A hand-held optical coherence tomography angiography scanner based on angiography reconstruction transformer networks

Liao, Jinpeng; Yang, Shufan; Zhang, Tianyu; Li, Chunhui; Huang, Zhihong

Published in:
Journal of Biophotonics

DOI:
10.1002/jbio.202300100

Publication date:
2023

Licence:
CC BY

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (APA):
Liao, J., Yang, S., Zhang, T., Li, C., & Huang, Z. (2023). A hand-held optical coherence tomography angiography scanner based on angiography reconstruction transformer networks. *Journal of Biophotonics*, [e202300100]. https://doi.org/10.1002/jbio.202300100

Received: 24 March 2023 | Revised: 18 May 2023 | Accepted: 24 May 2023

DOI: 10.1002/jbio.202300100

RESEARCH ARTICLE

JOURNAL OF BIOPHOTONICS

A hand-held optical coherence tomography angiography scanner based on angiography reconstruction transformer networks

Jinpeng Liao[1] | Shufan Yang[2,3] | Tianyu Zhang[1] | Chunhui Li[1] | Zhihong Huang[1]

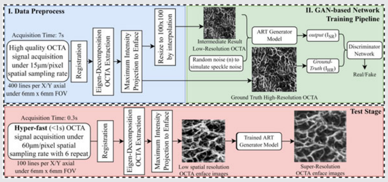
[1]School of Science and Engineering, University of Dundee, Scotland, UK

[2]School of Computing, Engineering and Built Environment, Edinburgh Napier University, Edinburgh, UK

[3]Research Department of Orthopaedics and Musculoskeletal Science, University College London, UK

Correspondence
Chunhui Li, School of Science and Engineering, University of Dundee, Scotland, UK.
Email: c.li@dundee.ac.uk

Abstract


I. Data Preprocess
Acquisition Time: 7s
High quality OCTA signal acquisition under 15μm/pixel spatial sampling rate
400 lines per X/Y axial under 6mm x 6mm FOV
Registration
Eigen-Decomposition OCTA Extraction
Maximum Intensity Projection to Enface
Resize to 100x100 by interpolation
II. GAN-based Network Training Pipeline
Intermediate Result Low-Resolution OCTA
Random noise (n) to simulate speckle noise
ART Generator Model
output (I_SR)
Ground-Truth (I_HR)
Discriminator Network
Real/Fake
Ground Truth High-Resolution OCTA
Test Stage
Acquisition Time: 0.3s
Hyper-fast (<1s) OCTA signal acquisition under 60μm/pixel spatial sampling rate with 6 repeat
100 lines per X/Y axial under 6mm x 6mm FOV
Registration
Eigen-Decomposition OCTA Extraction
Maximum Intensity Projection to Enface
Low spatial resolution OCTA enface images
Trained ART Generator Model
Super-Resolution OCTA enface images

Optical coherence tomography angiography (OCTA) has successfully demonstrated its viability for clinical applications in dermatology. Due to the high optical scattering property of skin, extracting high-quality OCTA images from skin tissues requires at least six-repeated scans. While the motion artifacts from the patient and the free hand-held probe can lead to a low-quality OCTA image. Our deep-learning-based scan pipeline enables fast and high-quality OCTA imaging with 0.3-s data acquisition. We utilize a fast scanning protocol with a 60 μm/pixel spatial interval rate and introduce angiography-reconstruction-transformer (ART) for 4× super-resolution of low transverse resolution OCTA images. The ART outperforms state-of-the-art networks in OCTA image super-resolution and provides a lighter network size. ART can restore microvessels while reducing the processing time by 85%, and maintaining improvements in structural similarity and peak-signal-to-noise ratio. This study represents that ART can achieve fast and flexible skin OCTA imaging while maintaining image quality.

KEYWORDS

deep learning, optical coherence tomography angiography, single image super-resolution

1 | INTRODUCTION

The skin microvasculature stores a large quantity of information on both cutaneous and systemic disorders [1]. Optical coherence tomography angiography (OCTA) is a noninvasive imaging modality, which is capable of providing vascular images at capillary-level resolution and can assist to identify disease by assessing the distribution of the vasculature rather than based on structural images [2]. In recent years, microvascular images provided by OCTA have proven to be of clinical important evidence in various skin diseases, such as acne [3], inflammatory disease (e.g., papule) [4], basal cell carcinoma, [5] actinic keratosis [6], and wound healing [7]. The lessons learned in the clinic are currently spurring a new set of advances in the laboratory that will expand the clinical use of OCTA by improving image quality and increasing acquisition speeds.

A commonly used method of generating OCTA scan is to acquire immediate succession changes of intensity over times of optical coherence tomography B-scans at the same location. The traditional OCTA extraction algorithms are based on the difference in phase-signal [8], intensity-signal [9], and complex-signal [10]. Since the connectivity of the microvasculature and the signal-to-noise ratio (SNR) of the OCTA images were directly proportional to the repeated number of OCT scans. Inadequate repeated time will deliver a high-level speckle noise [11]. In ophthalmology, a two-repeated scan is typically sufficient to produce high-quality retina OCTA images. However, in the case of skin OCTA scans, due to the intricate structure of the skin, a six-repeated scan strategy is necessary to strike a good compromise between the SNR and data acquisition time [12]. Nevertheless, with a 200 kHz A-scan rate swept-source-OCT scan device, the data acquisition time for a six-repeated OCT scan with 6 mm^2 FOV and 15 μm/pixel transverse resolution is ~7 s. Within those 7 s, the unexpected motion artifacts from the patient's movement are highly increased and lead to a low-quality OCTA image when applying an in vivo hand-held based OCTA scan in dermatology [13]. It is not effective if only increasing the swept rate of the light source at the data acquisition stage. Since the speckle noise will be increased with a shorter exposure time. Reducing the spatial sampling rate of the OCTA scan and recovering the high transverse resolution OCTA images from the low transverse resolution images by the neural network-based method can achieve a fast OCTA scan and maintain the six-repeated scan for skin application.

Alternatively, sparse representation matrix can recover the high-transverse resolution OCT structural image based on the low transverse images, but those methods require a complicated design of the regularizes for specific OCT data and are not suitable for OCTA applications because the capillaries flow signals are extracted from the reflected scattered signal of the red blood cells [14, 15]. Although prediction based algorithms (e.g., bicubic) are the extensive and fast methods to upsample the transverse resolution of the OCTA images, which had high applicability [16, 17], the quality of the upsample OCTA image from the prediction-based algorithms is unsatisfactory and the texture details (i.e., high-frequency signals) were lost.

A series of deep-learning-based methods were proposed to reconstruct the high-quality OCTA images based on the two-repeated OCT signals [18–20]. Those methods have achieved good results in OCTA image reconstruction while reducing the data acquisition time, but the data was acquired from the mice, which has a different domain from the skin OCT data. Furthermore, it is still challenging to extract the skin microvascular images from a two-repeated in vivo OCT signal. Regarding the single-image super-resolution (SISR) in the OCT data, a series of convolution neural network (CNN) models have proposed to super-resolve the OCT structural image resolution acquired by a low-spatial sampling rate, such as the U-Net [21] and SRResNet [22, 23]. However, these methods were specifically designed for super-resolving OCT structural images and thus, are not optimally suited for OCTA images. On the other hand, Kim et al. [24] demonstrated fast (~1.3 s) OCTA imaging by employing a CNN-based DenseReconstGAN to super-resolve low-resolution OCTA images. While CNN-based methods have shown competitive results in SISR for OCT, they inherently rely on the convolution operation, which restricts the receptive field (e.g., 3 × 3) and presents challenges in learning long-term information [25].

To facilitate the speed of in vivo skin OCTA scan and reduce the motion artifact, and maintain the six-repeated OCT scan, we propose an angiography reconstruction transformer (ART) to recover the high transverse resolution OCTA images from the low transverse resolution counterparts. The ART is based on the self-attention mechanism [25], which can provide long-term information and a large receptive field for feature extraction. In terms of the scan protocol, the FOV was maintained at 6 mm^2, and the transverse resolution was 15 and 60 μm/pixel for the high- and low-quality scan, separately. To maximize the quality of the resultant vascular image, the eigen-decomposition (ED)-OCTA method was used for data preprocessing, following an approach based on generative adversarial training to provide the adversarial loss and improve the performance of ART [26].

This paper's main contributions are (1) a new ART backbone network is proposed to form a deep-learning pipeline for a fast OCTA scan pipeline with high transverse resolution OCTA image reconstruction solutions. (2) To provide systematic quantity comparison for OCTA image reconstruction based on other CNN, transformer based neural networks. (3) To provide a better explain the ability of how a transformer block works, the visualization of the reconstruction convolutional transformer block in OCTA image reconstruction is reconstructed.

1.1 | Related work

1.1.1 | Convolution neural network (CNN)

In SISR, CNN-based methods were proven to have well performance and robustness, which achieve better super-resolved images than conventional interpolation methods [27–32]. Following the first proposed lightweight SRCNN to use trainable weights for image super-resolution [27], based on the residual learning strategy [33], the SRResNet [30], EDSR [28], and DRRN [31] increased the network depth and enhanced the performance of image super-resolution. To enhance the quality of the super-resolved images, ESRGAN [29] used the residual dense block and adversarial training in SISR for the natural images.

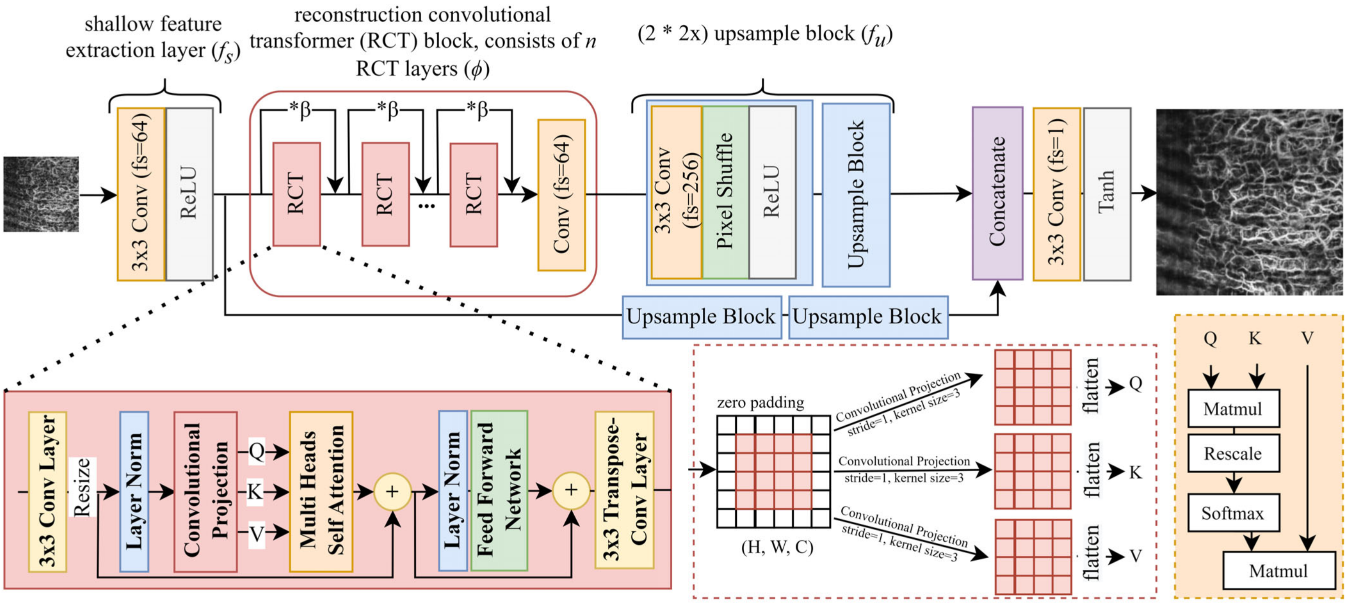
1.1.2 | Vision transformer

By flattening the two-dimension (2D) images into a one-dimension (1D) sequence of elements, Vision Transformer (ViT) has achieved a good competitive result in the Imagenet2K classification task [25]. Based on a pre-trained ViT model, IPT [34] proved that the transformer framework had the potential for image super-resolution. However, the sequences were generated by the nonoverlapping patch extraction layer, which reduced the position relation between the neighbor pixels. SwinIR [35] improved the SISR performance by using a hierarchical shift-window transformer [36], which introduce a local receptive field to the transformer block. However, the computing cost will be increased when using a larger size window in SwinIR for performance improvement. Although those methods achieved the prospective SISR results, the flattened projection for 1D sequences was done by a fully connected neural layer, which brings a high computing cost. Moreover, the training of the IPT and SwinIR required a lot of datasets (e.g., ImageNet2K, JFT300), which is impractical for the skin OCTA SISR because of the limited volume of datasets.

2 | METHOD

2.1 | Angiography reconstruction transformer (ART)

The architecture of ART model (as shown in Figure 1) consists of three parts: shallow feature extraction layer, reconstruction convolutional transformer block (RCTB), and upsample blocks.



(a) Modified Convolutional Transformer Layers (b) Convolutional Projection Layer (c) Self-Attention

FIGURE 1 The network architecture of the angiography reconstruction transformer (ART). The f_s is the filter size. The β is the parameter to control the residual scaling in the reconstruction convolutional transformer (RCT) block, and the n is the number of RCT layers inside the RCT block. All the convolution layers in ART have the same numerical value of f_s with the hidden size used in the RCT block. All convolution layers in ART have a filter size of 64 with strides 1. The blue block is upsample block, which is combined by a 3 × 3 convolution layer ($f_s = 256$) with strides 1, a pixel shuffle layer for tensor shape upsampling, and a ReLU activation layer. The output convolution layer ($f_s = 1$) is set as strides 1 and activated by a Tanh activation layer to enhance contrast. (A) is the modified convolutional transformer layer [37]. Before the layer normalization operation, a 3 × 3 convolution layer with strides 2 was used to downscale the size of feature maps. The Q, K, and V are then obtained by the convolutional projection (B) and then split with multi heads before being sent into the attention operation (C). The feed-forward network was combined with two groups of fully connected layers and the GELU activation layer. After processing by feed-forward network and add operation, a transpose convolution layer with 3 × 3 kernel size and strides 2 is used to upscale the shape of the feature map.

2.1.1 | Shallow feature extraction layer (f_s)

Shallow feature extraction layer (f_s) consists of a 3×3 convolution layer and a ReLU activation layer, which aims to extract the shallow feature (F_S) of low transverse resolution OCTA images (I_{LR}). The convolution layer has a stride of 1 and 64 filter sizes. Assumed the input is I_{LR}, the operation is:

$$F_s = ReLU(f_s(I_{LR})) \tag{1}$$

F_S was then used as an input of the RCT block, which consists of n RCT layers, and the output of each RCT layer was residual connected by multiple with a residual scaling parameter β [29]. A forward propagation for an RCT layer (f_ϕ) can be written as $f_\phi(F) = (\phi(F) * \beta + F)$, and F is the input of the RCT. And hence the processing of the RCT block is shown as (2):

$$F_n = f_c\left(f_{\phi_n}\left(\ldots f_{\phi_1}(F_s)\right)\right) \tag{2}$$

where F_n is the output of the RCT block, f_c is Conv-block, and n is the number of RCT layers in RCTB. Finally, F_n and F_s were simultaneously fed into the upsample blocks (f_u) to obtain the upscaled features (F_u). After being processed by a final layer and a Tanh activation layer, the predicted super-resolution OCTA images (I_{SR}) are obtained from ART.

$$I_{SR} = Tanh(f(concat(f_u(F_s), f_u(F_n)))) \tag{3}$$

2.1.2 | Reconstruction convolutional transformer (RCT)

The function of RCT is to extract the vascular texture feature, which is essential to OCTA image super-resolution. We modified the architecture proposed by Wu et al [37], as Figure 1(A) depicts, a 3 × 3 convolution layer with strides 2 is first applied to input features to reduce the tensor size, for computing cost reduction and improvement of the computing efficiency. Different from the transformer layer used in SwinIR [35] and ViT [25], the fully connected layer is not used to project the feature maps into 1D sequences in the RCT. As shown in Figure 1(B), the convolutional projection is then used to flatten the 2D feature maps to 1D sequences, which is aimed at increasing the related position information between the neighboring pixels. The output (i.e., Q, K, and V sequences) of convolutional projection is then split into multiheads and fed into the self-attention layer (Figure 1(C)) and processed by a feed-forward network. Finally, a deconvolution layer was used to expand the shape.

Figure 2 shows the heatmaps from the different RCT layers of the trained ART model with six RCT layers. In Figure 2(B,C) show that the first and third RCT blocks aim to extract the main backbone of the vascular texture details; (C) and (E) focus on the high-frequency details, which are concentrated to reconstruct a sharp OCTA image; the distribution of the weights in (F) and (G) is mainly on the micro-vessel, which benefit to the vessel connectivity reconstruction.

2.1.3 | Upsampling blocks

Pixel-shuffle layer and deconvolution layer were the two most used trainable upsample layers in SISR task, according to [38]. The deconvolution layer was widely used in the U-shape model for image reconstruction and denoizing [20], but the checkboard artifact was easy to appear if using the deconvolution layer for upsampling in SISR [39]. Hence, the pixel shuffle was introduced to ART as upsampling layers to reduce the checkboard artifacts and provide trainable weights for the mapping relationship between the I_{LR} and I_{HR}.

2.2 | OCTA pipelines

The overall pipeline consists of training and testing stages, as illustrated in Figure 3. In the training stage, a pre-processing strategy (blue zone) and a GAN-based training pipeline (green zone) for ART were performed. In this stage, the ground-truth OCTA images (I_{HR}) were acquired with a 15 μm/pixel spatial sample rate (with 6 mm^2 FOV and 6 repeated scans). The counterpart input OCTA images for ART training were then obtained by using a bilinear interpolation method to simulate the low transverse resolution images (i.e., 60 μm/pixel with 6 mm^2 FOV). An ED-OCTA algorithm is then used to separate the OCT signal into static structure signal, movement vessel signal, and noise signal [40], which shows less sensitivity to tissue motion and could suppress clutter efficiently [41]. The formulation of ED-OCTA is shown as follows:

$$E \wedge E^H = \sum\nolimits_{i=1}^{N} \lambda_B(i) e_B(i) e_B^H(i) \tag{4}$$

where $E = [e_B(1), e_B(2), \ldots, e_B(N)]$ is the $N \times N$ unitary matrix of eigenvectors, $\wedge = [\lambda_B(1), \lambda_B(2), \ldots, \lambda_B(N)]$ is the $N \times N$ diagonal matrix of eigenvalues, and H is the Hermitian transpose. The eigenvalues $\wedge$ are sorted in descending order. The structural signals are mainly at the

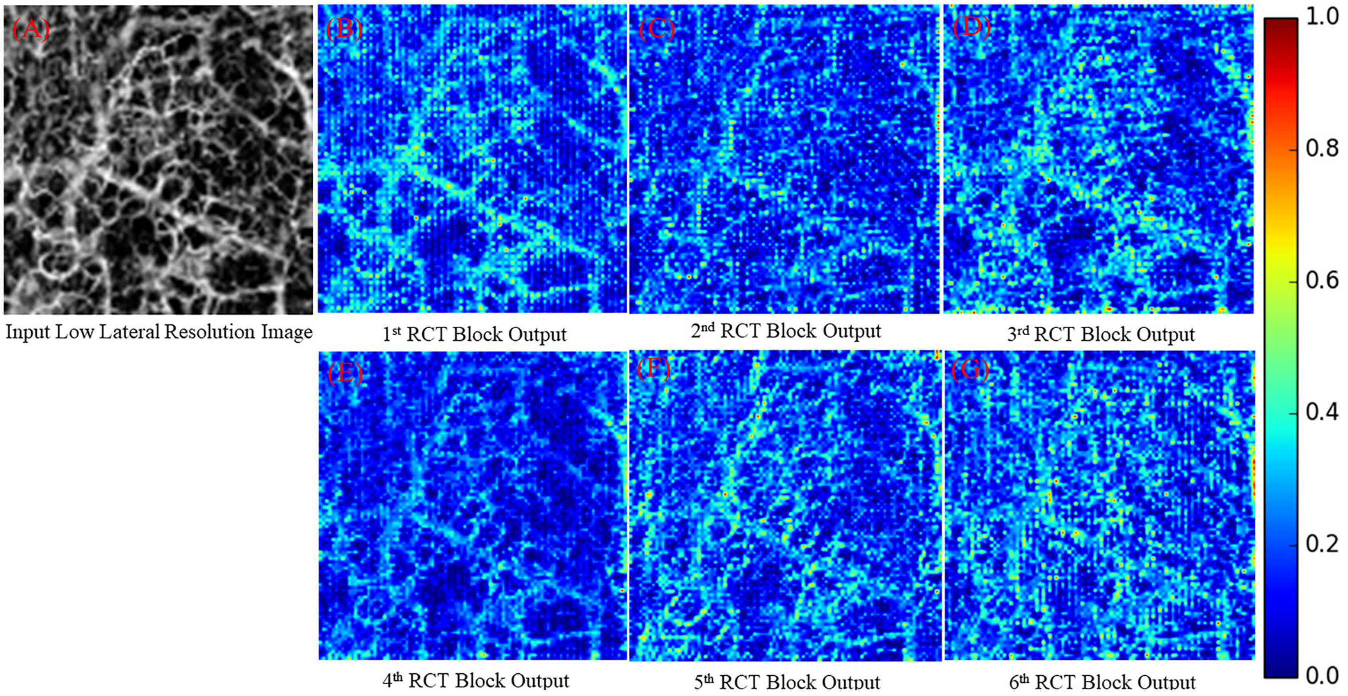


FIGURE 2 Heatmap of the different reconstruction convolutional transformer blocks from block 1 (B) to block 6 (G) in trained ART. (A) is the input low transverse resolution OCTA en face image. The range of the color bar is from 0 to 1, representing the weights of extracted features in the heatmap.

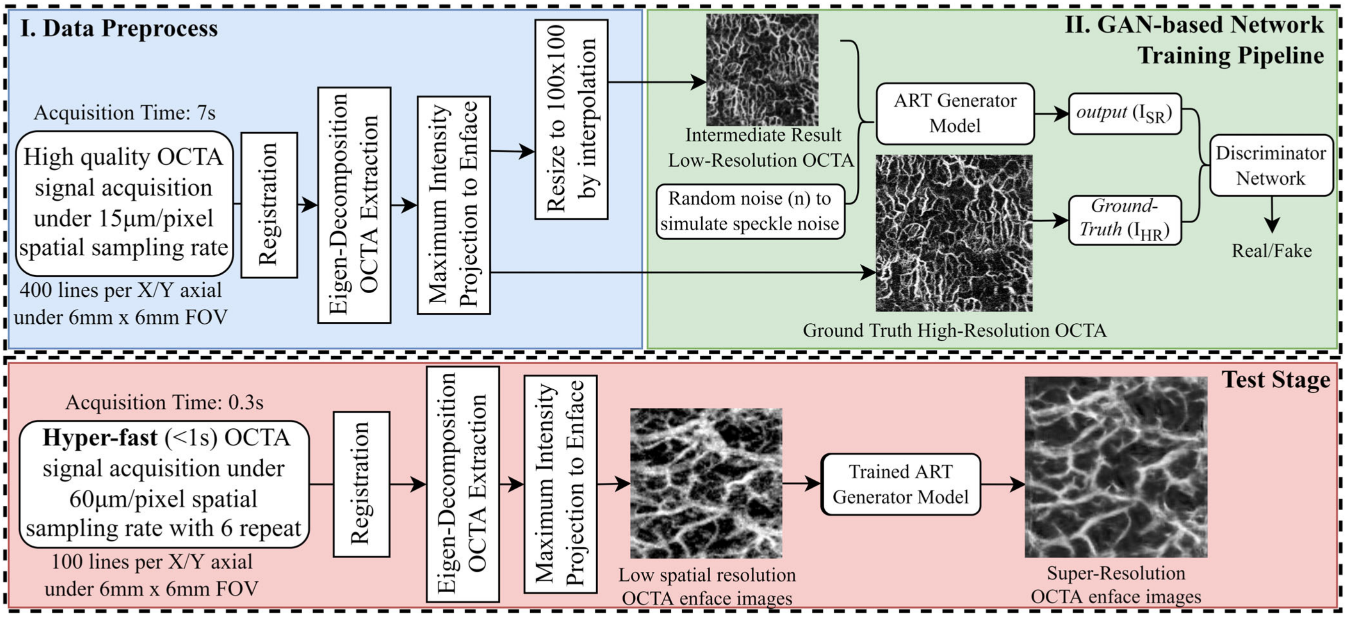

I. Data Preprocess
Acquisition Time: 7s
High quality OCTA signal acquisition under 15µm/pixel spatial sampling rate
400 lines per X/Y axial under 6mm x 6mm FOV
Registration
Eigen-Decomposition OCTA Extraction
Maximum Intensity Projection to Enface
Resize to 100x100 by interpolation
II. GAN-based Network Training Pipeline
Intermediate Result Low-Resolution OCTA
Random noise (n) to simulate speckle noise
ART Generator Model
output (I_SR)
Ground-Truth (I_HR)
Discriminator Network
Real/Fake
Ground Truth High-Resolution OCTA
Test Stage
Acquisition Time: 0.3s
Hyper-fast (<1s) OCTA signal acquisition under 60µm/pixel spatial sampling rate with 6 repeat
100 lines per X/Y axial under 6mm x 6mm FOV
Registration
Eigen-Decomposition OCTA Extraction
Maximum Intensity Projection to Enface
Low spatial resolution OCTA enface images
Trained ART Generator Model
Super-Resolution OCTA enface images

FIGURE 3 Stage-I: the data collection and preprocessing pipeline. Stage II: the GAN-based training pipeline for ART. Test Stage: In the proposed hyper-fast OCTA scan pipeline with trained neural networks, the OCTA acquisition time is 0.3 s in this scanning protocol.

first kth eigenvectors. The extraction of blood vessel signals can be written as (5):

$$X_v = \left[I - \sum_{i=1}^{K} e_B(i) e_B^H(i)\right] X \qquad (5)$$

where I is the identity matrix. The value of K is determined by the number of repeat scans ($k = 3$ in this study). $e_B(i)$ is the $1 \times N$ unitary matrix of eigenvectors. X_v are the vessel signals from the OCT signals after subtracting the static signals.

An adversarial training pipeline for the proposed ART to super-resolve OCTA images by learning the feature of high transverse resolution vascular from the ground truth. It consists of two networks with opposite goals: a generator (ART) that extracts the features of the I_{LR} and generates the corresponding I_{SR}; and a discriminator (i.e., VGG16 [42]) for binary classification, and tends to judge whether the synthetic image (I_{SR}) is from ART or not while taking the ground-truth image (I_{HR}) into account.

In the testing stage, the input of neural networks is the natural low transverse resolution OCT signal acquired under 60 μm/pixel spatial sampling rate with a 6 mm^2 FOV. After applying the volume registration and ED-OCTA algorithm mentioned in the last paragraph, a series of low-transverse resolution OCTA images are obtained. The high-quality and high transverse resolution OCTA images were then predicted by a trained ART model.

2.3 | Loss function

Although L_2 loss can provide a better peak-signal-to-noise ratio (PSNR) in SISR results, the L_1 loss was used in this study because L_1 loss results are less blurring than L_2 loss due to its force low-frequency correctness [43, 44]. And the L_1 loss is defined as Equation (6):

$$L_1(I_{HR}, I_{SR}) = \frac{1}{n}\sum_{i=1}^{n}|I_{HR} - I_{SR}| \tag{6}$$

where I_{HR} is the high-resolution ground-truth image, and I_{SR} is the output super-resolved OCTA image from the ART model. n is the total number of pixels in the image, and i means the $No.i$ pixel of the image. To better perform the high-frequency details of the super-resolved OCTA image, the VGG19-based perceptual loss was used [29, 35], and defined as follows:

$$L_p = \frac{1}{w} * \frac{1}{h}\sum_{i=1}^{h,w}(\varphi(I_{HR}) - \varphi(I_{SR}))^2 \tag{7}$$

where φ represents the block5 layer4 of the ImageNet pretrained VGG19 model. w and h are the weight and height of the feature maps generated by the φ, and i is the pixel number of the feature map. Besides the pixel-based and perceptual-based loss functions, an adversarial loss is introduced to improve the performance of ART. In the GAN, the calculation of adversarial loss is based on the properties of the input image is I_{HR} (real) or I_{SR} (fake), and the loss for the discriminator (L_D) and generator (L_G) were shown as follows:

$$\begin{aligned} L_D(I_{HR}, I_{SR}) = {} & \mathrm{E}_{I_{HR} \sim P_{data}(I_{HR})}(\log(D(I_{HR}))) \\ & + \mathrm{E}_{I_{SR} \sim P_{data}(I_{SR})}(\log(1 - D(I_{SR}))) \end{aligned} \tag{8}$$

$$L_G(I_{HR}, I_{SR}) = \mathrm{E}_{I_{SR} \sim P_{data}(I_{SR})}(1 - D(I_{SR})) \tag{9}$$

G represents the ART model; and D represents the VGG16 discriminator model, which has the same backbone as [42] for binary classification. $D(I_{HR})$ and $D(I_{SR})$ are the outputs of the discriminator. $E_{I_{\mathrm{HR}}}$ and $E_{I_{SR}}$ are the expected value over the real data distributions $P_{data}(I_{HR})$ and $P_{data}(I_{SR})$. Finally, the combined loss function for the ART model can be written as:

$$\begin{aligned} L_C(I_{HR}, I_{SR}) = {} & \eta \times L_G(I_{HR}, I_{SR}) + \lambda \times L_1(I_{HR}, I_{SR}) + \gamma \\ & \times L_p(I_{HR}, I_{SR}) \end{aligned} \tag{10}$$

where η, λ, and γ are hyper-parameter coefficients introduced to find the trade-off between the separated losses.

3 | EXPERIMENT SETUP

3.1 | Swept-Source (SS)-OCT and data acquisition

Human skin data were collected in-vivo using a lab-built SS-OCT 1310 nm system to test the performance of the proposed algorithm. The details of the SS-OCT (Figure 4) setup as well as its performance parameters were described elsewhere [45]. The study was approved by the School of Science and Engineering Research Ethics Committee of the University of Dundee (No. UOD_SS-REC_PGR_2022_003), which also conformed to the tenets of the Declaration of Helsinki. In the skin OCT data, arm and hand back (representative "thin" skin), palm (representative "thick" skin), and face and neck (representative difficult-to-reach regions) were taken from 10 subjects (four females and six males) aged between 20 and 35 years old, none of whom had any skin conditions. A free hand-held scan probe was used to collect the data from normal skin. To prevent the influence of the motion artifacts and increase the sample size of the OCTA images, the repetition of each position of participants is three.

Table 1 demonstrates the scan protocol of the SSOCT device during the OCT data acquisition. In Table 1, the data stage was related to the stages in Figure 3 for neural network training and testing, respectively. As for the imaging protocol in Table 1, both train and test data covered a volume of 6 × 6 × 1.5 mm^3 ($x \times y \times z$). In train data, one OCT data scanned from the participants consists of 6 × 400 × 400 × 300 (n, x, y, z) voxels. In test data, one OCT data consists of 6 × 100 × 100 × 300

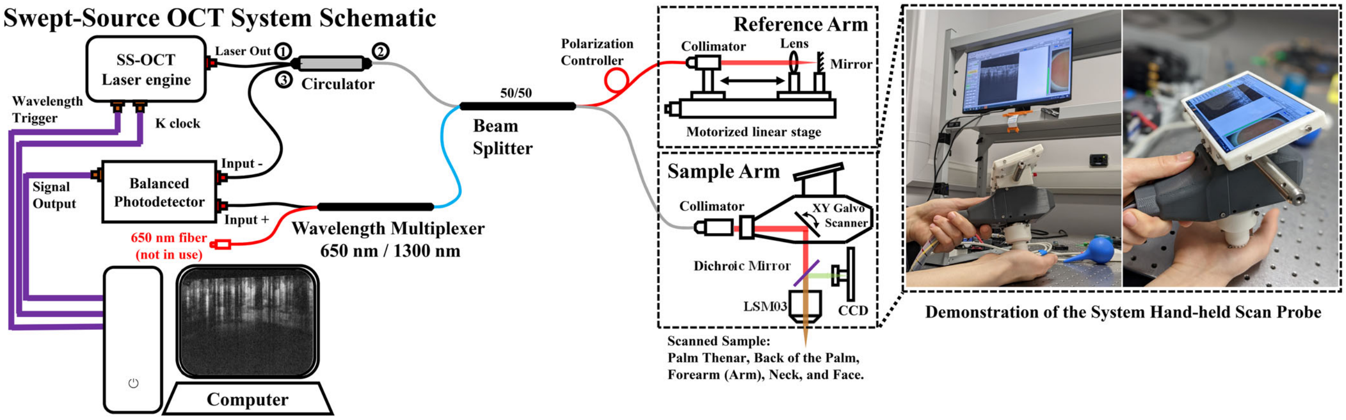


FIGURE 4 The system schematic of the lab-built swept-source optical coherence tomography (SSOCT) system. The laser engine used in this system is Thorlab SL1310, which has a 200 kHz swept rate, and a wavelength of 1310 nm with a 100 nm bandwidth. The axial resolution given by this system is ~8 μm in human skin tissue. The reference arm is consisting of a motorized linear stage to adjust the reference length to ensure the coherence signal is moderate for OCT scanning. The sample arm is consisting of a charge-coupled device (CCD) and a small screen to assist in the localization of the interested scanning area, a pair of 2D galvo-scanners, and an LSM03 lens. Moreover, the sample arm is made as a flexible hand-held scan probe (shown in the right figure), to image difficult-to-reach regions (e.g., face).

TABLE 1 Parameter setup of SSOCT system for OCT data acquisition.

Data (stage)	FOV (enface)	Repeated scan	Transverse resolution	Scan time
Train	6 mm^2	6	15 μm/pixel	~7 s
Test	6 mm^2	6	60 μm/pixel	~0.3 s

(n, x, y, z). Where n is the repeated time of scans, x and y are the transverse axis, and z is the axial axis. Theoretically, depending on the laser used in this SSOCT system, the axial resolution is ~7.8 μm.

After the image processing (Figure 3 blue area) and abandoning the nonsignal enface images, a total of 5840 pairs of the enface OCTA images were selected as training datasets; among 4700 pairs of images were used as a training I_{LR} and I_{HR}, and the remaining 1140 pairs of images for validation datasets. In terms of the test data, to investigate the feasibility of the hand-held fast OCTA scan pipeline, OCTA images were captured and processed by the pipeline mentioned in Figure 3 red zone. After the maximum intensity projection, a volume of 200 low-transverse resolution enface OCTA images was selected as the test data to visually evaluate the performance of the pipeline.

3.2 | Implementation details

3.2.1 | Data argument

A series of data enhancement methods were used in the train datasets to improve the robustness of a trained model, including image flipping up/down and right/left, and image contrast changing with a random factor between 0.8 and 1.2. The data-augmented methods are not applied to the validation set and are only deployed during the training.

3.2.2 | Comparative studies

To better evaluate the performance of the proposed ART, a series of state-of-the-art SISR models were employed to conduct comparative studies in the field of OCTA image super-resolution, and those models are SRResNet [30], SRDenseNet [29], DenseReconstGAN [24], EDSR [28], DRRN [31], ESRT [46], and SwinIR [35].

3.2.3 | ART architecture

To investigate the performance of the proposed ART under different architectures, four versions of ART were proposed under different sizes: tiny, base, large and huge, and the details were in Table 2. Three parameters were configured: the number of heads to control the number of heads used in multihead attention to split the 1D sequences; the RCT layer number to control the numerical value of the RCT layer used in RCT block; the hidden

TABLE 2 Implementation details of different ART models.

ART type	Tiny	Base	Large	Huge
Heads	2	4	8	8
RCT layer (n)	2	4	6	12
Hidden size	32	64	128	192

size to control the filter size (f_s) used in convolutional projection and the hidden size for the Q, K, V sequences before being split into multihead attention.

3.2.4 | Implementation details

In ART model, the kernel size of all convolution layers was set as 3 × 3. The filter sizes of the convolution layers before upsample block were the same as the hidden size (in Table 2) used in the RCT block. The filter sizes of the convolution layers used in upsample block were set as 256 and the output layer was set as 64. Besides the down-sample and deconvolution layers used in RCT layers, which had strides 2, the convolution layers used in ART were set as stride 1. The residual scaling parameter β was set as 0.4, and the n to control the number of RCT layers was different in the four types of ART. In a feed-forward network (FFN), the numerical value of units used in the first fully connected layer was 4 × hidden size, and 1 × hidden size in the second layer. The architecture of FFN is the same as the vision transformer proposed [25]. An Adam optimizer with a learning rate of 0.0001 was used [47]. The training epochs were set as 800, and the early stopping was used when the value of the metrics did not improve in 50 epochs. The batch size was set as 16 for ART_{Tiny}, ART_{Base}, and ART_{Large}, and was set as 8 for ART_{Huge}. The coefficients in the loss function (10) were $\eta = 0.001$, $\lambda = 1$, and $\gamma = 0.01$. The training was deployed under an NVIDIA RTX 3090 graphics card with 24GB memory, and an Intel i9-10980xe with 64GB RAM. The version of CUDA was 11.3, with TensorFlow 2.6.0.

3.2.5 | Ablation study

The ablation study was based on the ART_{Base} model because of the moderated network parameters (i.e., close to SwinIR) and limited GPU memory. Table 3 depicted the four parameters and relevant setups in the ablation study. In Table 3, the setups with the underline were the control group, which has the same implementation details as the ART_{Base}.

TABLE 3 Experiment setup for the ablation study in ART_{Base} model.

Study	Parameters setup				
RCT layer	2	4	6	8	16
Heads	2	4	8	16	32
Hidden size	32	64	96	128	192
Data usage	20%	40%	60%	80%	100%

3.3 | Evaluation metrics

To quantitatively compare the performance of different neural networks, two performance metrics are used in the experiments: PSNR and structural similarity index measure (SSIM) [48].

$$PSNR = 20log_{10}\left(\frac{I_{max}}{\sqrt{MSE}}\right) \tag{11}$$

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

$$\mathrm{MSE} = \frac{1}{M \times N}\sum\nolimits_{m=0}^{M-1}\sum\nolimits_{n=0}^{N-1}\left(I_{GT}(m,n) - I_{SR}(m,n)\right)^2 \tag{12}$$

where I_{GT} and I_{SR} are the ground truth and the super-resolved OCTA images, respectively. The term I_{max} is set as 1 in this evaluation, which refers to the maximum possible value in the image data. 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.

$$SSIM = \frac{(2\mu_{GT}\mu_{SR} + k1)(2\sigma_{cov} + k2)}{(\mu^2{}_{GT} + \mu^2{}_{SR} + k1)(\sigma^2{}_{GT} + \sigma^2{}_{SR} + k2)} \tag{13}$$

Here, μ_{GT} and (σ_{GT}) and μ_{SR} and (σ_{SR}) 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 | RESULTS

4.1 | Comparison with state-of-the-art networks

Table 4 is the comparative results between the state-of-the-art networks and the proposed ART model. Figure 5

TABLE 4 Quantitative comparison (average ± standard deviation of PSNR/SSIM) with state-of-the-art methods for OCTA image super-resolution.

Method	Scale	Type	#Params (m)	Inference time	PSNR	SSIM
Bicubic	×4	Interpolation	/	0.0004 s/image	18.82 ± 1.01	0.3175 ± 0.0602
SRResNet [30]	×4	CNN	0.714	0.0038 s/image	21.29 ± 1.03	0.3550 ± 0.0636
SRDenseNet [29]	×4	CNN	6.976	0.0091 s/image	21.79 ± 1.04	0.3577 ± 0.0722
EDSR [28]	×4	CNN	8.423	0.0072 s/image	21.56 ± 1.09	0.3681 ± 0.0770
DRRN [31]	×4	CNN	2.154	0.0485 s/image	21.77 ± 1.00	0.3452 ± 0.0574
DenseReconstGAN [24]	×4	CNN	5.984	0.0421 s/image	21.44 ± 1.01	0.3460 ± 0.0623
ESRT [46]	×4	Transformer	14.429	0.0702 s/image	21.86 ± 1.03	0.3517 ± 0.0575
SwinIR [35]	×4	Transformer	1.473	0.0208 s/image	22.08 ± 1.04	0.3826 ± 0.0700
ART_{Tiny}	×4	Transformer	0.639	0.0056 s/image	21.74 ± 1.02	0.3693 ± 0.0595
ART_{Base}	×4	Transformer	1.591	0.0092 s/image	22.15 ± 1.04	0.3930 ± 0.0727
$\text{ART}_{\text{Large}}$	×4	Transformer	6.430	0.0188 s/image	22.42 ± 1.03	0.4189 ± 0.0929
ART_{Huge}	×4	Transformer	25.514	0.0379 s/image	22.80 ± 1.05	0.4633 ± 0.1086

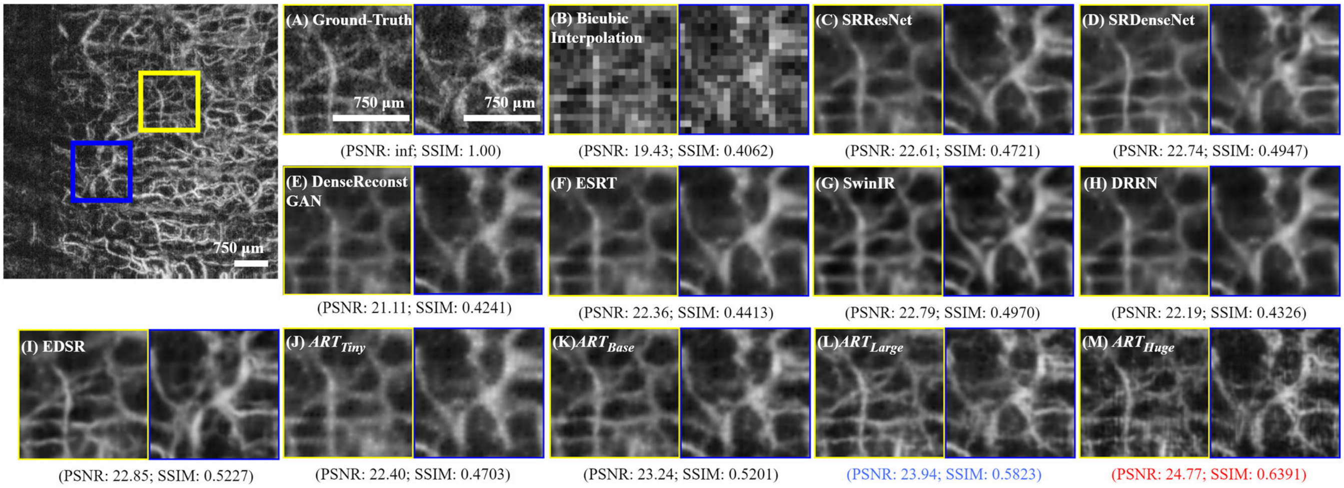

750 µm
(A) Ground-Truth
750 µm
750 µm
(PSNR: inf; SSIM: 1.00)
(B) Bicubic Interpolation
(PSNR: 19.43; SSIM: 0.4062)
(C) SRResNet
(PSNR: 22.61; SSIM: 0.4721)
(D) SRDenseNet
(PSNR: 22.74; SSIM: 0.4947)
(E) DenseReconst GAN
(PSNR: 21.11; SSIM: 0.4241)
(F) ESRT
(PSNR: 22.36; SSIM: 0.4413)
(G) SwinIR
(PSNR: 22.79; SSIM: 0.4970)
(H) DRRN
(PSNR: 22.19; SSIM: 0.4326)
(I) EDSR
(PSNR: 22.85; SSIM: 0.5227)
(J) ART_Tiny
(PSNR: 22.40; SSIM: 0.4703)
(K)ART_Base
(PSNR: 23.24; SSIM: 0.5201)
(L)ART_Large
(PSNR: 23.94; SSIM: 0.5823)
(M) ART_Huge
(PSNR: 24.77; SSIM: 0.6391)

FIGURE 5 Visual comparison of OCTA image super-resolution based on the test datasets. (A) ground-truth image; (B) Bicubic interpolation method; (C) SRResNet; (D) SRDenseNet; (E) DenseReconstGAN; (F) ESRT; (G) SwinIR; (H) DRRN; (I) EDSR; (J) ART_{Tiny}; (K) ART_{Base}; (L) ART_{Large}; (M) ART_{Huge}. The scale bar is 750 μm. The red result is the best, and the blue result is the second-best result. The scale bar is shown as a white label in the figure.

is the visual results of the networks (full size images are available in Figure A1). Based on the experiment observation, the bicubic interpolation method is the fastest (0.0004 s/image), but the results are worst (PSNR: 18.82). In the CNN-type models, the inference time of the SRResNet, SRDenseNet, and EDSR is lower than 0.01 s/image, but the mean PSNR is lower than 21.8, and the mean SSIM is lower than 0.37. DRRN and DenseReconstGAN that used a densely connected architecture have a high inference time (i.e., 0.0485 and 0.0421 s/image, respectively), while the performances of the PSNR are lower than 22, and the SSIM are lower than 0.35. In the transformer-type models, the proposed ART_{Tiny} had fewer parameters (0.639 M) than SRResNet, SRDenseNet, and EDSR and the best SSIM result (SSIM: 0.3693). Compared with the SwinIR, the proposed ART_{Base} can provide better quantitative results (PSNR: 22.15 > 22.08; SSIM: 0.3930 > 0.3826) of the OCTA image super-resolution, while the parameters are similar and had a faster inference time (0.0092 s/image). In terms of the ART_{Large} and ART_{Huge}, those two models provide the best competitive results in this study (PSNR: 22.42/SSIM: 0.4189 and PSNR: 22.80/SSIM:0.4633); however, the parameters are larger than 6 million and required more inference time than 0.01 s for an image. Based on the visual observation in Figure 5, the bicubic interpolation result is burly and the vascular texture details are hard to classify (PSNR: 19.43; SSIM: 0.4062). The results from CNN and transformer models can provide a high-quality texture detail for super-resolved large vessels; however, only the results in (I), (K), (L), and (M) had a better quantitative result (PSNR >22.8; SSIM >0.52) and can provide good vessel connectivity and high-quality texture details for small vessels. The ART_{Tiny} had the fewest network parameters, but the super-resolved image in Figure 5 (J) is with obvious artifacts (PSNR: 22.4; SSIM: 0.4703). Compared with the ground-truth image (A), based on the visual observation, the results from (C) to (M) have a lower noise level and we hypothesize that this is because of the utilization of L_1 and L_2 losses for network training.

4.2 | Ablation study

Figure 8 is the quantitative results of the ablation study based on the ART_{Base}. In Table 4, compared with the *SwinIR* result, the ART_{Base} has a faster inference time (0.0092 s < 0.0208 s) and better competitive results (PSNR: 22.15 > 22.08; SSIM: 0.393 > 0.3826). Therefore, compared with the ART_{Large} and ART_{Huge} models, the ART_{Base} with moderated network parameters (1.591 M) was used for the ablation study.

Figure 8 shows that the more hidden size used in the ART_{Base} model, the higher performance of the ART model in OCTA image super-resolution (PSNR from 21.97 to 22.69; SSIM from 0.3706 to 0.4481). Furthermore, the performance of the ART_{Base} is proportional to the number of RCT layers (PSNR from 21.95 to 22.41; SSIM from 0.3746 to 0.4185). However, compared with the control group (the head number is 4), the smaller or higher the numerical value of the head will decrease the performance of ART_{Base} (best scores when the number of the head is 4, PSNR: 22.15; SSIM: 0.3931). In terms of the influence of the data size, the performance of ART_{Base} is decreased seriously (SSIM from 0.3931 to 0.3504; PSNR from 22.15 to 21.72) when 20% of train datasets (1086 pairs of images) are used for network training. Nevertheless, the performance of ART_{Base} is degraded slightly (PSNR from 21.95 to 22.15) when the data size is from 40% to 100% of datasets.

4.3 | Proposed fast OCTA scan pipeline

The system setup was mentioned in the last paragraph, and the scanning protocol for the fast OCTA scan was in Figure 3 (red zone). With the proposed pipeline, the data acquisition time was reduced from 7 s to 0.3 s (reduced

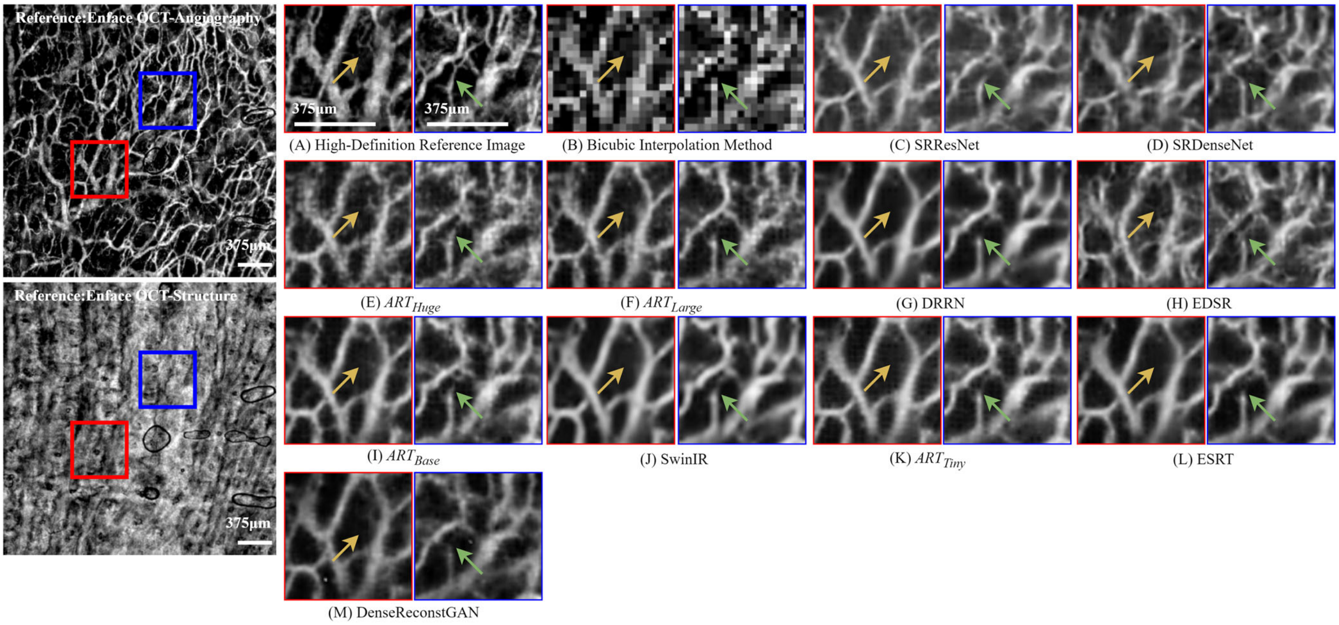
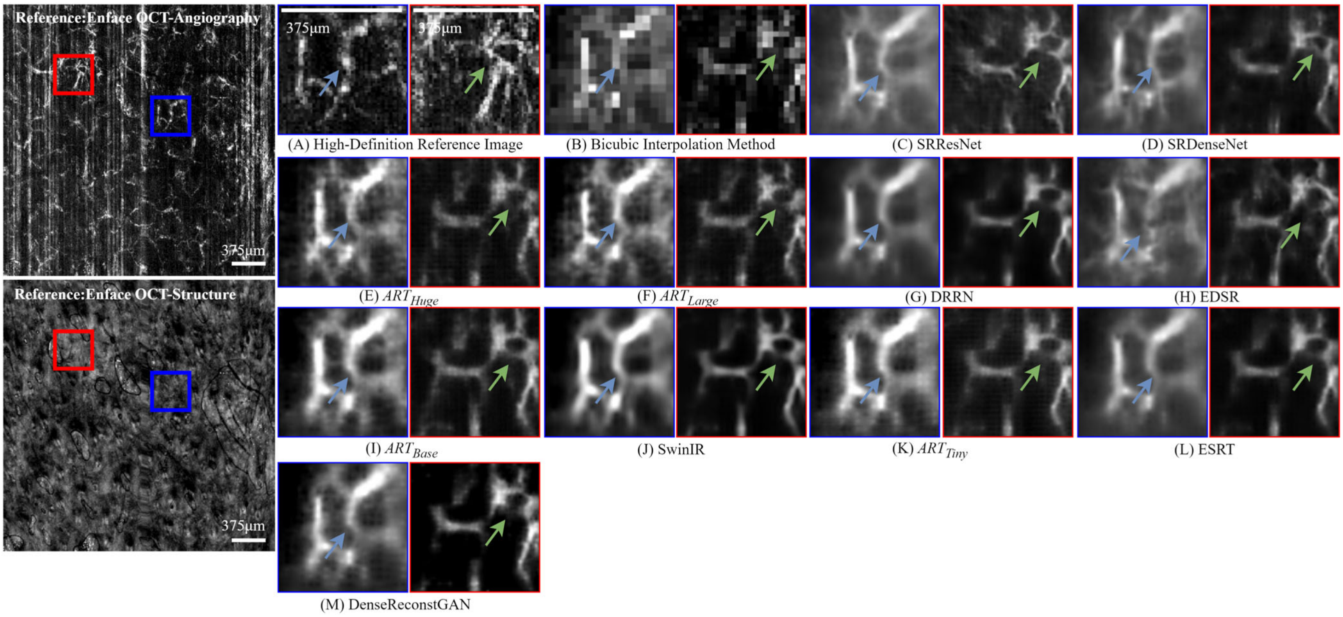


FIGURE 6 Visual comparison of OCTA image super-resolution based on the real low-spatial sampling rate OCTA imaging (palm area); the scanning protocol is the same as the test stage in Figure 3. (A) is the reference image acquired under 15 μm/pixel spatial sampling rate, in this stage, (A) is only used as a reference image for visual comparison. (B): Bicubic interpolation method; (C) SRResNet; (D) SRDenseNet; (E) ART_{Huge}; (F) ART_{Large}; (G) DRRN; (H) EDSR; (I) ART_{Base}; (J) SwinIR; (K) ART_{Tiny}; (L) ESRT; (M) DenseReconsGAN. The scanning area was palm with a 6 mm^2 field of view. The scale bar is shown as a white label in the figure.


Reference:Enface OCT-Angiography
375μm
Reference:Enface OCT-Structure
375μm
375μm
375μm

(A) High-Definition Reference Image
(B) Bicubic Interpolation Method
(C) SRResNet
(D) SRDenseNet
(E) ART_{Huge}
(F) ART_{Large}
(G) DRRN
(H) EDSR
(I) ART_{Base}
(J) SwinIR
(K) ART_{Tiny}
(L) ESRT
(M) DenseReconstGAN

FIGURE 7 Visual comparison of OCTA image super-resolution based on the real low-spatial sampling rate OCTA imaging (face area); the scanning protocol is the same as the test stage in Figure 3. (A) is the reference image acquired under 15 μm/pixel spatial sampling rate, in this stage, (A) is only used as a reference image for visual comparison. (B): Bicubic interpolation method; (C) SRResNet; (D) SRDenseNet; (E) ART_{Huge}; (F) ART_{Large}; (G) DRRN; (H) EDSR; (I) ART_{Base}; (J) SwinIR; (K) ART_{Tiny}; (L) ESRT. (M) DenseReconsGAN. The scanning area was palm with a 6 mm^2 field of view. The scale bar is shown as a white label in the figure.

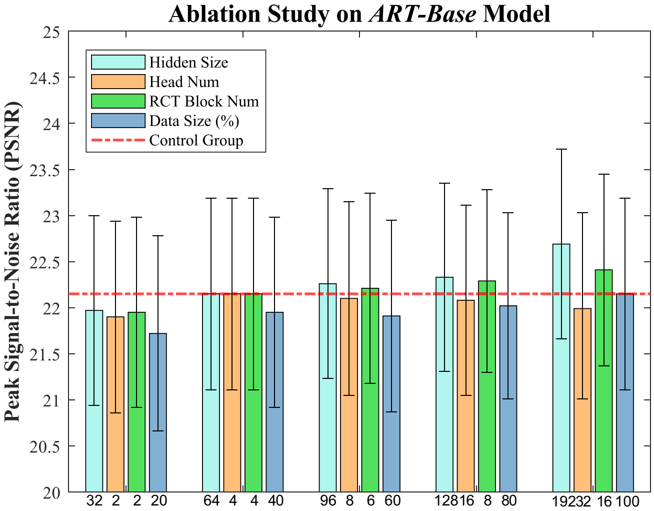


FIGURE 8 Ablation Study based on the proposed ART_{Base} Model. The quantitative evaluation metrics are demonstrated as PSNR in mean ± standard deviation. Details of the setup of the ablation study are in Table 3..

by 95%), while the FOV and the number of repeated scans are maintained at 6 mm^2 and 6, and efficiency prevents the motion artifacts from participants and hand-held (Figure 4) scan probe. Moreover, with the reduction of data size acquired by the fast OCTA scan pipeline, the data processing time was reduced from 2 min to 15 s (Platform: Intel i7-8700 with 32G RAM). The overall time required for the proposed pipeline is less than 20 s, which is 6.5 times faster than the normal pipeline (~130 s).

Figures 6 and 7 depict the real-world super-resolved OCTA images (B–M) with a reference high-resolution (A) OCTA image (Figure A2 is the full size of Figure 6, and Figure A3 is the full size of Figure 7). In this stage, the quantitative comparison is unavailable because the low transverse resolution OCTA images were not degraded from the counterpart high transverse resolution images, but from the hand-held scan probe with a low transverse resolution scanning protocol, as mentioned in Table-1 Test.

To simulate the OCTA scan in the clinical environment, the data acquisition area was defined as an easily obtained area (less motion palm area in Figure 6) and a hard obtained area (more motion face area in Figure 7) with the hand-held scan probe. The high-definition (15 μm/pixel) and low-spatial sampling rate (60 μm/pixel) OCT enface images were scanned under the close area with slight movement. The red and blue areas of the reference image (A) in Figures 6 and 7 were selected manually, which can be used as high-definition references but not ground truth, and the OCT structure image was also available.

In Figure 6, the reconstruction of the microvessel texture details was hard because those vessels were unavailable to be resolved under the low-spatial sampling rate scanning protocol. Based on the experiment observation, compared with the reference (A), the super-resolved image (E) from ART_{Huge} contains the micro-vessel (yellow arrow) while the super-resolved results in (B)–(D) and (F)–(M) do have not any vessel signal in yellow arrow. In terms of the connectivity of the vessel (represented in green arrow), the results in (E), (F), (I), and (J) show good connectivity between the micro-vessel; however, the results in (B)–(D), (G), and (L)–(M) have a bad vessel connection in green arrow area. Although (H) represents a good super-resolved result in the visual aspect, compared with the reference image (A), the morphology and connection between the vessel are incorrect, and that might because of the overfitting of the trained EDSR network. Compared with SwinIR (J), the results from ART_{Base} (I), ART_{Large} (F), and ART_{Huge} (E) can represent a sharper image, which is close to the reference image (A).

In Figure 7, the quality of reference image (A) is relatively low, and the motion artifacts (i.e., the vertical white line) are obvious. While those artifacts are disappeared in the super-resolved images because the scanning time is 0.3 s in the hyper-fast OCTA scanning pipeline. The blue and green arrows in Figure 7 are marks to find the correlated position of the vessel. Compared with reference (A), the results from ART_{Base} (I), ART_{Tiny} (K), and *SwinIR* (J) are less noise and higher contrast; furthermore, the result (I) has better vessel connectivity (blue and green arrow) than (J). In terms of the blue zoom-in area, the results from the CNN-based models ((C), (D), (G), (H), and (M)) have less contrast than the results from the transformer-based models ((E), (F), (I), (J), (K), and (L)). Although the result from EDSR (H) contains rich blood flow signals, those signals are mismatched with reference (A) in the aspect of morphology. In the visual aspects, the results from ART_{Large} (F) and ART_{Huge} (F) are performances worse than ART_{Base} (I) and ART_{Tiny} (K) in face OCTA super-resolution, we suppose that is because of the network overfitting. With the benefit of the faster OCTA scanning (0.3 s), the super-resolved results ((E), (F), (H), (I)-(M)) from neural networks represent a better vessel connection and contrast than the high-definition OCTA scan, which need the participants to keep still under 7 s.

5 | DISCUSSION

In this work, we proposed a novel ART model to achieve a fast OCTA imaging pipeline, which includes a free hand-held scan probe to scan the interesting area of skin, a trained ART model for the image super-resolution, and a pre-processing workflow to generate a low-spatial sampling rate enface OCTA images. With the proposed ART-based OCTA scan pipeline, the data processing time was reduced from ~300 s to ~20s, and the data acquisition time was reduced from 7 s to 0.3 s. Furthermore, the proposed pipeline is efficient to reduce the motion artifacts during the OCTA scan. This study has opened up a number of directions for future possible studies. In the current study, OCTA clinical workflow focused on ophthalmology tasks which are the main primary structures assessment with fixed position. We observed that few operators looked at soft tissue such as the skin. It would be interesting to acquire free-hand OCTA scans covering a wider range of scans and use the methodology reported in this paper to assist in understanding the standardization of the first free-hand skin OCTA scanner. The second direction of study might utilize the current knowledge obtained from this study to provide the justification for, and subsequent evaluation of assistive tools for GAN network-based image reconstruction for other domains such as photoacoustic, and 3D ultrasound.

Regarding the proposed ART model, with the convolutional transformer architecture, the neighboring information in low-spatial sampling OCTA images (i.e., input) is utilized well for feature extraction. Furthermore, the parameters are reduced by the downsample and upsample layers in the RCT layer (Figure 1(A)). The ablation study on data size (Figure 8 blue bar) shows that the ART_{Base} model has good competitive results under the 1880 pairs of images (40% of the training dataset). In Table 4 and Figure 5, the ART_{Huge} can achieve the best competitive results but has large network parameters, and the visual result is worse than ART_{Large} in Figure 7. We suppose that this is because of limited training data size (<5 k image) and a large network (>25 M parameters). In ART_{Base} and ART_{Large}, both two networks can achieve better competitive results than *SwinIR* in inference time and results (Table 4). Although the ART_{Tiny} can achieve a fast inference time (0.0056 s/image), the results (Figure 5 (I)) include artifacts and perform worse than *SwinIR* (F) in quantitative results (PSNR: $22.4 < 22.79$; SSIM: $0.4703 < 0.4970$).

Based on the visual comparison in Figure 5, the results from neural networks (Figure 5 C-M) are well based on the validation dataset from the downsampling operation. Still, the results (Figures 6 and 7C–M) have a gap in the microvessel reconstruction with the reference (Figure 6A) when the input image is acquired under real-world low-spatial sampling rate scanning protocol. We hypothesize that might be because the microvessel cannot be resolved under the low-spatial sampling rate (60 μm/pixel), and the trained neural networks are unavailable to super-resolve the micro-vessel from the nonexistent blood flow signal.

Our study has limitations. First, the input shape was fixed for the transformer-type neural network, which means the change of the OCTA image acquisition protocol will require the re-train of the model. Although the reshaping of the input image can solve this limitation, the vessel signal might be lost during the reshaping operation. Secondly, the GPU in this study has only 24G memory, which limits the batch size of the model training. Hence, we will further develop a lightweight transformer-based model with scalable input, which can achieve a shorter time of processing and higher super-resolved image quality. Thirdly, high-quality OCTA data acquisition is hard with a hand-held scan probe, and the quality of OCTA images will influence the performance of trained neural networks; we will investigate a higher efficient method to acquire high-quality OCTA images.

6 | CONCLUSION

Our proposed method has achieved a good competitive result in enfaces OCTA images super-resolution. In the future, we will inform the hyper-fast OCTA scan pipeline into the field of the oral OCT/OCTA scan, face and head OCT/OCTA scan, and the wide-field retinal OCT/OCTA scan. Moreover, with the trained ART models, the wide field (15 × 15 mm^2 FOV) OCTA scan for the skin application is available while the scan time is not changed (i.e., use low-spatial sampling rate for wide FOV scan) and the image quality is not degraded seriously.

ACKNOWLEDGMENTS

I would like to express my gratitude to Yilong Zhang, Yubo Ji, and Thomas Zhang for their supports.

FUNDING INFORMATION

This research received no external funding.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The data presented in this study are available on request from the corresponding author. The data are not publicly available due to ethical restrictions.

ORCID

Jinpeng Liao https://orcid.org/0000-0001-6287-8079
Chunhui Li https://orcid.org/0000-0003-2186-5137

REFERENCES

[1] A. J. Deegan, R. K. Wang, *Phys. Med. Biol.* **2019**, *64*, 07TR01.

[2] B. Zabihian, Z. Chen, E. Rank, C. Sinz, M. Bonesi, H. Sattmann, J. R. Ensher, M. P. Minneman, E. E. Hoover, J. Weingast, *J. Biomed. Opt.* **2016**, *21*, 096011.

[3] U. Baran, Y. Li, W. J. Choi, G. Kalkan, R. K. Wang, *Lasers Surg. Med.* **2015**, *47*, 231.

[4] A. J. Deegan, F. Talebi-Liasi, S. Song, Y. Li, J. Xu, S. Men, M. M. Shinohara, M. E. Flowers, S. J. Lee, R. K. Wang, *Lasers Surg. Med.* **2018**, *50*, 183.

[5] L. F. di Ruffano, J. Dinnes, J. J. Deeks, N. Chuchu, S. E. Bayliss, C. Davenport, Y. Takwoingi, K. Godfrey, C. O'Sullivan, R. N. Matin, *Cochrane Database System. Rev.* **2018**, *12*. https://www.cochranelibrary.com/cdsr/doi/10.1002/14651858.CD013189/information

[6] A. Marneffe, M. Suppa, M. Miyamoto, V. del Marmol, M. Boone, *Exp. Dermatol.* **2016**, *25*, 684.

[7] A. J. Deegan, W. Wang, S. Men, Y. Li, S. Song, J. Xu, R. K. Wang, *Quant. Imaging Med. Surg.* **2018**, *8*, 135.

[8] J. Fingler, D. Schwartz, C. Yang, S. E. Fraser, *Opt. Express* **2007**, *15*, 12636.

[9] E. Jonathan, J. Enfield, M. J. Leahy, *J. Biophotonics* **2011**, *4*, 583.

[10] J. Xu, S. Song, Y. Li, R. K. Wang, *Phys. Med. Biol.* **2017**, *63*, 015023.

[11] J. M. Schmitt, S. H. Xiang, K. M. Yung, *J. Biomed. Opt.* **1999**, *4*, 95.

[12] R. K. Wang, A. Zhang, W. J. Choi, Q. Zhang, C. L. Chen, A. Miller, G. Gregori, P. J. Rosenfeld, *Opt. Lett.* **2016**, *41*, 2330.

[13] B. Baumann, C. W. Merkle, R. A. Leitgeb, M. Augustin, A. Wartak, M. Pircher, C. K. Hitzenberger, *Biomed. Opt. Express* **2019**, *10*, 5755.

[14] M. Asif, M. U. Akram, T. Hassan, A. Shaukat, R. Waqar, *Eighth International Conference on Graphic and Image Processing (ICGIP 2016)*, SPIE, Bellingham, Washington, USA, **2017**, p. 204.

[15] X. Zhang, Z. Li, N. Nan, X. Wang, *Opt. Express* **2022**, *30*, 5788.

[16] W.-C. Siu, K.-W. Hung, *Proceedings of The 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference*, IEEE, Hollywood, California, **2012**, p. 1.

[17] D. Han, *Conference of the 2nd International Conference on Computer Science and Electronics Engineering (ICCSEE 2013)*, Atlantis Press, Hangzhou, China, **2013**, p. 1556.

[18] X. Liu, Z. Huang, Z. Wang, C. Wen, Z. Jiang, Z. Yu, J. Liu, G. Liu, X. Huang, A. Maier, Q. Ren, Y. Lu, *J. Biophotonics* **2019**, *12*, e201900008.

[19] Z. Jiang, Z. Huang, B. Qiu, X. Meng, Y. You, X. Liu, M. Geng, G. Liu, C. Zhou, K. Yang, A. Maier, Q. Ren, Y. Lu, *IEEE Trans. Med. Imaging* **2020**, *40*, 688.

[20] Z. Jiang, Z. Huang, B. Qiu, X. Meng, Y. You, X. Liu, G. Liu, C. Zhou, K. Yang, A. Maier, Q. Ren, Y. Lu, *Biomed. Opt. Express* **2020**, *11*, 1580.

[21] B. Qiu, Y. You, Z. Huang, X. Meng, Z. Jiang, C. Zhou, G. Liu, K. Yang, Q. Ren, Y. Lu, *J. Biophotonics* **2021**, *14*, e202000282.

[22] S. Cao, X. Yao, N. Koirala, B. Brott, S. Litovsky, Y. Ling, Y. Gan, *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, IEEE, Montreal, Chair, **2020**, p. 1879.

[23] Y. Huang, Z. Lu, Z. Shao, M. Ran, J. Zhou, L. Fang, Y. Zhang, *Opt. Express* **2019**, *27*, 12289.

[24] G. Kim, J. Kim, W. J. Choi, C. Kim, S. Lee, *Sci. Rep.* **2022**, *12*, 1289.

[25] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *arXiv Preprint arXiv:2010.11929* **2020**. https://arxiv.org/abs/2010.11929

[26] I. Goodfellow, J. Pouget–Abadie, M. Mirza, B. Xu, D. Warde–Farley, S. Ozair, A. Courville, Y. Bengio, *Adv. Neural Inf. Process. Syst.* **2014**, *27*, 2672.

[27] C. Dong, C. C. Loy, K. He, X. Tang, *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *38*, 295.

[28] B. Lim, S. Son, H. Kim, S. Nah, K. M. Lee, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* **2017**, 136.

[29] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, C. Change Loy, *Proc. Eur. Conf. Comput. Vision (ECCV)* **2018**, *11133*, 63.

[30] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, *Proc. IEEE Conf. Comput. Vision Pattern Recognit.* **2017**, 4681.

[31] Y. Tai, J. Yang, X. Liu, *Proc. IEEE Conf. Comput. Vision Pattern Recognit.* **2017**, 3147.

[32] W.-S. Lai, J.-B. Huang, N. Ahuja, M.-H. Yang, *Proc. IEEE Conf. Comput. Vision Pattern Recognit.* **2017**, 624.

[33] K. He, X. Zhang, S. Ren, J. Sun, *Proc. IEEE Conf. Comput. Vision Pattern Recognit.* **2016**, 770.

[34] H. Chen, Y. Wang, T. Guo, C. Xu, Y. Deng, Z. Liu, S. Ma, C. Xu, C. Xu, W. Gao, *Proc. IEEE/CVF Int. Conf. Comput. Vision* **2021**, 12299.

[35] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, R. Timofte, *Proc. IEEE/CVF Int. Conf. Comput. Vision* **2021**, 1833.

[36] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, *Proc. IEEE/CVF Int. Conf. Comput. Vision* **2021**, 10012.

[37] H. Wu, Xiao, N. Codella, M. Liu, X. Dai, L. Yuan, L. Zhang, *Proc. IEEE/CVF Int. Conf. Comput. Vision* **2021**, 22.

[38] Z. Wang, J. Chen, S. C. H. Hoi, *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *43*, 3365.

[39] A. Odena, V. Dumoulin, C. Olah, *Distill* **2016**, *1*, e3.

[40] S. Yousefi, Z. Zhi, R. K. Wang, *IEEE Trans. Biomed. Eng.* **2011**, *58*, 2316.

[41] Q. Zhang, J. Wang, R. K. Wang, *Quant. Imaging Med. Surg.* **2016**, *6*, 557.

[42] K. Simonyan, A. Zisserman. very deep convolutional networks for large-scale image recognition. **2014** Available: http://arxiv.org/abs/1409.1556

[43] M. Wang, W. Zhu, K. Yu, Z. Chen, F. Shi, Y. Zhou, Y. Ma, Y. Peng, D. Bao, S. Feng, L. Ye, D. Xiang, X. Chen, *IEEE Trans. Med. Imaging* **2021**, *40*, 1168.

[44] Y. Ma, X. Chen, W. Zhu, X. Cheng, D. Xiang, F. Shi, *Biomed. Opt. Express* **2018**, *9*, 5129.

[45] Y. Ji, K. Zhou, S. H. Ibbotson, R. K. Wang, C. Li, Z. Huang, *J. Biophotonics* **2021**, *14*, e202100152.

[46] Z. Lu, J. Li, H. Liu, C. Huang, L. Zhang, T. Zeng, *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit.* **2022**, 457.

[47] D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization," *arXiv Preprint arXiv:1412.6980*, **2014**.

[48] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli, *IEEE Trans. Image Process.* **2004**, *13*, 600.

How to cite this article: J. Liao, S. Yang, T. Zhang, C. Li, Z. Huang, *J. Biophotonics* **2023**, e202300100. https://doi.org/10.1002/jbio.202300100

APPENDIX A

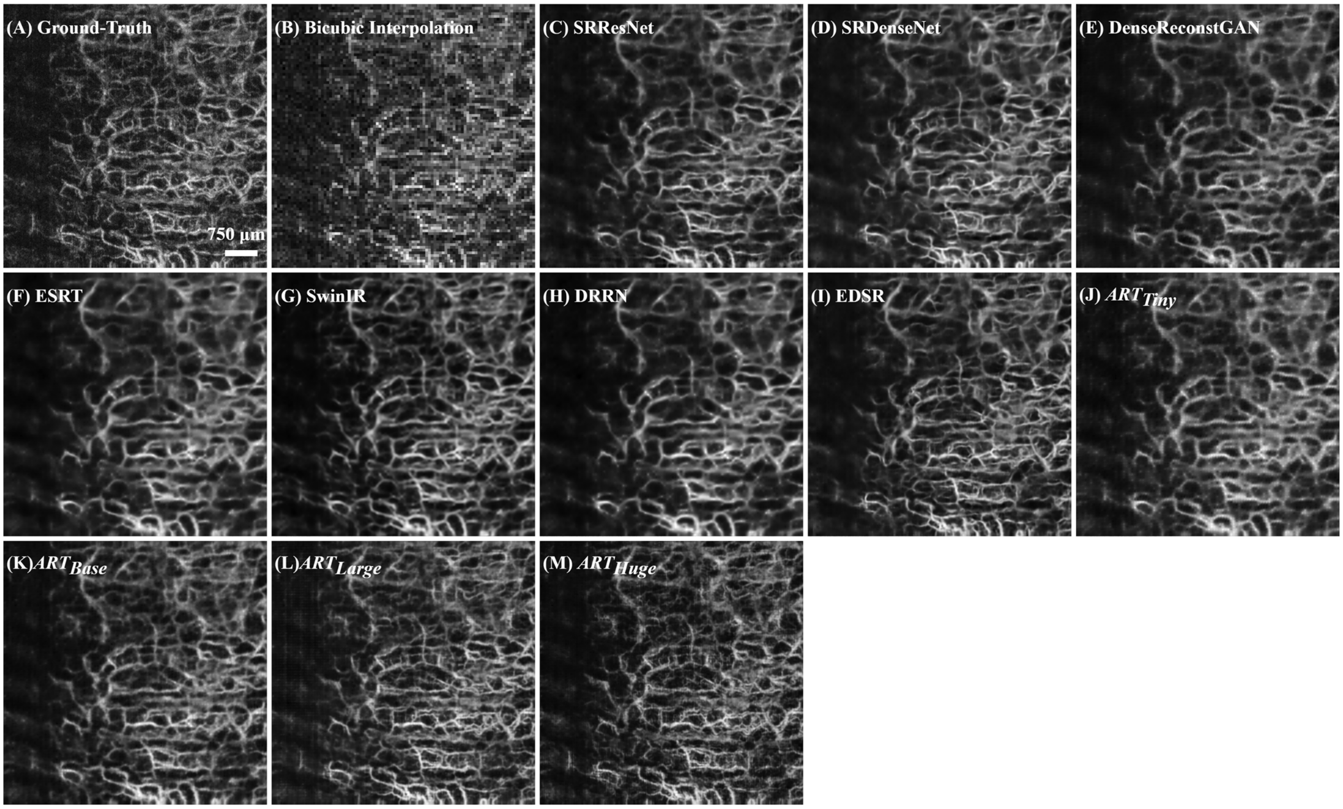


FIGURE A1 The full-size images of Figure 5 results for visual comparison of OCTA image super-resolution based on the test dataset. (A) ground-truth image; (B) Bicubic interpolation method; (C) SRResNet; (D) SRDenseNet; (E) DenseReconstGAN; (F) ESRT; (G) SwinIR; (H) DRRN; (I) EDSR; (J) ARTTiny; (K) ARTBase; (L) ARTLarge; (M) ARTHuge. The scale bar is shown as a white label in the figure.

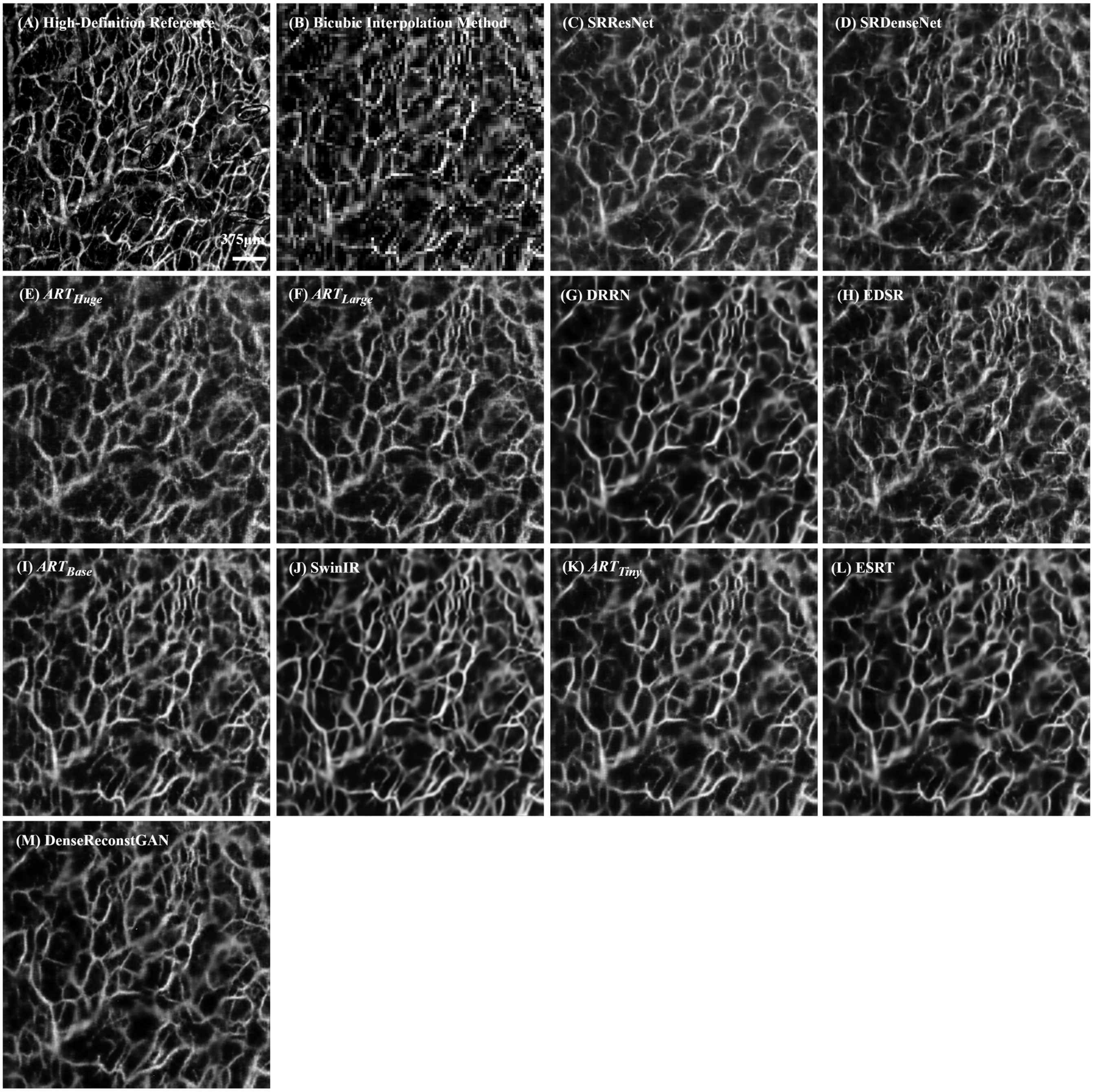


FIGURE A2 The full-size images of Figure 6 result for visual comparison of OCTA image super-resolution based on the real low-spatial sampling rate OCTA imaging (palm area). The scanning protocol is the same as the test stage in Figure 3. (A) is the reference image acquired under 15 μm/pixel spatial sampling rate, in this stage, (A) is only used as a reference image for visual comparison. (B): Bicubic interpolation method; (C) SRResNet; (D) SRDenseNet; (E) ARTHuge; (F) ARTLarge; (G) DRRN; (H) EDSR; (I) ARTBase; (J) SwinIR; (K) ARTTiny; (L) ESRT; (M) DenseReconsGAN. The field of view is set as 6 mm × 6 mm. The scale bar is shown as a white label in the figure.

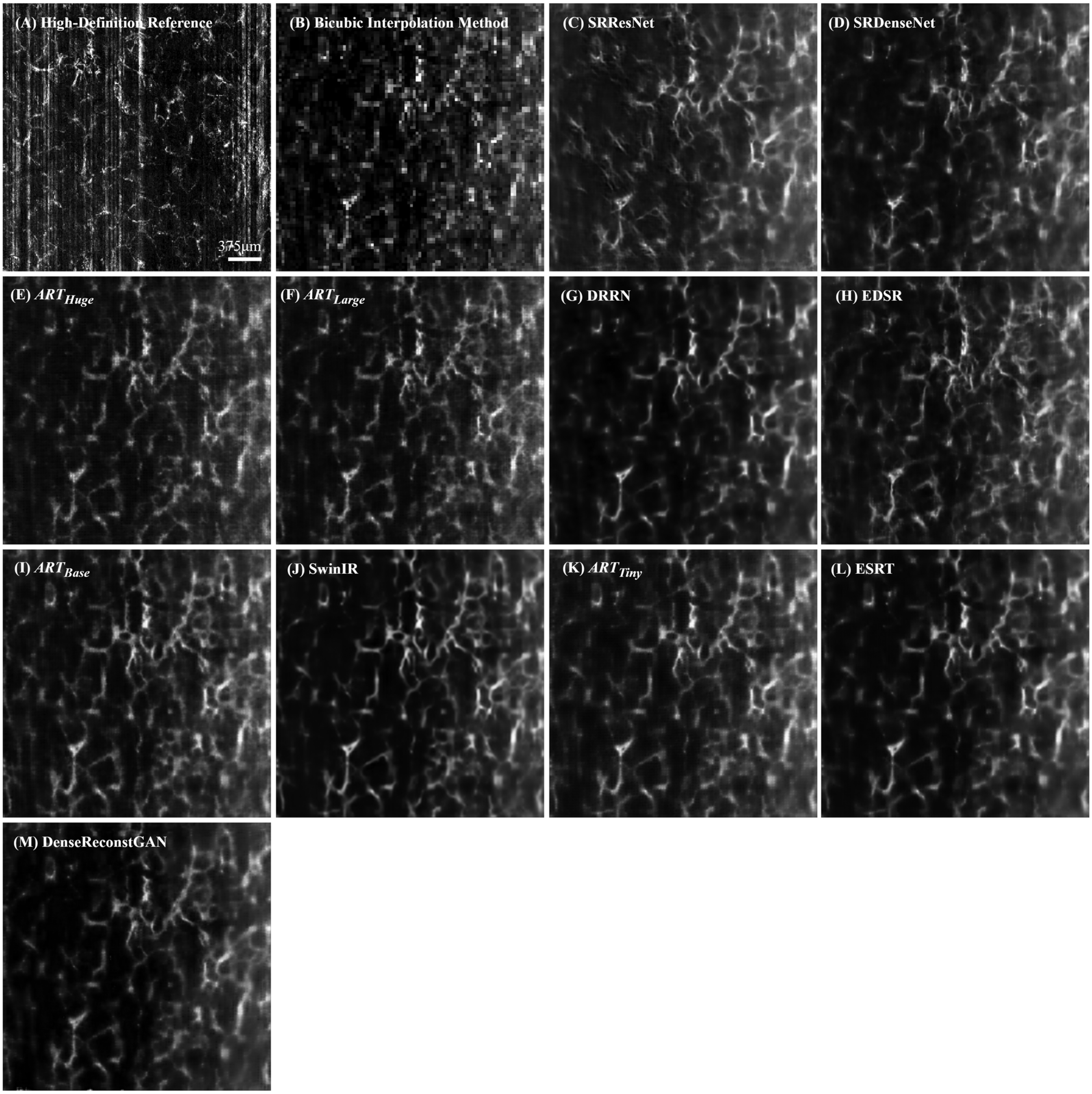


FIGURE A3 The full-size images of Figure 7 result for visual comparison of OCTA image super-resolution based on the real low-spatial sampling rate OCTA imaging (face area). The scanning protocol is the same as the test stage in Figure 3. (A) is the reference image acquired under 15 μm/pixel spatial sampling rate, in this stage, (A) is only used as a reference image for visual comparison. (B): Bicubic interpolation method; (C) SRResNet; (D) SRDenseNet; (E) ARTHuge; (F) ARTLarge; (G) DRRN; (H) EDSR; (I) ARTBase; (J) SwinIR; (K) ARTTiny; (L) ESRT. (M) DenseReconsGAN. The field of view is set as 6 mm × 6 mm. The scale bar is shown as a white label in the figure.