Biomedical Engineering Advances 6 (2023) 100099

Contents lists available at ScienceDirect

Biomedical Engineering Advances

journal homepage: www.journals.elsevier.com/biomedical-engineering-advances

A Hybrid Approach for retinal image super-resolution

Alnur Alimanov [*,a], Md Baharul Islam [a,b], Nirase Fathima Abubacker [c]

[a] Department of Computer Engineering, Bahcesehir University, Yildiz, Ciragan Cd., Besiktas, Istanbul, 34349, Turkey
[b] College of Data Science & Engineering, American University of Malta, Triq Dom Mintoff, Bormla, BML 1013, Malta
[c] School of Computing, Asia Pacific University of Technology & Innovation, Bukit Jalil, Kuala Lumpur, 57000, Malaysia

ARTICLE INFO

Keywords:
Retinal images
Single image super-Resolution
Adaptive patch embedding layer
Locality self-Attention
Vision transformer
Convolutional neural network

ABSTRACT

Experts require large high-resolution retinal images to detect tiny abnormalities, such as microaneurysms or issues of vascular branches. However, these images often suffer from low quality (e.g., resolution) due to poor imaging device configuration and misoperations. Many works utilized Convolutional Neural Network-based (CNN) methods for image super-resolution. The authors focused on making these models more complex by adding layers and various blocks. It leads to additional computational expenses and obstructs the application in real-life scenarios. Thus, this paper proposes a novel, lightweight, deep-learning super-resolution method for retinal images. It comprises a Vision Transformer (ViT) encoder and a convolutional neural network decoder. To our best knowledge, this is the first attempt to use a transformer-based network to solve the issue of accurate retinal image super-resolution. A progressively growing super-resolution training technique is applied to increase the resolution of images by factors of 2, 4, and 8. The prominent architecture remains constant thanks to the adaptive patch embedding layer, which does not lead to additional computational expense due to increased upscaling factors. This patch embedding layer includes 2-dimensional convolution with specific values of kernel size and strides that depend on the input shape. This strategy has removed the need to append additional super-resolution blocks to the model. The proposed method has been evaluated through quantitative and qualitative measures. The qualitative analysis also includes vessel segmentation of super-resolved and ground truth images. Experimental results indicate that the proposed method outperforms the current state-of-the-art methods.

1. Introduction

Medical images are essential in diagnosing various diseases, and retinal images are no exception. The retina is a layer of nerve tissue located at the eye's rear side and is responsive to incoming light. This light is converted into neural signals to the brain's visual cortex for visual perception. Digital retinal imaging allows the detection of various conditions directly related to eye diseases, e.g., diabetic retinopathy, hypertensive retinopathy, retinal tear and detachment, papilledema, optic atrophy, microaneurysm, etc. Microaneurysms are too tiny to be easily noticed; on average, the diameter of an aneurysm ranges between 43 and 266 microns with a mean value of only 104 microns [1]. These issues may lead to more serious and severe health conditions, e.g., diabetic retinopathy is a consequence of diabetes, hypertensive retinopathy caused by hypertension, papilledema may lead to a brain tumor, meningitis, and stroke. Segmented retinal images can also be used in diagnosing. By analyzing the vessel width, curvature, and structure ophthalmologists can detect various health issues.

Detecting these retinal issues at early stages will greatly assist medical treatment. Deep learning is broadly used in medicine; for example, Usui et al. [2] developed a Denoising network structure of CNN (DnCNN) to reduce the effect of noise in low-dose computed tomography images, Tang et al. [3] proposed an algorithm for retinal image segmentation. In addition, several methods based on deep learning techniques have recently been proposed to detect retinal diseases. For example, Grassmann et al. [4] trained an ensemble of 6 different deep-learning architectures to detect and classify 13 age-related retinal diseases, while Maji and Sekh [5] developed a CNN for automatic grading of retinal blood vessel maps. However, these methods heavily rely on high-quality fundus images for analysis. Hence, few traditional and learning-based methods were proposed for retinal image enhancement. These methods have produced low-resolution images due to the complexity of the task. Thus, retinal image super-resolution is an important task that requires an accurate technique.

Most deep learning-based methods utilized CNN architectures. For example, the recent Frequency-based Enhancement Network (FENet)

* Corresponding author.
E-mail addresses: mr.allnur_55@live.com (A. Alimanov), bislam@gmail.com (M.B. Islam), nfathima.abubacker@gmail.com (N.F. Abubacker).

https://doi.org/10.1016/j.bea.2023.100099
Received 14 April 2023; Received in revised form 23 May 2023; Accepted 21 June 2023
Available online 25 June 2023

[6] is composed of CNN that increases the resolution of images by factors of 2, 4, and 8; however, as the super-resolution factors increase, the model generates blurry results. Fig. 1 shows the output of our model along with ground truth and low-resolution input that was up-scaled using bicubic interpolation [7]; as we can observe, the resulted image is not only accurate compared to the Ground Truth (G.T), but also it is not affected by the blurriness and noise. As a result, lately, vision transformer models have taken over precedence [8]. Moreover, several CNN in image classification tasks regarding the accuracy and computational efficiency [9]. However, up-scaling natural and medical images requires different objectives. For instance, in natural image super-resolution, the objective is to generate high-resolution, realistic images. To do so, the models may generate realistic-looking results far from the ground truth. This cannot be applied to medical data, as it is crucial not to have features in the ground truth data. Additionally, the superiority of the ViT network comes at a price since ViT requires a significant amount of computational memory compared to CNN and cannot upscale the input. When ViT networks process the input, they do not change its shape; thereby, these models would require some up-scaling module, such as convolution or any interpolation method, that would reshape inputs after each ViT block. Thus, to increase the resolution of a single image standalone ViT model would need to repeat up-sampling steps several times; as the super-resolution factor increases, the number of these steps also increases, which obstructs the application of these methods in real life.

In this connection, we have focused on solving the problems of using standalone CNN or ViT, which include lower up-sampling accuracy of CNN and high computational complexity of ViT. Therefore, we intend to develop a computationally efficient way (in terms of memory and time) of utilizing ViT for extensive and accurate image generation. We use a combination of the ViT encoder and CNN decoder. Thus, we propose a novel lightweight fixed progressive encoder-decoder model that combines ViT and CNN for single image super-resolution. The model is trained, validated, and tested in the EyeQ dataset [10]. The quantitative metrics include PSNR, SSIM [11], and single image testing time (SITT). For further analysis, we used three other datasets such as DRIVE [12], STARE [13], and CHASE DB1 [14]. These datasets contain retinal images and corresponding manually segmented retinal vessels. The quantitative evaluation metrics involve the Jaccard similarity coefficient, Matthew's correlation coefficient (MCC), Cohen's kappa, F1-score, precision, recall, and accuracy and are compared with the state-of-the-art methods. The main contributions are summarized below.

- We propose a novel fixed progressively growing super-resolution model that combines a vision transformer encoder and a convolutional neural network decoder. To our knowledge, this is the first attempt to utilize a transformer-based technique in medical image super-resolution. This combination has led to higher computational, quantitative, and qualitative performance than standalone CNN and ViT models.
- We design a novel super-resolution training technique (Adaptive Patch Embedding Layer) that keeps the prominent architecture constant. This block has led to better quantitative and qualitative performance than progressively growing architecture.
- We adapt Locality Self-Attention instead of standard Multi-Head Self-Attention in our ViT encoder, which was designed for small- and mid-sized datasets, such as the EyeQ dataset. This has helped to achieve better performance.
- The presented hybrid model is lightweight and outperforms the state-of-the-art models in retinal image super-resolution on diverse datasets. In addition, an extensive ablation study has been performed to demonstrate the effectiveness of ViT and Adaptive Patch Embedding Layer, including statistical analysis.

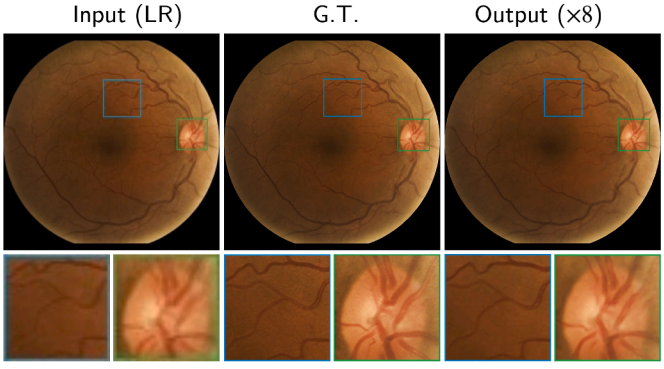

Fig. 1. An example of our super-resolution method's performance. Left to right: low-resolution input, Ground Truth (G.T.), and super-resolved output of our model. The resolution was increased by a factor of 8. The bottom images zoom in on the highlighted parts by 7 times.

The remainder of the paper is organized as follows. In Section 2, retinal image enhancement and super-resolution methods in the literature have been reviewed. A description of the proposed methodologies and implementation is provided in Section 3. In Section 4, we share information about utilized datasets, model evaluation techniques, and our experimental setup. Section 5 demonstrates the results obtained by our method and the comparison with other models. Finally, the conclusion and plans are presented in Section 6.

2. Related works

2.1. Medical image restoration methods

Recently, many deep learning methods [15–24] have been developed for medical image enhancement. For example, Deng et al. [15] developed a transformer-based Generative Adversarial Network and a paired images dataset. In some other works [16,21,24], the authors utilized cycle-consistent GAN with Convolutional Block Attention Module (CBAM) [25] for unpaired image-to-image translation, low- to high-quality. The authors in [17,18] focused on manually degrading high-quality images to get paired data and used conditional GAN. Subramani and Veluchamy [19] developed an adaptive fuzzy gray level difference histogram equalization algorithm for medical image enhancement and denoising. Cao et al. [20] developed a retinal image enhancement algorithm for deblurring and contrast improvement. Li et al. [22] proposed another CNN-based method that restores cataract-free images from simulated cataract-like retinal images. In the double pass fundus reflection (DPFR) model [23], the authors proposed a retinal image enhancement method that improves the contrast. These models have been trained to improve the overall quality of the retinal images by decreasing the effects of blurriness, poor illumination, image artifacts, and noise. However, due to the complexity of the task and models' sizes, the generated images often happen to be of low resolution, with image sizes ranging between 256×256 and 512×512 pixels; therefore, retinal image super-resolution models would greatly increase the applicability of the retinal image restoration models.

2.2. Single image super-resolution methods

A lot of works utilized CNN-based models for SISR [6,26–37]. Liu et al. [36] developed a super-resolution model with the non-local self-similarity of images, cross-resolution similarity, and structural sparsity. The authors of Frequency-based Enhancement Network (FENet) [6] proposed a deep learning-based super-resolution method focusing more on restoring the high-frequency information while forwarding the low-frequency details to the output. Hu et al. [37] developed a multi-scale information cross-fusion network (MSICF) that consists of cascaded subnetworks that infer the high-resolution features. The authors of a medium-sized dense network (MADNet) [26] have built

a residual multi-scale module with an attention mechanism for better focus on more informative features. Ledig et al. [27] was the first to utilize GAN, originally proposed by Goodfellow et al. [38], in single-image super-resolution. They developed Super-Resolution GAN (SRGAN) and Super-Resolution Residual Network (SRResNet); in addition, they used perceptual loss using pre-trained VGG19 [39]. Later, Wang et al. [28] introduced Enhanced SRGAN (ESRGAN), in which they utilized Residual-in-Residual Dense Block (RRDB) as the main building block of the model. Dong et al. [40] offered a lightweight CNN-based method for SISR (SRCNN). Dai et al. [29] developed a deep CNN network with second-order channel attention (SOCA) mechanism, which uses second-order feature statistics to learn feature inter-dependencies. The authors of multi-scale GAN (MSGAN) [41] embedded multi-scale Pyramid module into generator model to improve feature extraction.

In another work [30], the authors collected a dataset (RealSR) with real-world low-resolution images along with a Laplacian pyramid-based kernel prediction network (LP-KPN). He et al. [31] concentrated on utilizing numerical methods with CNN. They combined Leapfrog (LF-block) and Runge-Kutta (RK-block) methods with CNN architectures. Hui et al. [32] developed another CNN-based model, namely an information distillation network (IDN), that consists of feature extractors, information distillation blocks, and a reconstructor. Park et al. [33] constructed a generative adversarial network that works in the feature domain with an additional discriminator. Wang et al. [34] proposed another CNN-based progressively growing generative adversarial network (ProGanSR), which progressively learns to increase the resolution of images up to $\times$ 8. The authors of a channel-wise and spatial feature modulation (CSFM) network [35] stacked densely connected feature modulation memory (FMM) modules for better attention. Recently, in some works, the authors utilized only ViT-based models for image super-resolution. For example, Yang et al. [42] proposed a texture learning ViT for natural image super-resolution. In their work, they trained a model to learn how to increase the resolution of images using high-resolution reference images. In other words, it transfers the relevant features of high-resolution to low-resolution images. Zamir et al. [43] developed a ViT-based method for generating high-resolution images from low-quality noisy, blurry images. Recently, Lu et al. [44] developed an Efficient Super-Resolution Transformer (ESRT) method which consists of a Lightweight CNN Backbone (LCB) and a Lightweight Transformer Backbone (LTB); LTB also includes Efficient Multi-Head Attention (EMHA). Gao et al. [45] proposed an efficient CNN-Transformer Cooperation Network (CTCNet) that combines a novel Local-Global Feature Cooperation Module (LGCM) for facial image super-resolution. In another work [46], they proposed a Lightweight Bimodal Network (LBNet) consisting of an effective Symmetric CNN and a Recursive Transformer. However, most of these works have utilized pure CNN architectures. Only recently, most researchers proposed pure ViT networks, which require a significant amount of computational memory and are computationally inefficient.

2.3. Medical image super-resolution methods

Medical image super-resolution methods are also generally represented by CNN-based SISR models [47–54]. Zhang et al. [47] developed a CNN-based fast medical image super-resolution (FMISR) method. In another work [48], researchers developed a GAN-based lesion-focused SR (LFSR) method for brain tumor Magnetic resonance images (MRI) super-resolution. The authors of another research work [49] developed a GAN-based multi-level densely connected super-resolution network (mDCSRN) with a generative adversarial network-guided training. Mahapatra et al. [50] presented a CNN-based progressively growing GAN for retinal images super-resolution with a novel triplet loss function that compares generated images from previous stages with original high-resolution images. The authors of Multiple Improved Residual Network (MIRN) [51] built their CNN-based model with eight improved residual blocks and up-sampling modules. The authors of the improved generative adversarial network (IGAN) [52] proposed a GAN model that has modified residual blocks with attention layers for retinal image super-resolution. Lv et al. [53] developed dense and ReZero (residual with zero initialization) residual networks for super-resolution of retinal images (DRRN). Mahapatra et al. [54] presented another CNN-based GAN model with local saliency maps to build a saliency loss function. Tian et al. [55] proposed a deep learning network based on ESRGAN with their Mixed Attention Block for retinal image super-resolution. The authors of Sparse-based domain Adaptation Super-Resolution network [56] developed a method for super-resolution of retinal optical coherence tomography angiography, proposing a novel adversarial learning-based model with a sparse edge-aware loss function. All medical-related super-resolution methods are developed using CNN architectures. In addition, most of these works have utilized adversarial training to increase the realism of generated images; however, generated medical images must be the same or similar to ground truth images. In this paper, we have built a ViT-based deep learning super-resolution method that does not rely on adversarial learning. Instead, we focused on preserving the structure of medical images rather than classifying if it is real or generated. To do so, we utilized SSIM loss that trains the network to keep the images' structure, which is crucial for medical images.

3. Methodology

3.1. Model overview

The proposed method consists of a vision transformer encoder and a convolutional neural network decoder, as shown in Fig. 2. In this work, we propose a novel training strategy, Adaptive Convolutional Patch Embedding (ACPE) layer, to keep the architecture of the ViT encoder and CNN decoder fixed for all up-scaling factors, starting from 2 to 8. The only part that changes is the patch embedding block that includes one 2D convolutional layer with various kernel and stride sizes for different image sizes and up-scaling factors.

3.2. Adaptive convolutional patch embedding layer

At first, the low-resolution image is fed into the Adaptive Convolutional Patch Embedding (ACPE) layer as shown in Fig. 3; it is utilized to split the input image into non-overlapping patches. Additionally, it is used to embed the resulting image patches, as ViT models lack prior positional information of image patches. The number of input channels changes from 3 to 2048, and both the height and width of resulted features are equal to 64. ACPE has eliminated the need to add additional layers to the model while increasing the super-resolution factor, since we only need to change the patch sizes when increasing the super-resolution factors. This has led to better quantitative and qualitative results and computational performance. The output of patch embedding is fed into the transformer encoder. The output of the ViT network is up-scaled and converted back to a 2048 $\times$ 2048 image in the 5 transposed CNN blocks followed by 2D convolution and a Tanh activation function. To calculate the kernel size and strides, or patch size, of a given input image, we used the following formula:

$$PS = K \times S = \frac{H}{NP} \times \frac{W}{NP} \tag{1}$$

where PS is the patch size, NP is the number of patches equal to 64, K is kernel size, S is strides, H and W are the image height and width, respectively. For example, for image of size 1024 $\times$ 1024 the patch size would be $PS = \frac{1024}{64} \times \frac{1024}{64} = 16 \times 16$. NP has been derived in experiments in which we trained the model with NP being equal to 32, 64, and 128.

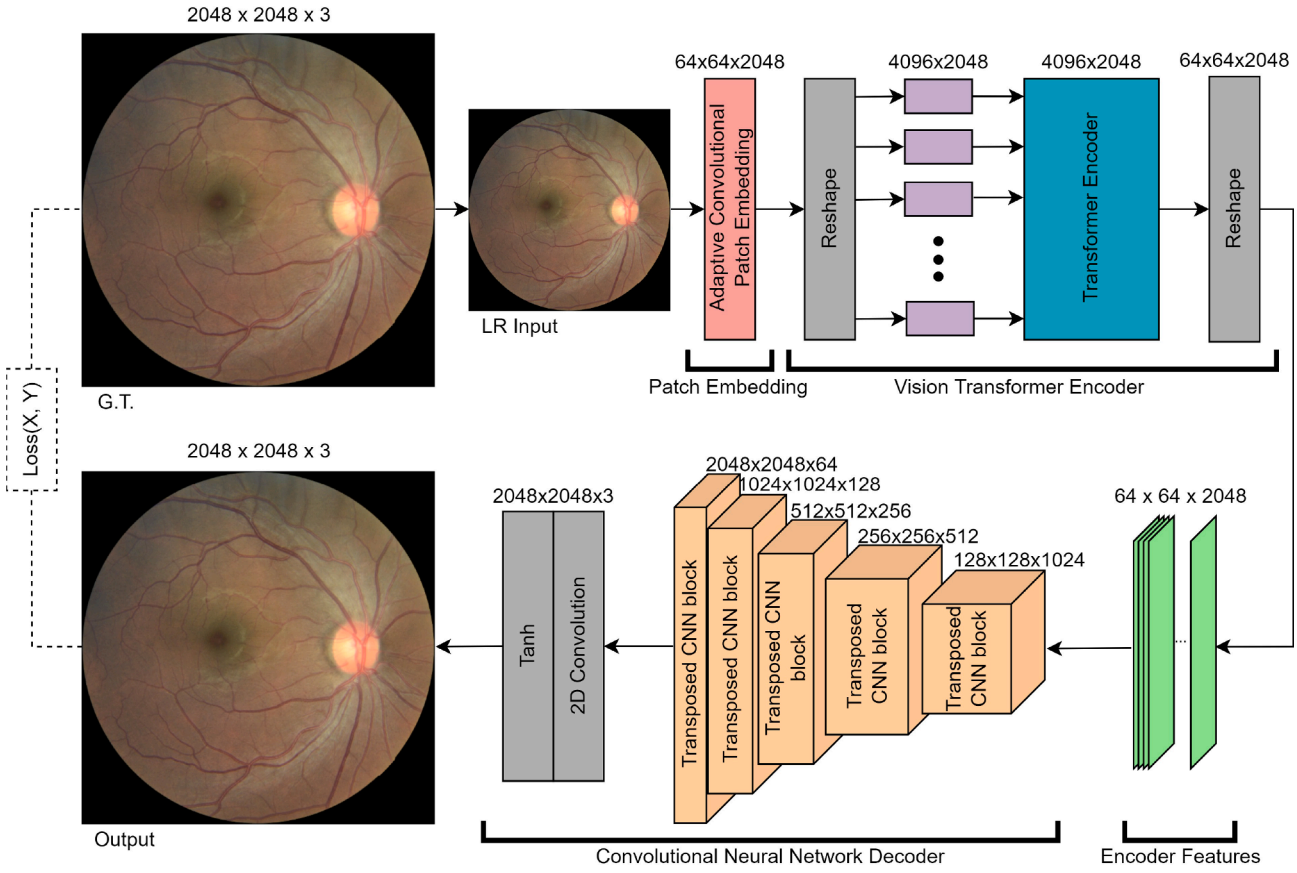

Fig. 2. The proposed super-resolution model consists of a vision transformer encoder and a convolutional neural network decoder. The model remains the same for different up-scaling factors except for the patch embedding. The patch size increases as we increase the image size. For example, for input images of size 256 × 256, the patch size is 4 × 4 to perform × 8 super-resolution.

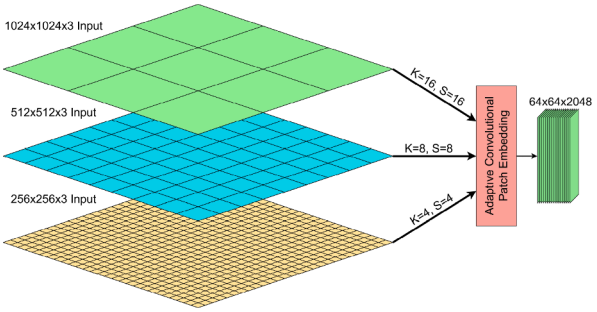

Fig. 3. Adaptive Convolutional Patch Embedding (ACPE). After the input images of different sizes are split into patches by the ACPE layer, the output features have the same shape of 64 × 64 × 2048. K is the kernel size, and S is the value of strides.

3.3. Transformer encoder

The encoder includes locality self-attention [57] (LSA), layer normalization [58], and multilayer perceptron (MLP). LSA was initially proposed for small-sized datasets. However, it has also shown better results than multi-head self-attention when trained with medium-sized datasets thanks to learnable temperature scaling. In this work, we have decided to use LSA for our model to be efficiently trained on the EyeQ dataset, which is considered mid-sized.

The input image $(x) \in \mathbb{R}^{B\times H\times W\times C}$ with number of samples (batch size) B, height H, width W and number of channels C is divided into patches using convolutional layer, flattened across H and W, and transposed. As a result, we obtain patch sequences $(x)_p \in \mathbb{R}^{B\times N\times D}$ with $N = H \times W$ denoting the number of latent vectors and $D = 2048$ denoting the length of each latent vector. The value of D has been derived through several experiments, in which we trained the model with various values, such as 512, 1024, 2048, and 3072; the most optimal result was achieved with 2048. The input patch sequences are divided into queries, keys, and values using a linear layer, as shown in Fig. 4. This layer changes the sequence length D to be equal to hidden vector length D_h, which is equal to the multiplication of the number of heads and their length ($NH \times LH$); NH and LH have been derived during experiments, and the optimal values are 64 and 4, respectively. The queries and keys matrices are multiplied to obtain R_{ij}. LSA uses diagonal masking to improve the attention process. The masking is defined in Eq. (3). The diagonal elements are set to $-\infty$. Using the formula 4, we get attention values for R. After the dropout layer, the output is multiplied

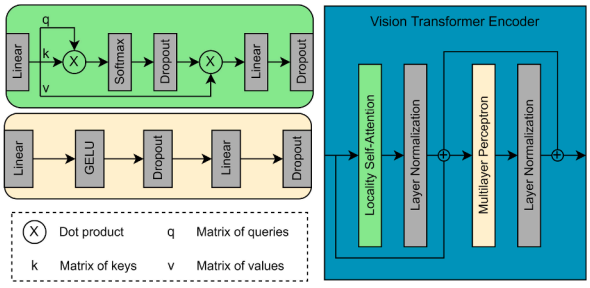

Fig. 4. The architecture of the ViT encoder consists of local self-attention, layer normalization, and multilayer perceptron layers. After dividing the image into patches, the features are fed into the ViT encoder.

by values. The resulting outcome is sent to the final linear and dropout layers. Linear layer changes the shape of attention maps, restoring it to its original size of 4096 × 2048. In our model, the dropout rate is 0. The output features are normalized using layer normalization using the following formula:

$$\widehat{X} = \frac{X - \mu(X)}{\sqrt{\sigma(X) + \epsilon}} * \gamma + \beta \tag{2}$$

where $\widehat{X}$ is the normalized features, X is the input features, $\mu(X)$ is the mean, $\sigma(X)$ is the standard deviation, ϵ is a small value to avoid division by zero, γ and β are learnable parameters for affine transformations that are derived during training.

$$R_{i,j}^{Masked} = \begin{cases} R_{i,j} & i \neq j \\ -\infty & i = j \end{cases} \tag{3}$$

where $R_{i,j}$ represents the components of the dot product between the matrix of queries and transposed matrix of keys $R = q \times k^T$.

$$LSA = Softmax\left(\frac{R^{Masked}}{\tau}\right) \tag{4}$$

where τ is learnable temperature scaling, another feature of LSA that helps the softmax calculate its temperature during training.

The normalized features are concatenated with the encoder input and sent to MLP. The structure of MLP is presented in Fig. 4, the first linear layer changes the dimension D of input features from 2048 to hidden dimension of MLP $D_h^{MLP} = 2048 \times 4$, where 4 is the MLP ratio that was derived in experiments. Next, they are sent to the Gaussian Error Linear Unit (GELU) [59] activation function and dropout. Another linear layer reduces dimensions to the initial size D with dropout. Dropout rates are equal to 0. The depth of our ViT encoder is equal to 1 after testing it with depths of 2, 5, and 7.

3.4. CNN Decoder

The proposed method utilizes CNN for decoding the ViT output features. The training was performed with images of different sizes, such as 256 × 256, 512 × 512, and 1024 × 1024. The patch embedding layer has been adjusted according to the image size. For example: for 256 × 256 image size, the patch size would equal 4 × 4 according to 1, and for 512 × 512, the patch size would be 8 × 8. It allowed fixing our model's encoder and decoder parts, changing only the patch embedding layer. Our convolutional decoder consists of 5 consecutive 2D transpose convolution blocks and a final 2D convolution with the tanh activation function as presented in Fig. 5. The structure of all blocks is the same and includes 2D transpose convolution, instance normalization, and ReLU activation function; the number of kernels is 3, strides are 2, and padding is 1. Each CNN block up-scales the input by a factor of 2. For instance, the output shape of an encoder is 2048 × 64 × 64. After the first block, it will be of shape 1024 × 128 × 128, and the final block produces a 64 × 2048 × 2048 matrix. In the end, it will be fed into the last layer with 2D convolution and tanh activation function, which keeps the input's height and width but reduces the channels to 3. Tanh activation function makes the input values in the range between -1 and 1.

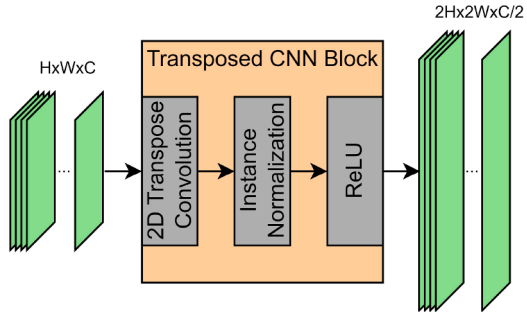

Fig. 5. The structure of the CNN decoder includes 2-dimensional transpose convolution, instance normalization, and ReLU activation function layers. Each block increases the dimensions of the features (green) by a factor of 2 and reduces the number of these features by 2. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

3.5. Loss functions

To train our retinal image super-resolution model, a combination of two loss functions was used: SSIM loss and L1 loss. The SSIM loss can be expressed using generated image X and original high-resolution image Y:

$$SSIM_{Loss}(X, Y) = \frac{1 - SSIM(X, Y)}{2} \tag{5}$$

where $SSIM(X, Y)$ is calculated using the following formula:

$$SSIM(X, Y) = \frac{(2\mu_X\mu_Y + C_1)(2\sigma_{XY} + C_2)}{(\mu_X^2 + \mu_Y^2 + C_1)(\sigma_X^2 + \sigma_Y^2 + C_2)} \tag{6}$$

where μ is the mean, σ is the standard deviation, C_1 and C_2 are variables that stabilize the division. The window size is equal to 3, which means the loss is computed with 3 × 3 windows, and the average value is derived.

The second loss function that we used is L1 loss. In L1 loss, the average value of absolute difference between Y and X is derived using the following equation:

$$L1_{Loss}(X, Y) = \frac{1}{N}\sum_{n=1}^{N} |Y - X| \tag{7}$$

where N is the total number of pixels.

The total loss function is calculated using the following mathematical expression:

$$L = \lambda_1 \frac{1 - \frac{(2\mu_X\mu_Y + C_1)(2\sigma_{XY} + C_2)}{(\mu_X^2 + \mu_Y^2 + C_1)(\sigma_X^2 + \sigma_Y^2 + C_2)}}{2} + \lambda_2 \frac{\sum_{n=1}^{N} |Y - X|}{N} \tag{8}$$

where λ_1 and λ_2 are weights equal to 10 as optimal value through experiment.

4. Datasets and evaluation metrics

4.1. Datasets

To train our super-resolution network, we used the publicly available EyeQ dataset [10]. The dataset consists of 28,792 retinal images divided into three classes according to their qualities. There are 16,818 "Good" quality, 6434 "Usable," and 5538 "Bad" retinal images in this dataset. We used "Good" and "Usable" images for training, validation, and testing. We used 22,152 images for training, 100 for validation at each epoch, and 1000 for quantitative and qualitative testing. EyeQ images have a size of 2048 × 2048 pixels. During training, the images have been augmented using random horizontal and vertical flipping and normalized to be in the range between -1 and 1.

To further validate the performance of our model in the second testing stage, we used three other publicly available retinal vessel segmentation datasets, such as DRIVE [12], STARE [13], and CHASE DB1 [14]. The low-resolution images have been up-scaled by a factor of 4 and segmented. The vessel segmentation results have been qualitatively and quantitatively compared with high-resolution segmented retinal images. DRIVE dataset includes 40 image pairs of resolution 584 × 565 pixels. STARE dataset comprises 20 image pairs with a resolution of 700 × 605

pixels. CHASE DB1 has 28 image pairs with a size of 999 ×960 pixels. Datasets details are presented in Table 1. All images have been resized to 512 × 512 and up-sampled to 2048 × 2048.

4.2. Evaluation metrics

To validate and compare the proposed method results, we performed qualitative analysis and utilized quantitative evaluation metrics such as PSNR, SSIM, SITT, Jaccard similarity index measure, MCC, kappa, F1-score, precision, recall, and accuracy. The PSNR, SSIM, and SITT are calculated for the EyeQ dataset. In contrast, Jaccard, F1-score, MCC, kappa, precision, recall, and accuracy are calculated between manually segmented and up-scaled segmented images. Jaccard, MCC, and kappa are more accurate at evaluating the classification methods since they do not suffer from sensitivity to class imbalance and asymmetricity, unlike precision, recall, F1, and accuracy. The SSIM and PSNR are calculated using 6 and 9, respectively.

$$PSNR(X,Y) = 10\log_{10}\left(\frac{MAX_X^2}{MSE}\right) \tag{9}$$

where MAX_X^2 is the maximum value of a pixel and MSE is calculated in the following formula:

$$MSE(X,Y) = \frac{1}{N}\sum_{n=1}^{N}(Y-X)^2 \tag{10}$$

where N is the total number of pixels.

The jaccard index shows the similarity between two sets and is defined in the following equation:

$$J(X,Y) = \frac{|X \cap Y|}{|X \cup Y|} \tag{11}$$

Precision and recall, F1-score, and accuracy are calculated using Eqs. (12); while MCC and kappa coefficient are computed using Eq. (16) and Eq. (17), respectively.

$$Precision = \frac{TP}{TP+FP} \tag{12}$$

$$Recall = \frac{TP}{TP+FN} \tag{13}$$

$$F1 = 2\frac{Precision * Recall}{Precision + Recall} \tag{14}$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{15}$$

where TP is a true positive, TN means true negative, FP is a false positive, and FN means false negative values.

$$MCC = \frac{TP*TN - FP*FN}{\sqrt{(TP+FP)(TP+FN)(TN+FN)(TN+FN)}} \tag{16}$$

$$kappa = \frac{Accuracy - Accuracy^R}{1 - Accuracy^R} \tag{17}$$

Table 1
Details of datasets used for the experiment.

Name	Number of Images	Image Size	Task
EyeQ [10]	22,152	2048 × 2048	Train
	100	2048 × 2048	Validation
	1000	2048 × 2048	Test
DRIVE [12]	40	584 × 565	Test
STARE [13]	20	700 × 605	Test
CHASE DB1 [14]	28	999 × 960	Test

where random accuracy $Accuracy^R$ is calculated in the following way:

$$Accuracy^R = p_1 \times p_2 + (1-p_1) \times (1-p_2) \tag{18}$$

$$p_1 = \frac{TP+FN}{TP+TN+FP+FN} \tag{19}$$

$$p_2 = \frac{TP+FP}{TP+TN+FP+FN} \tag{20}$$

To perform qualitative analysis, we used high-resolution ground truth images from the EyeQ dataset [10]. We compared them with up-scaled output images of our and other state-of-the-art methods. In addition, we resized images of DRIVE [12], STARE [13], and CHASE DB1 [14] datasets to 512 × 512 pixels and up-scaled them by a factor of 4 to 2048 × 2048. Then, we performed vessel segmentation of these images using UNet [60]. The resulting vessel trees have been qualitatively compared with manually segmented ground truth images of vessel trees.

4.3. Training details

The model has been implemented using the PyTorch framework. The hardware configurations are an Intel Core i7-10700f CPU, 32 GB of RAM, and an NVIDIA GeForce RTX 2080 SUPER 8 GB. The model was set to train for 100 epochs for all factors, including × 2, × 4, and × 8 with early stopping. If validation SSIM and PSNR do not increase for consecutive 4 epochs, the training with the current factor stops. As a result, the training processes with a factor of 2 took 41 epochs, a factor of 4 needed for 35 epochs, and 8 lasted for 8 epochs. The training took 10 days in total for all super-resolution factors.

In our experiments, we used the Adam optimization function, batch size of 1, and learning rate of 10^{-5}. During training input images were resized to 1024 × 1024, 512 × 512, 256 × 256 for × 2, × 4, and × 8 super-resolution, respectively; data augmentation techniques include random horizontal and vertical flipping. In the transformer encoder, latent vector length D is equal to 2048, the number of latent vectors N is 4096, the number of heads is 4, the dimension of each head is 64, and the MLP ratio is 4.

5. Results and discussion

The proposed method has been qualitatively and quantitatively evaluated and compared with state-of-the-art methods, including SRResNet [27], SRGAN [27], ESRGAN [28], and recent FENet [6], and ESRT [44]. These super-resolution methods have been trained and tested in the same experimental setup for a fair comparison. The quantitative results are presented in Tables 2–4.

5.1. Computational performance

In Table 2, it is evident that the proposed method is the least time-consuming than other methods. Additionally, our method has the highest number of parameters due to embedding a ViT encoder that normally has more parameters than CNN. However, the proposed model maintains comparatively low GPU memory consumption for all super-resolution factors. In particular, for × 2, the proposed model is second in terms of memory efficiency after ESRGAN [28]; for ×4 super-resolution, it comes after ESRGAN [28] and FENet [6]; finally, for ×8 the proposed model still outperforms another ViT-based method ESRT [44]. On average, the proposed model takes testing time of 0.132, 0.122, and 0.117 seconds per image for 2, 4, and 8 up-scaling factors, respectively. Our proposed method significantly differs from the other existing methods [6,27,28]. Particularly, ESRT [44] and ESRGAN [28] require more than 6.4 and 4.8 seconds for ×2 single image super-resolution, respectively; this is almost 49 and 37 times longer than our result. Both SRGAN [27] and SRResNet [27] up-sample one image in

Table 2
Comparison of Single Image Testing Time of EyeQ test set [10] (SITT) in seconds, GPU Memory consumption and Number of Parameters. The **bold** and underlined numbers represent the best and second-best results.

Factor	Method	SITT ↓	Memory ↓	Parameters
× 2	SRResNet [27]	0.484	7.8 Gb	27,400,001
	SRGAN [27]	0.484	7.8 Gb	24,965,508
	ESRGAN [28]	4.874	**6.3** Gb	40,222,852
	FENet [6]	0.273	7.9 Gb	638,086
	ESRT [44]	6.443	7.8 Gb	677,783
	Ours	**0.132**	7.0 Gb	62,399,620
× 4	SRResNet [27]	0.24	7.3 Gb	27,547,457
	SRGAN [27]	0.246	7.1 Gb	25,113,284
	ESRGAN [28]	1.267	**4.7** Gb	40,259,780
	FENet [6]	0.255	5.0 Gb	675,462
	ESRT [44]	1.801	7.8 Gb	751,767
	Ours	**0.122**	7.0 Gb	61,219,972
× 8	SRResNet [27]	0.181	6.0 Gb	27,694,913
	SRGAN [27]	0.19	5.8 Gb	25,261,060
	ESRGAN [28]	0.381	**4.1** Gb	40,296,708
	FENet [6]	0.247	5.3 Gb	712,838
	ESRT [44]	0.678	7.8 Gb	825,751
	Ours	**0.117**	7.0 Gb	60,925,060

Table 3
Quantitative results of EyeQ test set [10]. The **bold** and underlined numbers represent the best and second-best results.

Factor	Method	PSNR ↑	SSIM ↑	SITT ↓
× 2	SRResNet [27]	37.858 dB	0.917	0.484
	SRGAN [27]	37.355 dB	0.933	0.484
	ESRGAN [28]	38.324 dB	0.95	4.874
	FENet [6]	40.01 dB	0.96	0.273
	ESRT [44]	46.131 dB	**0.983**	6.443
	Ours	**46.146** dB	0.982	**0.132**
× 4	SRResNet [27]	34.755 dB	0.866	0.24
	SRGAN [27]	35.392 dB	0.909	0.246
	ESRGAN [28]	34.706 dB	0.895	1.267
	FENet [6]	35.607 dB	0.918	0.255
	ESRT [44]	42.376 dB	0.963	1.801
	Ours	**43.336** dB	**0.977**	**0.122**
× 8	SRResNet [27]	33.087 dB	0.879	0.181
	SRGAN [27]	33.88 dB	0.897	0.19
	ESRGAN [28]	32.574 dB	0.844	0.381
	FENet [6]	31.015 dB	0.876	0.247
	ESRT [44]	38.679 dB	0.941	0.678
	Ours	**38.89** dB	**0.942**	**0.117**

about 0.49 seconds, which is almost 4 times longer than the proposed method. FENet [6] up-scales an image by a factor of 2 in 0.273 seconds. As we can see, our model outperforms all other methods in terms of SITT for all super-resolution factors by a large margin.

5.2. Quantitative results

In Table 3, we can observe that our model achieves the highest PSNR and SSIM values in most cases for all given up-scaling factors; only for × 2 super-resolution it is slightly (0.001) outperformed by ESRT [44] in terms of SSIM. For × 2, × 4, ×8 super-resolution, the average testing PSNR values are 46.146 dB, 43.336 dB, 38.89 dB, and SSIM values are 0.982, 0.977, 0.942, respectively. The proposed method, along with ESRT [44], which is also based on ViT, have shown very close results that are higher than those of other models. This is due to embedding ViT into the models as it has proven to outperform CNN in computer vision tasks; the experiments have proven this. The SSIM values of our model do not change dramatically when the up-scaling factors increase. It is significant for medical images as the overall image structure is generally unchanged. As we can observe from Fig. 6, SSIM loss has gradually decreased across all epochs. The green line represents SSIM loss values for × 2 up-scaling, the red curve demonstrates × 4 up-sampling progress, and the blue line shows × 8 super-resolution loss values; the lengths of these lines determine the number of training epochs.

In addition, we performed vessel segmentation using UNet [60] to validate the quantitative and qualitative results of the proposed method. Firstly, we resized the images to 512 × 512 pixels and up-scaled by a factor of ×4 to 2048 × 2048. Table 4 presents the evaluation results with other methods [6,27,28,44]. The presented evaluation metrics include the Jaccard similarity index, F1-score, MCC, kappa, precision, recall, and accuracy. Almost all the models performed well, and there is no significant difference in evaluation metrics. However, if we take into account the computational time from Table 2, in which our model significantly outperforms all other models in terms of SITT, we may conclude that the proposed model utilizes its performance more efficiently compared to other methods as SITT is notably lower and our evaluation results are competing. For the DRIVE dataset [12], our method has shown competitive results with the model ESRGAN [28] and SRResNet [27]; in particular, our method obtains the highest values of Jaccard, MCC, and kappa which demonstrates strongest positive correlation and similarity between classes. For the STARE dataset [13], the model ESRGAN [28] has slightly outperformed our method with minor differences in its results while up-scaling the images. However, our proposed model performs better at up-sampling images when compared with all the other models used for CHASE DB1 [14], having the highest Jaccard, MCC, and kappa rates. In this evaluation stage, the proposed hybrid model along with ESRGAN [28], SRResNet [27], and ESRT [44] have shown favorable outcomes; the state-of-the-art methods have deeper and more complex architectures, which leads to higher computational time compared to ours.

5.3. Qualitative results

Fig. 7 demonstrates the qualitative results of given SR methods, where we can see the superiority of our model. All methods performed satisfactorily on ×2 super-resolution. However, a significant difference is noticeable in ×4 and ×8 super-resolution. For instance, SRGAN [27], SRResNet [27], ESRGAN [28], and ESRT [44] generate good-quality images making the blood vessels visible. On the other hand, if we compare these images with original high-resolution ground truth images, we can see the high presence of noise, while ESRT produces slightly over-smoothed images. FENet [6] produces blurry images, leading to the loss of important information, especially tiny blood vessels. Our method produces images closest to the ground truth and has the highest SSIM value, which is crucial for medical images.

For qualitative comparison, the ground truth images and the images obtained using all the other methods used for comparative analysis are given in Fig. 8. As we can notice from Fig. 8, the visual illustration of comparison of our model along with SRGAN, ESRGAN, and SRResNet are similar to the original (ground truth) images. However, FENet has less visible retinal blood vessels caused by blurred output. Hence, taking into account the computational efficiency in terms of time and memory usage of our proposed model that is demonstrated in Table 2, we can consider it efficient based on its quantitative and qualitative analysis results.

5.4. Ablation study

To further validate the effectiveness of the ViT encoder, we performed an ablation study. To do so, we have built two additional models to validate the effectiveness of embedding the ViT encoder and utilizing adaptive patch embedding. As a result, we removed the ViT part from the proposed model with ViT and an adaptive patch layer (AP) to develop the first model. To build the second model, we used the proposed model and fixed the patch size to equal 16. As the super-resolution factor increases, we apply progressively growing architecture appending an additional layer to the CNN decoder part. We trained both models on the same data in the same conditions. The results are quantitatively and

Table 4
Quantitative results comparison for ×4 up-scaled images of DRIVE [12], STARE [13], CHASE DB1 [14] datasets. The **bold** and underlined numbers represent the best and second-best results.

Data	Method	Jaccard ↑	MCC ↑	Kappa ↑	F1-score ↑	Precision ↑	Recall ↑	Accuracy ↑
DRIVE [12]	SRResNet [27]	0.6642	0.7765	0.7764	**0.7982**	0.8080	0.7832	0.9653
	SRGAN [27]	0.6489	0.7692	0.7686	0.7869	**0.8191**	0.7593	0.9643
	ESRGAN [28]	0.6641	0.7813	0.7811	0.798	0.8171	0.7835	**0.9656**
	FENet [6]	0.4935	0.6344	0.6325	0.6602	0.7134	0.6183	0.9448
	ESRT [44]	0.6575	0.7770	0.7756	0.7942	0.7540	**0.8405**	0.9646
	Ours	**0.6643**	**0.7820**	**0.7818**	0.7981	0.8186	0.7834	**0.9656**
	Original images	0.6659	0.7833	0.7830	0.7992	0.8213	0.7829	0.9659
STARE [13]	SRResNet [27]	0.6501	0.7709	0.7709	0.791	0.7813	0.795	**0.9682**
	SRGAN [27]	0.6122	0.7419	0.7418	0.7576	0.7686	0.7539	0.9643
	ESRGAN [28]	0.6509	**0.7744**	**0.7744**	**0.7913**	**0.7854**	0.7981	0.9681
	FENet [6]	0.5153	0.6560	0.6556	0.6771	0.7034	0.6595	0.9534
	ESRT [44]	**0.6544**	0.7674	0.7673	0.7902	0.7746	**0.799**	0.9614
	Ours	0.6481	0.7692	0.7692	0.7848	0.7805	0.793	0.9676
	Original images	0.6656	0.7839	0.7839	0.7979	0.7997	0.8010	0.9697
CHASE DB1 [14]	SRResNet [27]	0.6086	0.7419	0.7399	0.7565	0.7090	0.814	0.9664
	SRGAN [27]	0.5881	0.7240	0.7228	0.7404	0.7035	0.7843	0.9647
	ESRGAN [28]	0.6052	0.7240	0.7371	0.7538	0.7108	0.8055	0.9662
	FENet [6]	0.4697	0.6158	0.6158	0.6389	0.6501	0.6307	0.9542
	ESRT [44]	0.6096	0.7390	0.7365	0.7599	0.7009	0.8159	**0.9671**
	Ours	**0.6135**	**0.7460**	**0.7440**	**0.7603**	**0.7138**	**0.8164**	0.967
	Original images	0.6077	0.7409	0.7391	0.7558	0.7116	0.8088	0.9665

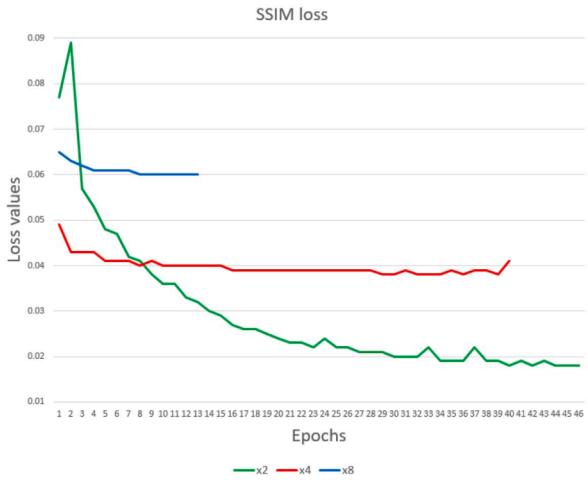

Fig. 6. Validation loss computed for each super-resolution factor at each epoch using formula 5. In this figure, the red line is for × 2, the green line is for × 4, and the blue line is for × 8 super-resolution loss.. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

qualitatively compared. Fig. 9 demonstrates the validation performance of 3 models across all training epochs for × 4 super-resolution factor. As we can observe, the proposed model incorporating a ViT encoder with AP and CNN decoder gets higher validation PSNR and SSIM values.

As we can observe, the ViT-based model with AP outperforms the ViT-based model with fixed patch size and the CNN-based method in terms of PSNR and SSIM as presented in Table 5. As we can see in Table 6, the CNN-based method is superior to the ViT-based model with fixed patch size in the DRIVE [12] dataset in terms of Jaccard, MCC, and kappa. However, it is still outperformed by the ViT model with AP. When testing with STARE [13] dataset, the proposed model has the highest Jaccard value, but it is slightly outperformed by the CNN-based method regarding MCC and kappa. Finally, the proposed method has shown superior results in CHASE DB1 [14] dataset. The proposed method has outperformed other models of Jaccard, MCC, and kappa, proving its superior similarity and positive correlation between segmentation classes. The CNN-based method takes the least time for a single image super-resolution. The average difference between the proposed and CNN-based methods in SITT is only 0.013 seconds. Particularly, the proposed method takes 0.132, 0.122, and 0.117 seconds for each image to do × 2, × 4, and ×8 super-resolution, respectively, which is slightly more than the CNN model that took 0.12, 0.109, 0.101 seconds, respectively. We have also performed statistical analysis, including a one-way variance analysis (ANOVA). In our experiments, a p-value less than 0.05 is considered statistically significant. The results are shown as mean ± standard deviation. As we can observe from Fig. 10 that shows statistical results, utilizing the ViT encoder with AP has shown better performance. In particular, for × 4 super-resolution, the values of mean and standard deviation for PSNR are 43.336 ± 1.80, 41.4 ± 1.67, 40.723 ± 2.03 for ViT+AP, ViT, and CNN+AP models, respectively. For SSIM, the values are 0.977 ± 0.01, 0.953 ± 0.02, and 0.949 ± 0.02 for ViT+AP, ViT, and CNN+AP models, respectively. In all cases, the p-value equals 0, which proves a significant difference between these models.

As can be inferred from Fig. 11, ViT-based models perform better at preserving the structures of the retinal images, particularly retinal blood vessels. The resulting images are smoother and contain less noise than the CNN model. In addition, the CNN-based model produces less sharp, blurry results, missing some tiny vessels compared to our method. Our method that includes ViT and AP has shown higher quality results than the model with ViT but with no AP since some small vessels are less visible than the proposed method's outcomes. The difference can be noticed in ×8 up-scaling, where some small blood vessels are not as straightforward as in the proposed ViT-based model's output. Therefore, these results prove the effectiveness of embedding the ViT encoder and AP layer into our model.

5.5. Limitations

The proposed method can accurately increase retinal images' resolution by factors such as × 2, × 4, and × 8, keeping high SSIM values. However, the model might perform worse when it comes to up-sampling images with tiny vessels with bright backgrounds as we increase the up-scaling factor. An example is demonstrated in Fig. 12. Increasing the super-resolution factor, small blood vessels in the disc cup area become less visible. An unbalanced dataset causes it since most images do not have bright optic disc regions. This challenge can be solved by increasing the high-contrast images with tiny vessels by synthesizing them and efficiently using more advanced methods. We plan to address this issue in future works.

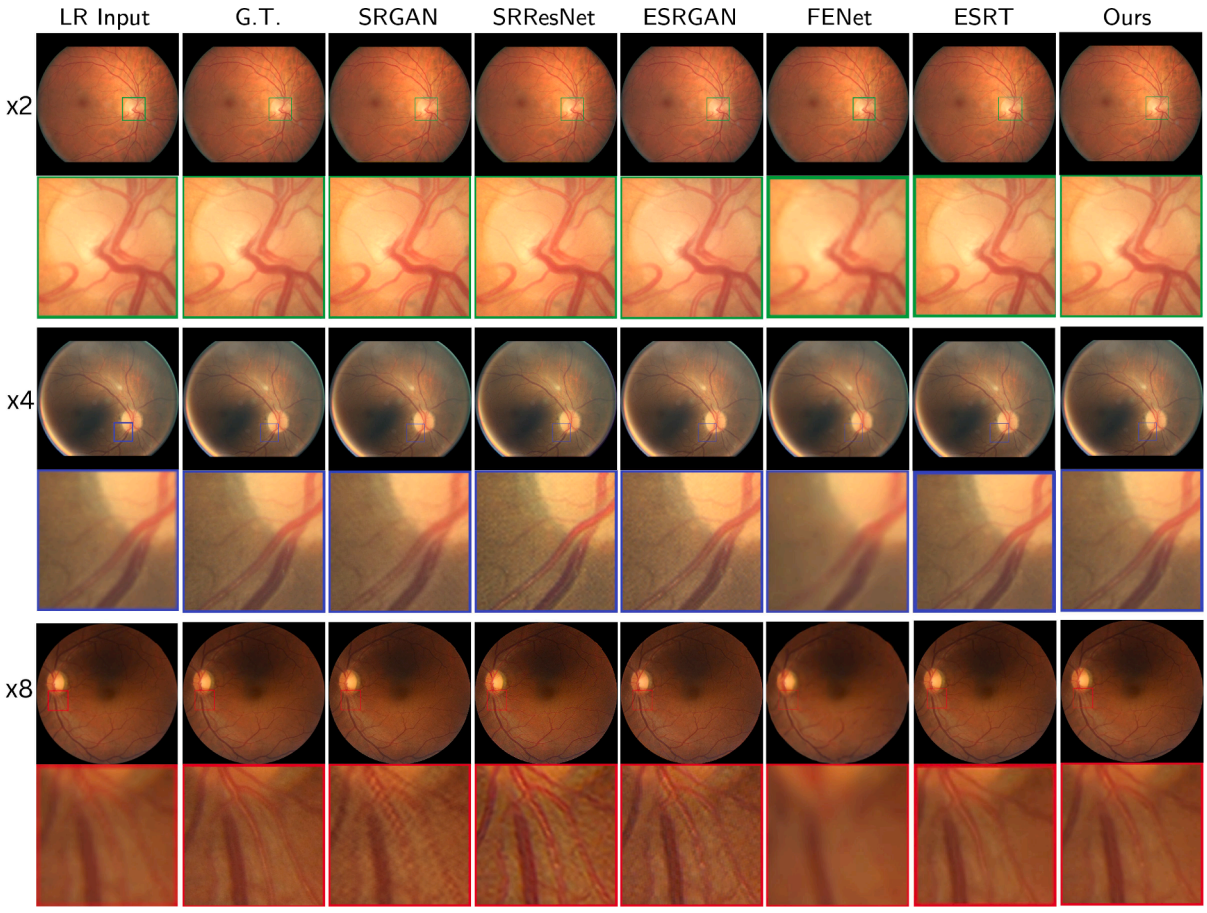

Fig. 7. Qualitative comparison of Ground Truth (G.T.) with SRGAN [27], SRResNet [27], ESRGAN [28], FENet [6], ESRT [44] and ours. For better visualization, we cropped parts of high-resolution images.

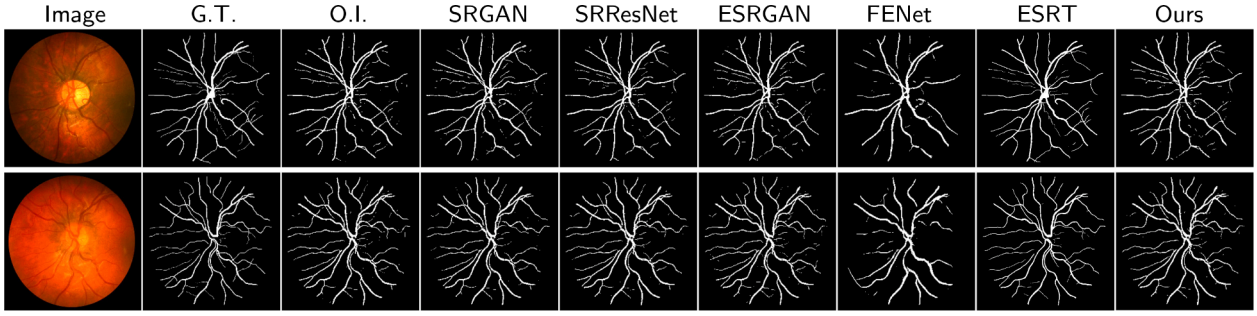

Fig. 8. Qualitative comparison of segmentation results of Ground Truth (G.T.) with Original Image, SRGAN [27], SRResNet [27], ESRGAN [28], FENet [6], ESRT [44] and ours.

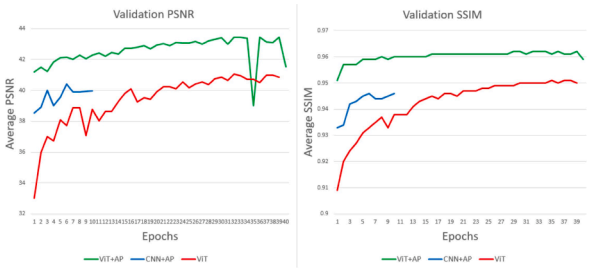

Fig. 9. Comparison of validation results between the proposed ViT with AP, CNN with AP, and ViT without AP models during training..

Table 5
Quantitative ablation study results of EyeQ test set [10]. The **bold** and underlined numbers represent the best and second-best results.

Factor	ViT	AP	PSNR ↑	SSIM ↑	SITT (sec) ↓
× 2	✓	✓	**46.146** dB	**0.982**	0.132
	✓	×	**46.146** dB	**0.982**	0.132
	×	✓	41.619 dB	0.949	**0.12**
× 4	✓	✓	**43.336** dB	**0.977**	0.122
	✓	×	41.4 dB	0.953	0.114
	×	✓	40.723 dB	0.949	**0.109**
× 8	✓	✓	**38.89** dB	**0.942**	0.117
	✓	×	38.527 dB	0.934	0.108
	×	✓	37.906 dB	0.936	**0.101**

Table 6
Quantitative ablation study results ×4 up-scaled images of DRIVE [12], STARE [13], CHASE DB1 [14] datasets. The **bold** and underlined numbers represent the best and second-best results.

Dataset	ViT	AP	Jaccard ↑	MCC ↑	Kappa ↑	F1-score ↑	Precision ↑	Recall ↑	Accuracy ↑
DRIVE [12]	×	✓	0.6622	0.7793	0.7791	0.7938	0.8155	0.7819	0.9649
	✓	×	0.6035	0.7095	0.7091	0.7322	0.7584	0.7110	0.9546
	✓	✓	**0.6643**	**0.7820**	**0.7818**	**0.7981**	**0.8186**	**0.7834**	**0.9656**
	Original images		0.6659	0.7832	0.7829	0.7992	0.8213	0.7829	0.9659
STARE [13]	×	✓	0.645	0.7675	0.7675	0.7835	0.7798	**0.7915**	0.9667
	✓	×	0.6471	**0.7694**	**0.7693**	0.7843	**0.7824**	0.7915	**0.9676**
	✓	✓	**0.6481**	0.7692	0.7692	**0.7848**	0.7805	**0.793**	**0.9676**
	Original images		0.6656	0.7839	0.7839	0.7979	0.7997	0.8010	0.9697
CHASE DB1 [14]	×	✓	0.6099	0.7429	0.7410	0.7576	0.7125	0.8119	0.9666
	✓	×	0.6107	0.7434	0.7416	0.7581	0.7132	0.8121	0.9667
	✓	✓	**0.6135**	**0.7460**	**0.7440**	**0.7603**	**0.7138**	**0.8164**	**0.967**
	Original images		0.6077	0.7409	0.7391	0.7558	0.7116	0.8088	0.9665

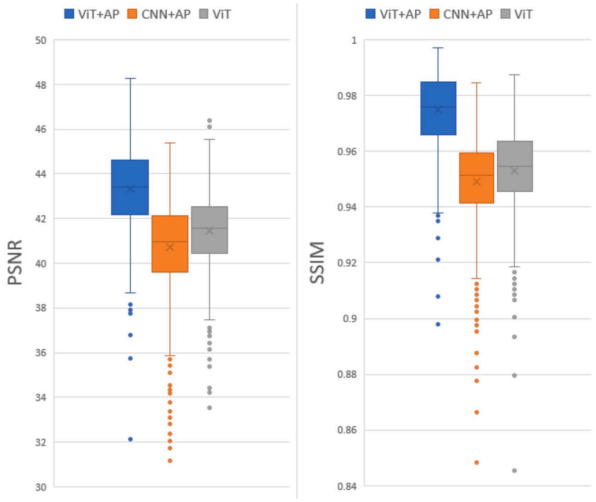

Fig. 10. Comparison of the image quality of each model for × 4 super-resolution. From left to right: box plots of PSNR and SSIM values..

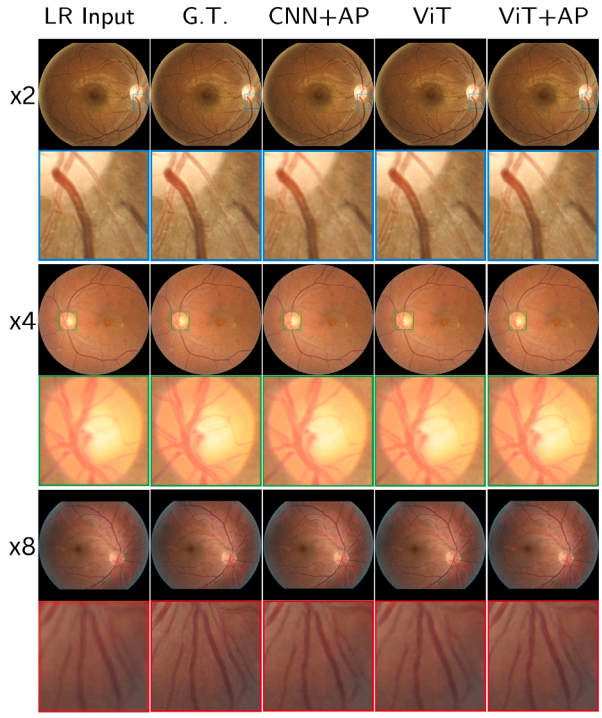

Fig. 11. Qualitative comparison of ablation study. Left to right: Low-Resolution Input, Ground Truth, the model with CNN and AP, ViT encoder without AP, and the proposed method with ViT and AP. From top to bottom: × 2, × 4, ×8 up-scaling.

6. Conclusion

This study focused on accurate and efficient retinal image super-resolution by proposing a model that combines the ViT encoder and CNN decoder. To perform × 2, × 4, and ×8 super-resolution, we offer an adaptive patch embedding layer that removes the need to build additional decoding layers to increase the super-resolution factors. Our method has led to significantly greater quantitative and qualitative results than current state-of-the-art methods. It can accurately up-sample retinal images by factors of × 2, × 4, and ×8 without an increase in testing time and without significant loss of structural information of retinal images. In addition, we performed an extensive ablation study. The results presented in this paper prove the effectiveness of embedding the ViT encoder and non-fixed patch size. Compared with other methods, our model outperforms them regarding PSNR, SSIM, and testing time. Our method has shown very competitive results during the evaluation of segmented images. In future works, we plan to solve this model's current limitations related to generating tiny vessels in retinal images' bright disc and cup region. We intend to solve it by normalizing the dataset, additional image pre-processing, retinal image synthesis, and more advanced methods. In the future, we intend to develop a more advanced method that is still computationally efficient, fast, and accurate for single-image super-resolution.

Funding

This work is partially supported by the Scientific and Technological Research Council of Turkey (TUBITAK) under the 2232 Outstanding Researchers program, Project No. 118C301.

Abbreviations

CNN: Convolutional Neural Network; ViT: Vision Transformer; SSIM: Structural Similarity Index Measure; PSNR: Peak Signal-to-Noise Ratio; FENet: Frequency-based Enhancement Network; SITT: Single Image Testing Time; GAN: Generative Adversarial Network; SISR: Single Image Super-Resolution; SRGAN: Super-Resolution Generative Adversarial Network; SRResNet: Super-Resolution Residual Network; ESRGAN: Enhanced Super-Resolution Generative Adversarial Network; MLP: Multilayer Perceptron; GELU: Gaussian Error Linear Unit; MSE: Mean Squared Error; LSA: Locality Self-Attention, MCC: Matthew's correlation coefficient, ESRT: Efficient Super-Resolution Transformer.

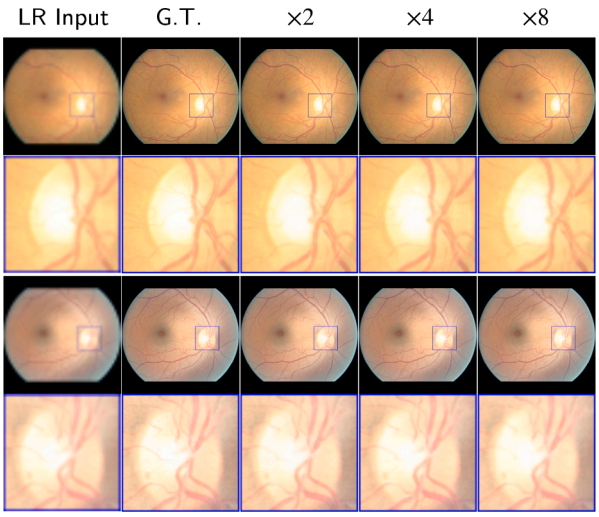

Fig. 12. From left to right: Low-Resolution Input, Ground truth, Generated × 2, Generated × 4, and Generated × 8. An example of the limitations of our model. When the background is bright, tiny vessels are less visible.

Availability of data and materials

The source code for this work is available at https://github.com/AAleka/Retinal-Image-Super-Resolution-ViT-CNN.

Ethics approval and consent to participate

Not applicable.

Consent for publication

All authors read and approved the final manuscript.

CRediT authorship contribution statement

Alnur Alimanov: Methodology, Formal analysis, Writing – original draft. **Md Baharul Islam:** Conceptualization, Investigation, Supervision, Writing – review & editing. **Nirase Fathima Abubacker:** Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be available after publication.

References

[1] H. Wang, J. Chhablani, W.R. Freeman, C.K. Chan, I. Kozak, D.-U. Bartsch, L. Cheng, Characterization of diabetic microaneurysms by simultaneous fluorescein angiography and spectral-domain optical coherence tomography, Am. J. Ophthalmol. 153 (5) (2012) 861–867.
[2] K. Usui, K. Ogawa, M. Goto, Y. Sakano, S. Kyougoku, H. Daida, Quantitative evaluation of deep convolutional neural network-based image denoising for low-dose computed tomography, Visual Comput. Ind. Biomed. Art 4 (1) (2021) 1–9.
[3] Z. Tang, J. Zhang, W. Gui, Selective search and intensity context based retina vessel image segmentation, J. Med. Syst. 41 (3) (2017) 47.
[4] F. Grassmann, J. Mengelkamp, C. Brandl, S. Harsch, M.E. Zimmermann, B. Linkohr, A. Peters, I.M. Heid, C. Palm, B.H.F. Weber, A deep learning algorithm for prediction of age-related eye disease study severity scale for age-related macular degeneration from color fundus photography, Ophthalmology 125 (9) (2018) 1410–1420.
[5] D. Maji, A.A. Sekh, Automatic grading of retinal blood vessel in deep retinal image diagnosis, J. Med. Syst. 44 (10) (2020) 180.
[6] P. Behjati, P. Rodriguez, C.F. Tena, A. Mehri, F.X. Roca, S. Ozawa, J. Gonzàlez, Frequency-based enhancement network for efficient super-resolution, IEEE Access 10 (2022) 57383–57397.
[7] R. Keys, Cubic convolution interpolation for digital image processing, IEEE Trans. Acoust. 29 (6) (1981) 1153–1160.
[8] Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al., An image is worth 16x16 words:transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020).
[9] S. Paul, P.-Y. Chen, Vision transformers are robust learners. Proceedings of the AAAI Conference on Artificial Intelligence volume 36, 2022, pp. 2071–2081.
[10] H. Fu, B. Wang, J. Shen, S. Cui, Y. Xu, J. Liu, L. Shao, Evaluation of retinal image quality assessment networks in different color-spaces. International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2019, pp. 48–56.
[11] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Trans. Image Process. 13 (4) (2004) 600–612.
[12] J. Staal, M.D. Abràmoff, M. Niemeijer, M.A. Viergever, B. Van Ginneken, Ridge-based vessel segmentation in color images of the retina, IEEE Trans. Med. Imag. 23 (2004) 501–509.
[13] A.D. Hoover, V. Kouznetsova, M. Goldbaum, Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response, IEEE Trans. Med. Imag. 19 (3) (2000) 203–210.
[14] M.M. Fraz, P. Remagnino, A. Hoppe, B. Uyyanonvara, A.R. Rudnicka, C.G. Owen, S.A. Barman, An ensemble classification-based approach applied to retinal blood vessel segmentation, IEEE Trans. Biomed. Eng. 59 (9) (2012) 2538–2548.
[15] Z. Deng, Y. Cai, L. Chen, Z. Gong, Q. Bao, X. Yao, D. Fang, W. Yang, S. Zhang, L. Ma, Rformer: transformer-based generative adversarial network for real fundus image restoration on a new clinical benchmark, IEEE J. Biomed. Health Inform. (2022).
[16] Q. You, C. Wan, J. Sun, J. Shen, H. Ye, Q. Yu, Fundus image enhancement method based on cycleGAN. 2019 41st annual international conference of the IEEE engineering in medicine and biology society (EMBC), IEEE, 2019, pp. 4500–4503.
[17] W. Li, G. Liu, Y. He, J. Wang, W. Kong, G. Shi, Quality improvement of adaptive optics retinal images using conditional adversarial networks, Biomed. Opt. Express 11 (2) (2020) 831–849.
[18] A.D. Pérez, O. Perdomo, H. Rios, F. Rodríguez, F.A. González, A conditional generative adversarial network-based method for eye fundus image quality enhancement. International Workshop on Ophthalmic Medical Image Analysis, Springer, 2020, pp. 185–194.
[19] B. Subramani, M. Veluchamy, Fuzzy gray level difference histogram equalization for medical image enhancement, J. Med. Syst. 44 (2020) 1–10.
[20] L. Cao, H. Li, Y. Zhang, Retinal image enhancement using low-pass filtering and α-rooting, Signal Process. 170 (2020) 107445.
[21] C. Wan, X. Zhou, Q. You, J. Sun, J. Shen, S. Zhu, Q. Jiang, W. Yang, Retinal image enhancement using cycle-constraint adversarial network, Front. Med. 8 (2021).
[22] H. Li, H. Liu, Y. Hu, R. Higashita, Y. Zhao, H. Qi, J. Liu, Restoration of cataract fundus images via unsupervised domain adaptation. 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), IEEE, 2021, pp. 516–520.
[23] S. Zhang, C.A.B. Webers, T.T. Berendschot, A double-pass fundus reflection model for efficient single retinal image enhancement, Signal Process. 192 (2022) 108400.
[24] A. Alimanov, M.B. Islam, Retinal image restoration and vessel segmentation using modified cycle-CBAM and CBAM-UNet. 2022 Innovations in Intelligent Systems and Applications Conference (ASYU), IEEE, 2022, pp. 1–6.
[25] S. Woo, J. Park, J.-Y. Lee, I.S. Kweon, Cbam: Convolutional block attention module. Proceedings of the European conference on computer vision (ECCV), 2018, pp. 3–19.
[26] R. Lan, L. Sun, Z. Liu, H. Lu, C. Pang, X. Luo, Madnet: a fast and lightweight network for single-image super resolution, IEEE Trans. Cybern. 51 (3) (2020) 1443–1453.
[27] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al., Photo-realistic single image super-resolution using a generative adversarial network. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4681–4690.
[28] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, C. Change Loy, Esrgan: enhanced super-resolution generative adversarial networks. Proceedings of the European Conference on Computer Vision (ECCV) Workshops, 2018, 0–0.
[29] T. Dai, J. Cai, Y. Zhang, S.-T. Xia, L. Zhang, Second-order attention network for single image super-resolution. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 11065–11074.
[30] J. Cai, H. Zeng, H. Yong, Z. Cao, L. Zhang, Toward real-world single image super-resolution: a new benchmark and a new model. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 3086–3095.
[31] X. He, Z. Mo, P. Wang, Y. Liu, M. Yang, J. Cheng, Ode-inspired network design for single image super-resolution. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 1732–1741.
[32] Z. Hui, X. Wang, X. Gao, Fast and accurate single image super-resolution via information distillation network. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 723–731.
[33] S.-J. Park, H. Son, S. Cho, K.-S. Hong, S. Lee, Srfeat: single image super-resolution with feature discrimination. Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 439–455.

[34] Y. Wang, F. Perazzi, B. McWilliams, A. Sorkine-Hornung, O. Sorkine-Hornung, C. Schroers, A fully progressive approach to single-image super-resolution. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018, pp. 864–873.
[35] Y. Hu, J. Li, Y. Huang, X. Gao, Channel-wise and spatial feature modulation network for single image super-resolution, IEEE Trans. Circuits Syst. Video Technol. 30 (11) (2019) 3911–3927.
[36] J. Liu, Y. Liu, H. Wu, J. Wang, X. Li, C. Zhang, Single image super-resolution using feature adaptive learning and global structure sparsity, Signal Process. 188 (2021) 108184.
[37] Y. Hu, X. Gao, J. Li, Y. Huang, H. Wang, Single image super-resolution with multi-scale information cross-fusion network, Signal Process. 179 (2021) 107831.
[38] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial networks, Commun. ACM 63 (11) (2020) 139–144.
[39] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556 (2014).
[40] C. Dong, C.C. Loy, K. He, X. Tang, Image super-resolution using deep convolutional networks, IEEE Trans. Pattern Anal. Mach. Intell. 38 (2) (2015) 295–307.
[41] J. Daihong, Z. Sai, D. Lei, D. Yueming, Multi-scale generative adversarial network for image super-resolution, Soft Comput. 26 (8) (2022) 3631–3641.
[42] F. Yang, H. Yang, J. Fu, H. Lu, B. Guo, Learning texture transformer network for image super-resolution. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 5791–5800.
[43] S.W. Zamir, A. Arora, S. Khan, M. Hayat, F.S. Khan, M.-H. Yang, Restormer: efficient transformer for high-resolution image restoration. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 5728–5739.
[44] Z. Lu, J. Li, H. Liu, C. Huang, L. Zhang, T. Zeng, Transformer for single image super-resolution. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 457–466.
[45] G. Gao, Z. Xu, J. Li, J. Yang, T. Zeng, G.-J. Qi, Ctcnet: a cnn-transformer cooperation network for face image super-resolution, IEEE Trans. Image Process. 32 (2023) 1978–1991.
[46] G. Gao, Z. Wang, J. Li, W. Li, Y. Yu, T. Zeng, Lightweight bimodal network for single-image super-resolution via symmetric cnn and recursive transformer, arXiv preprint arXiv:2204.13286 (2022).
[47] S. Zhang, G. Liang, S. Pan, L. Zheng, A fast medical image super resolution method based on deep learning network, IEEE Access 7 (2018) 12319–12327.
[48] J. Zhu, G. Yang, P. Lio, How can we make GAN perform better in single medical image super-resolution? a lesion focused multi-scale approach. 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), IEEE, 2019, pp. 1669–1673.
[49] Y. Chen, F. Shi, A.G. Christodoulou, Y. Xie, Z. Zhou, D. Li, Efficient and accurate MRI super-resolution using a generative adversarial network and 3d multi-level densely connected network. International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2018, pp. 91–99.
[50] D. Mahapatra, B. Bozorgtabar, R. Garnavi, Image super-resolution using progressive generative adversarial networks for medical image analysis, Computerized Med. Imag. Graphic. 71 (2019) 30–39.
[51] D. Qiu, L. Zheng, J. Zhu, D. Huang, Multiple improved residual networks for medical image super-resolution, Future Generat. Comput. Syst. 116 (2021) 200–208.
[52] D. Qiu, Y. Cheng, X. Wang, Improved generative adversarial network for retinal image super-resolution, Comput. Methods Programs Biomed. (2022) 106995.
[53] Y. Lv, H. Ma, J. Li, S. Liu, Fusing dense and rezero residual networks for super-resolution of retinal images, Pattern Recognit. Lett. 149 (2021) 120–129.
[54] D. Mahapatra, B. Bozorgtabar, S. Hewavitharanage, R. Garnavi, Image super resolution using generative adversarial networks and local saliency maps for retinal image analysis. International Conference on Medical Image Computing and Computer-assisted Intervention, Springer, 2017, pp. 382–390.
[55] C. Tian, J. Yang, P. Li, S. Zhang, S. Mi, Retinal fundus image superresolution generated by optical coherence tomography based on a realistic mixed attention GAN, Med. Phys. 49 (5) (2022) 3185–3198.
[56] H. Hao, C. Xu, D. Zhang, Q. Yan, J. Zhang, Y. Liu, Y. Zhao, Sparse-based domain adaptation network for OCTA image super-resolution reconstruction, IEEE J. Biomed. Health Inform. 26 (9) (2022) 4402–4413.
[57] Lee, S.H.; Lee, S.; Song, B.C., Vision transformer for small-size datasets. arXiv preprint arXiv:2112.13492 (2021).
[58] Ulyanov, D.; Vedaldi, A.; Lempitsky, V., Instance normalization: the missing ingredient for faststylization. arXiv preprint arXiv:1607.08022 (2016).
[59] Hendrycks, D.; Gimpel, K., Gaussian error linear units (gelus). arXiv preprint arXiv:1606.08415 (2016).
[60] O. Ronneberger, P. Fischer, T. Brox, U-net: convolutional networks for biomedical image segmentation. International Conference on Medical Image Computing and Computer-assisted Intervention, Springer, 2015, pp. 234–241.