14.131), compared to ground truth.

Our study has limitations. First, the performance of the proposed VET model may be impacted when using OCT data from additional diseased subjects, as our training data is from healthy participants. In the future, we plan to collect skin OCTA data from participants with various skin conditions and investigate the vasculature extraction pipeline for both healthy and diseased OCT data. Second, we did not apply adversarial training (e.g., generative adversarial network (GAN) [46]) to the VET model training, as it is challenging and can lead to unstable training. We aim to further explore adversarial training for the VET model using conditional GAN [47] and relativistic average (Ra)-GAN [48] to enhance vasculature extraction performance. Thirdly, based on the visual result comparison in this study, all neural network results struggle with micro-vasculature extraction when compared to high-quality ground truth images. In the future, we aim to develop a new architecture that can better extract micro-vasculature features by combining the advantages of local convolution layer and global attention mechanisms.

VI. Conclusion

In this study, we propose an end-to-end vasculature extraction pipeline and VET model that only uses a single OCT scan, demonstrating promising results for clinical applications. The VET model outperforms other deep-learning approaches in terms of efficiency (FLOPs: 27.57G) and performance metrics (PSNR: 18.03 in the healthy set, and 17.42 in the disease set). Despite the limitations in this study, our findings indicate that the proposed pipeline significantly reduces data acquisition time by 75%, while providing similar high-quality OCTA images compared to those obtained by the conventional ED-OCTA algorithm with four-repeated OCT scans. This makes it a valuable tool for fast skin OCTA imaging in clinical settings. In terms of network generalization and robustness, the VET consistently performs stable vasculature extraction across different positions (e.g., face, and lip) with varying conditions. In future work, we plan to introduce this fast OCTA scan pipeline to retinal scans, aiming to achieve high-quality OCTA imaging with minimal motion artifacts and rapid acquisition.

References

[1] A. J. Deegan and R. K. Wang, "Microvascular imaging of the skin," *Phys Med Biol*, vol. 64, no. 7, p. 07TR01, 2019.

[2] M. Roustit and J.-L. Cracowski, "Assessment of endothelial and neurovascular function in human skin microcirculation," *Trends Pharmacol Sci*, vol. 34, no. 7, pp. 373–384, 2013.

[3] L. A. Holowatz, C. S. Thompson-Torgerson, and W. L. Kenney, "The human cutaneous circulation as a model of generalized microvascular function," *J Appl Physiol*, vol. 105, no. 1, pp. 370–372, 2008.

[4] M. P. De Boer *et al.*, "Microvascular dysfunction: a potential mechanism in the pathogenesis of obesity-associated insulin resistance and hypertension," *Microcirculation*, vol. 19, no. 1, pp. 5–18, 2012.

[5] H. Debbabi, L. Uzan, J. J. Mourad, M. Safar, B. I. Levy, and E. Tibiriçà, "Increased skin capillary density in treated essential hypertensive patients," *Am J Hypertens*, vol. 19, no. 5, pp. 477–483, 2006.

[6] R. G. IJzerman *et al.*, "Individuals at increased coronary heart disease risk are characterized by an impaired microvascular function in skin," *Eur J Clin Invest*, vol. 33, no. 7, pp. 536–542, 2003.

[7] S. E. Kaiser, A. F. Sanjuliani, V. Estato, M. B. Gomes, and E. Tibiriçá, "Antihypertensive treatment improves microvascular rarefaction and reactivity in low-risk hypertensive individuals," *Microcirculation*, vol. 20, no. 8, pp. 703–716, 2013.

[8] R. K. Wang, Q. Zhang, Y. Li, and S. Song, "Optical coherence tomography angiography-based capillary velocimetry," *J Biomed Opt*, vol. 22, no. 6, p. 066008, 2017.

[9] A. Zhang, Q. Zhang, C.-L. Chen, and R. K. Wang, "Methods and algorithms for optical coherence tomography-based angiography: a review and comparison," *J Biomed Opt*, vol. 20, no. 10, p. 100901, 2015.

[10] A. S. Nam, I. Chico-Calero, and B. J. Vakoc, "Complex differential variance algorithm for optical coherence tomography angiography," *Biomed Opt Express*, vol. 5, no. 11, pp. 3822–3832, 2014, doi: 10.1364/BOE.5.003822.

[11] B. Zabihian *et al.*, "Comprehensive vascular imaging using optical coherence tomography-based angiography and photoacoustic tomography," *J Biomed Opt*, vol. 21, no. 9, p. 096011, 2016.

[12] A. J. Deegan *et al.*, "Optical coherence tomography angiography of normal skin and inflammatory dermatologic conditions," *Lasers Surg Med*, vol. 50, no. 3, pp. 183–193, 2018.

[13] Y. Ji, K. Zhou, S. H. Ibbotson, R. K. Wang, C. Li, and Z. Huang, "A novel automatic 3D stitching algorithm for optical coherence tomography angiography and its application in dermatology," *J Biophotonics*, vol. 14, no. 11, p. e202100152, 2021.

[14] L. Themstrup, G. Pellacani, J. Welzel, J. Holmes, G. B. E. Jemec, and M. Ulrich, "In vivo microvascular imaging of cutaneous actinic keratosis, Bowen's disease and squamous cell carcinoma using dynamic optical coherence tomography," *Journal of the European Academy of Dermatology and Venereology*, vol. 31, no. 10, pp. 1655–1662, 2017.

[15] A. Mariampillai *et al.*, "Speckle variance detection of microvasculature using swept-source optical coherence tomography," *Opt Lett*, vol. 33, no. 13, pp. 1530–1532, 2008.

[16] S. Yousefi, Z. Zhi, and R. K. Wang, "Eigendecomposition-based clutter filtering technique for optical microangiography," *IEEE Trans Biomed Eng*, vol. 58, no. 8, pp. 2316–2323, 2011.

[17] M. A. Linares, A. Zakaria, and P. Nizran, "Skin cancer," *Primary care: Clinics in office practice*, vol. 42, no. 4, pp. 645–659, 2015.

[18] B. Baumann *et al.*, "Signal averaging improves signal-to-noise in OCT images: But which approach works best, and when?," *Biomed. Opt. Express*, vol. 10, no. 11, pp. 5755–5775, Nov. 2019, doi: 10.1364/BOE.10.005755.

[19] X. Liu *et al.*, "A deep learning based pipeline for optical coherence tomography angiography," *J Biophotonics*, vol. 12, no. 10, p. e201900008, 2019.

[20] Z. Jiang *et al.*, "Weakly supervised deep learning-based optical coherence tomography angiography," *IEEE Trans Med Imaging*, vol. 40, no. 2, pp. 688–698, 2020.

[21] Z. Jiang *et al.*, "Comparative study of deep learning models for optical coherence tomography angiography," *Biomed Opt Express*, vol. 11, no. 3, pp. 1580–1597, 2020.

[22] H. Wu *et al.*, "Cvt: Introducing convolutions to vision transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 22–31.

[23] J. Xu *et al.*, "Deep-learning visualization enhancement method for optical coherence tomography angiography in dermatology," *J Biophotonics*, p. e202200366, 2023.

[24] M. Pan, Y. Wang, P. Gong, Q. Wang, and B. Cense, "Feasibility of deep learning-based polarization-sensitive optical coherence tomography angiography for imaging cutaneous microvasculature," *Biomed Opt Express*, vol. 14, no. 8, pp. 3856–3870, 2023.

[25] 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*, Apr. 2023, doi: 10.1364/BOE.486933.

[26] 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.

[27] H. Cao *et al.*, "Swin-unet: Unet-like pure transformer for medical image segmentation," *arXiv preprint arXiv:2105.05537*, 2021.

[28] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[29] Z. Liu *et al.*, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10012–10022.

[30] T. Xiao, M. Singh, E. Mintun, T. Darrell, P. Dollár, and R. Girshick, "Early convolutions help transformers see better," *Adv Neural Inf Process Syst*, vol. 34, pp. 30392–30400, 2021.

[31] Y. Ji *et al.*, "Deep-learning approach for automated thickness measurement of epithelial tissue and scab using optical coherence tomography," *J Biomed Opt*, vol. 27, no. 1, p. 015002, 2022. [1]

[32] Y. Cheng, Z. Chu, and R. K. Wang, "Robust three-dimensional registration on optical coherence tomography angiography for speckle reduction and visualization," *Quant Imaging Med Surg*, vol. 11, no. 3, p. 879, 2021.

[33] 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*, vol. 29, no. 1, pp. 196–205, 2009.

[34] Q. Zhang, J. Wang, and R. K. Wang, "Highly efficient eigen decomposition based statistical optical microangiography," *Quant Imaging Med Surg*, vol. 6, no. 5, p. 557, 2016.

[35] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[36] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising," *IEEE transactions on image processing*, vol. 26, no. 7, pp. 3142–3155, 2017.

[37] 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.

[38] C. Ledig *et al.*, "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.

[39] X. Wang *et al.*, "Esrgan: Enhanced super-resolution generative adversarial networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, p. 0.

[40] J. Chen *et al.*, "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.

[41] Z. Wang, X. Cun, J. Bao, W. Zhou, J. Liu, and H. Li, "Uformer: A general u-shaped transformer for image restoration," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 17683–17693.

[42] 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.

[43] J. Liao, C. Li, and Z. Huang, "A Lightweight Swin Transformer-Based Pipeline for Optical Coherence Tomography Image Denoising in Skin Application," *Photonics*, vol. 10, no. 4, 2023, doi: 10.3390/photonics10040468.

[44] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[45] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, Ieee, 2003, pp. 1398–1402.

[46] I. Goodfellow *et al.*, "Generative adversarial nets," *Adv Neural Inf Process Syst*, vol. 27, 2014.

[47] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[48] A. Jolicoeur-Martineau, "The relativistic discriminator: a key element missing from standard GAN," *arXiv preprint arXiv:1807.00734*, 2018.