symmetry

MDPI

Article

SwinDPSR: Dual-Path Face Super-Resolution Network Integrating Swin Transformer

Xing Liu [1], Yan Li [1], Miao Gu [1], Hailong Zhang [1,*], Xiaoguang Zhang [1], Junzhu Wang [2], Xindong Lv [2] and Hongxia Deng [2,*]

[1] China FAW Group Corporation, Changchun 130000, China; liuxing13@faw.com.cn (X.L.); liyan41@faw.com.cn (Y.L.); gumiao@faw.com.cn (M.G.); zhangxiaoguang1@faw.com.cn (X.Z.)
[2] College of Computer Science and Technology, Taiyuan University of Technology, Jinzhong 030600, China; wangjunzhu0439@link.tyut.edu.cn (J.W.); lvxindong0593@link.tyut.edu.cn (X.L.)
* Correspondence: zhanghailong3@faw.com.cn (H.Z.); denghongxia@tyut.edu.cn (H.D.)

Abstract: Whether to use face priors in the face super-resolution (FSR) methods is a symmetry problem.Various face priors are used to describe the overall and local face features, making the generation of super-resolution face images expensive and laborious. FSR methods that do not require any prior information tend to focus too much on the local features of the face, ignoring the modeling of global information. To solve this problem, we propose a dual-path facial image super-resolution network (SwinDPSR) fused with Swin Transformer. The network does not require additional face priors, and it learns global face shape and local face components through two independent branches. In addition, the channel attention ECA module is used to aggregate the global and local face information in the above dual-path sub-networks, which can generate corresponding high-quality face images. The results of face super-resolution reconstruction experiments on public face datasets and a real-scene face dataset show that SwinDPSR is superior to previous advanced methods both in terms of visual effects and objective indicators. The reconstruction results are evaluated with four evaluation metrics: peak signal-to-noise ratio (PSNR), structural similarity (SSIM), learned perceptual image patch similarity (LPIPS), and mean perceptual score (MPS).

Keywords: super-resolution; Transformer; attention

Citation: Liu, X.; Li, Y.; Gu, M.; Zhang, H.; Zhang, X.; Wang, J.; Lv, X.; Deng, H. SwinDPSR: Dual-Path Face Super-Resolution Network Integrating Swin Transformer. *Symmetry* **2024**, *16*, 511. https://doi.org/10.3390/sym16050511

Academic Editor: Sergei D. Odintsov

Received: 19 March 2024
Revised: 10 April 2024
Accepted: 12 April 2024
Published: 23 April 2024

1. Introduction

Face super-resolution (FSR) represents a specialized endeavor within the domain of image super-resolution (SR), specifically targeting the enhancement of facial imagery. It pertains to the process of reconstructing a high-resolution (HR) representation of a face from a provided low-resolution (LR) counterpart. This approach significantly enhances the visual fidelity of facial images by reinstating intricate details often lost in lower-quality renditions. Given the inherent symmetry commonly found in facial structures, FSR methodologies heavily leverage global contextual cues to refine the reconstruction process, thereby yielding more faithful representations of the original face.

Various deep learning-based FSR methods have emerged, categorized into three groups: general, prior-information-guided, and attribute-constrained approaches. General methods focus on streamlined neural networks for face super-resolution without specific facial features. For instance, BCCNN [1] by Zhou et al. introduced a CNN for LR to HR face image mapping. CDFH [2] by Liu et al. is a cascaded model that first denoises and restores low-frequency information, then compensates for high-frequency details. SPARNet [3] by Chen et al. incorporates spatial attention in the generator and employs a multi-scale discriminator for enhanced image quality. PCRCN [4] by Liu et al. uses progressive upsampling for gradual high-resolution image acquisition. DBTC [5] by Shi et al. combines Transformer and CNN for improved detail recovery. SCGAN [6] introduces semi-cycled generative adversarial networks to address real-world face super-resolution challenges.

While most general methods employ CNNs, limited by their local receptive fields, they struggle to model global information, necessitating improvements in face reconstruction naturalness and fidelity.

For prior-information-guided FSR methods, FSRNet [7] by Chen et al. utilizes facial image characteristics to construct a prior-knowledge-extraction network, extracting geometric prior information (facial parsing map) to enhance super-resolution effects. MS-FSR [8] introduces a novel face prior (face boundary) for progressive LR image processing, employing cascaded sub-networks for upsampling. JASRNet [9] by Yin et al. leverages prior estimates' correlation for face super-resolution, while FSRGFCH [10] integrates the prior estimation branch (PEB) directly into the super-resolution branch (SRB), splitting the PEB into distinct segments. EIPNet [11] by Kim et al. incorporates a lightweight edge block and identity information to mitigate distortion, preserving identity integrity via an identity loss function. Notably, the reconstruction performance of prior-information-guided FSR methods hinges on the quality of prior information, which can be resource-intensive to generate.

In addition to network architecture design and prior information utilization, face super-resolution incorporates face attribute data, known as the attribute-constrained FSR approach. Face attributes, as semantic data, provide valuable insights, such as whether an individual wears glasses, enhancing the super-resolution process. AGCycleGAN [12] replicates attribute vectors to match LR image dimensions, generating attribute maps concatenated with LR images for super-resolution. However, Lee et al. suggest potential disparities between LR image data and attributes, prompting the development of AACNN [13], featuring a feature extraction network, super-resolution model, and discriminator. In contrast, ATENet [14] and its enhanced version, ATSENet [15], employ an attribute transfer network to upsample LR features and fuse them with attributes, generating an upsampled LR image with consistent attributes. Reconstruction efficacy in attribute-constrained FSR methods directly hinges on face attribute accuracy, shaping reconstruction outcomes.

Making full use of the information that face is symmetrical, the global information of face image is introduced for image reconstruction. In this paper, we propose a new dual-path face super-resolution network fused with Swin Transformer, called SwinDPSR, to improve the naturalness and realism of face reconstruction results. The network learns global face shape and local face components through two independent branches. Specifically, we first construct an encoder to extract high-dimensional features of LR images, and take the high-dimensional features as the input of the local representation path (LRP) and global representation path (GRP). In the LRP, we use the spatial attention mechanism to construct the facial attention unit (FAU), which focuses on the local information of the face. In our approach, the GRP utilizes the self-attention mechanism from Swin Transformer to construct the Residual Swin Transformer Block (RSTB), capturing global face features without relying on CNN architectures like in [16]. A fusion and reconstruction module then merges features from the LRP and GRP, feeding the fused vector into the decoder for high-resolution face image generation. Training occurs through end-to-end supervision, combining multiple loss functions in a weighted sum. Experimental results, as depicted in Figure 1, demonstrate superior global information modeling compared to existing methods, thereby enhancing the quality of super-resolution face images. In summary, our main contributions are as follows:

- We propose a dual-path face super-resolution network fused with Swin Transformer, called SwinDPSR, to perform face super-resolution reconstruction by fusing local detail features and global face features. The proposed global representation path utilizes Transformer's self-attention mechanism to recover face global information. This is followed by feature fusion with a local representation path composed of facial attention units, thereby improving the representation ability and SR performance of the network.

- We jointly train the network with pixel loss, style loss, and SSIM loss to promote network convergence from the pixel level, perception level, and image structure level, respectively.
- In addition to traditional SR evaluation metrics like PSNR and SSIM, we incorporate learned perceptual image patch similarity (LPIPS), mean perceptual score (MPS), and identity similarity as network performance indicators. LPIPS is computed using AlexNet to measure the L2 distance between SR and HR image eigenvectors. Identity similarity, calculated using FaceNet, quantifies the cosine similarity between SR and HR image eigenvectors.

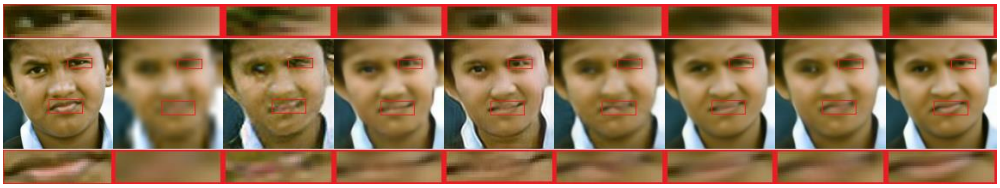

Figure 1. Visual results of different super-resolution methods using upscale factor of 8: (**a**) Ground truth; (**b**) Bicubic; (**c**) SRGAN; (**d**) FSRNet; (**e**) FSRGAN; (**f**) AACNN; (**g**) SPARNet; (**h**) EIPNet; (**i**) Ours.

2. Related Works

2.1. Attention Networks

The attention mechanism aims to scan the entire image, identifying key attention areas while suppressing irrelevant information, thereby enhancing the efficiency and accuracy of visual information processing. In recent years, this mechanism has found widespread application in high-level vision tasks like image segmentation and enhancement.

In the field of image segmentation, DANet [17], proposed by Fu et al., adaptively integrates local features and their global dependencies, and realizes scene segmentation task by capturing rich contextual correlations. Tao et al. [18] found that the predictions of network models at certain scales are better at resolving specific failure modes, resulting in better predictions. Therefore, they proposed an attention-based method to combine multi-scale information for segmentation, which improves the effect of semantic segmentation. Traditional attention mechanisms ignore the implicit semantic segmentation subtask and are constrained by the grid structure of convolution kernels. The SANet [19], proposed by Zhong et al., utilizes an efficient squeeze-and-attention (SA) module to account for the two salient features of segmented pixel-group attention and pixel-level prediction.

In the field of image enhancement, Zhang et al. [20] integrated channel attention into a deep residual network for super-resolution. Qin et al. proposed FFA-Net [21] to directly restore haze-free images, featuring channel and pixel attention for enhanced flexibility in processing varied information types. Tian et al.'s ADNet [22] employs sparse, feature enhancement, attention, and reconstruction blocks for image denoising. The attention block extracts hidden noise information from the background, while the feature enhancement and attention blocks collaboratively streamline noise model training and reduce complexity.

In the field of face super-resolution reconstruction, spatial-attention-guided convolutional layers can adaptively guide features related to key facial structures and pay less attention to those regions that are not rich in features. In order to better capture key face structures, for the local representation path (LRP), we utilize a Facial Attention Unit (FAU) to focus on face local information.

2.2. Vision Transformer

Transformer [23], originating from natural language processing (NLP), stacks multi-head self-attention and feed-forward MLP layers to capture long-range correlations among words. Inspired by its success, researchers have explored Transformer's advantages in various visual tasks, emphasizing global feature extraction. For instance, Dosovitskiy

et al. [24] introduced Vision Transformer, treating 16×16 image patches as a sequence and predicting image classes via a unique class token. Swin Transformer [25] combines CNN and Transformer strengths, utilizing local attention for large-scale image handling and a shifted-window scheme for long-term dependency modeling. SwinIR [26] employs residual Swin Transformer blocks (RSTBs) for deep feature extraction, achieving notable performance across diverse image denoising tasks. Swin-Unet [27] applies Swin Transformer in a UNet architecture for medical image segmentation, facilitating local-global semantic feature learning. DehazeFormer [28] enhances Swin Transformer-based single-image dehazing by improving normalization, activation functions, and spatial information aggregation, effectively removing uneven haze from real remote sensing datasets.

While pure Transformer networks excel at extracting global representations, they may struggle to capture local fine-grained details in images. To address this limitation, some approaches integrate CNNs into Vision Transformer, leveraging their unique local modeling and translation invariance capabilities. For instance, Carion et al. [29] employed a cascaded CNN and Transformer for end-to-end object detection. Similarly, Yang et al. [30] combined a CNN-based learnable texture extraction module with a Transformer-based embedding module for texture transfer and synthesis tasks, achieving visually objective results. In this study, we propose a hybrid face super-resolution reconstruction network that combines a spatial-attention-guided CNN with a self-attention-guided Transformer. This architecture effectively restores local face details while capturing the global face structure.

3. Proposed Method

In this section, we introduce SwinDPSR, a novel face super-resolution approach that integrates a dual-path architecture with Swin Transformer. The network is structured with an encoder–decoder architecture, which systematically extracts high-dimensional features of the face. It employs a local representation path, leveraging spatial attention, and a global representation path, utilizing Swin Transformer, to capture detailed facial features and global characteristics, respectively. Subsequently, we provide a comprehensive overview of SwinDPSR, including its overall structure, internal module architecture, and optimization strategy.

3.1. Overall Architecture

The overall architecture of SwinDPSR is shown in Figure 2. It consists of five parts: encoder, global representation path, local representation path, fusion and reconstruction module, and decoder. The input and output of the network are taken as I_{IR} and I_{SR}. First, a shallow feature extractor composed of 3 × 3 convolutional layers is used to extract a shallow feature $F_{shallow}$ containing rich structural information from the input image.

$$F_{shallow} = H_{Conv}^{3\times3}(I_{LR}) \tag{1}$$

where $H_{Conv}^{3\times3}$ is a convolutional layer with a convolution kernel size of 3 × 3. Then, use $F_{Shallow}$ as the input of the encoder to extract high-dimensional features $F_{Encoder}$ from the input image:

$$F_{Encoder} = H_{Encoder}(F_{Shallow}) \tag{2}$$

$H_{Encoder}$ serves as the downsampling encoder, while $F_{Encoder}$ is inputted into both the global representation path and local representation path to extract global structure and local details, respectively.

$$F_{global} = H_{global}(F_{Encoder}), F_{local} = H_{local}(F_{Encoder}) \tag{3}$$

where H_{global} and H_{local} are functions of the global representation path and local representation path, respectively. F_{global} and F_{local} denote the global features extracted by the global representation path and the local features extracted by local representation path, respectively. After obtaining the global and local features, feature fusion is performed using the

fusion and reconstruction module, which utilizes two ECA modules and a convolutional layer to fuse F_{global} and F_{local}. The fused feature is expressed as

$$F_{gl} = H_{ECA}(H_{Conv}^{3\times3}(H_{ECA}(H_{Cat}(F_{global}^{ft}, F_{local}^{ft})))) \tag{4}$$

where H_{Cat} represents the splicing function in the channel dimension. After obtaining F_{gl}, the decoder needs to perform an upsampling operation to obtain the feature $F_{Decoder}$.

$$F_{Decoder} = H_{Decoder}(F_{gl}) \tag{5}$$

where $H_{Decoder()}$ represents the upscale decoder. Finally, the enlarged features are reconstructed by convolutional layers, and the target high-resolution image I_{SR} is output.

$$I_{SR} = H_{Conv}^{3\times3}(F_{Decoder}) \tag{6}$$

where $H_{Conv}^{3\times3}$ is used to output RGB three-channel images.

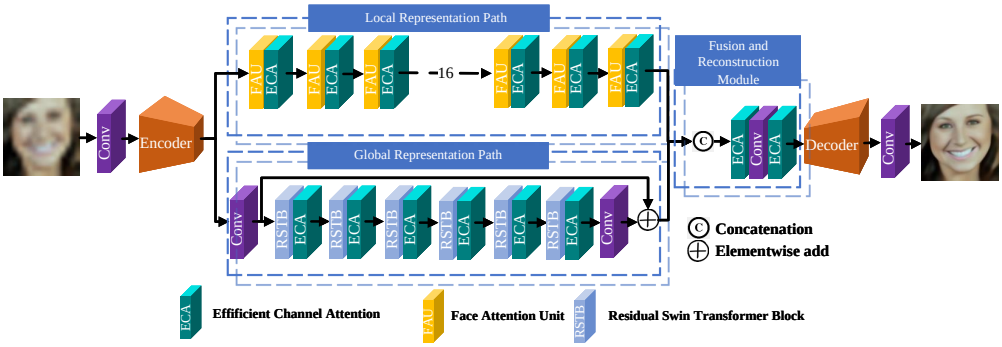

Figure 2. The overall architecture of SwinDPSR, which consists of five components: encoder, global representation path composed of RSTB, local representation path composed of Facial Attention Units (FAUs), fusion and reconstruction module (FRM), and decoder.

Algorithm 1 gives the algorithm process of SwinDPSR.

Algorithm 1 Training of SwinDPSR

Require: Set the batch size to 16, the amplification factor to 8, the epoch to 20, the network initialization method to Xavier, the learning rate to 4×10^{-4}, the learning rate decay strategy to linear decay, and the parameters β_1 in the Adam optimizer to 0.9 and β_2 to 0.99.

1: $iter \leftarrow 0$
2: **repeat**
3: $high\_dim\_features \leftarrow Encoder(low\_resolution\_image)$
4: $local\_features \leftarrow LPR(high\_dim\_features)$
5: $global\_features \leftarrow GRP(high\_dim\_features)$
6: Learns global face shape and local face components through two independent branches.
7: $fused\_features \leftarrow FusionAndReconstructionModule(local\_features, global\_features)$
8: $high\_resolution\_image \leftarrow Decoder(fused\_features)$
9: **Output:** Generate high_resolution_image
10: Calculate the loss update parameter
11: **until** end of Training

3.2. Details of SwinDPSR

3.2.1. Local Representation Path

Taking inspiration from SPARNet [3], our attention branch requires the extraction of multi-scale features. To achieve this, we introduce a Facial Attention Unit (FAU). Illustrated in Figure 3, the FAU utilizes an hourglass block and an additional Conv layer to generate attention maps. The hourglass block is renowned for its ability to capture information across multiple scales [31], demonstrating effectiveness in face analysis tasks such as face alignment [32] and face parsing [7]. Through the stacking of FAUs, critical features for facial SR images are continuously enhanced.

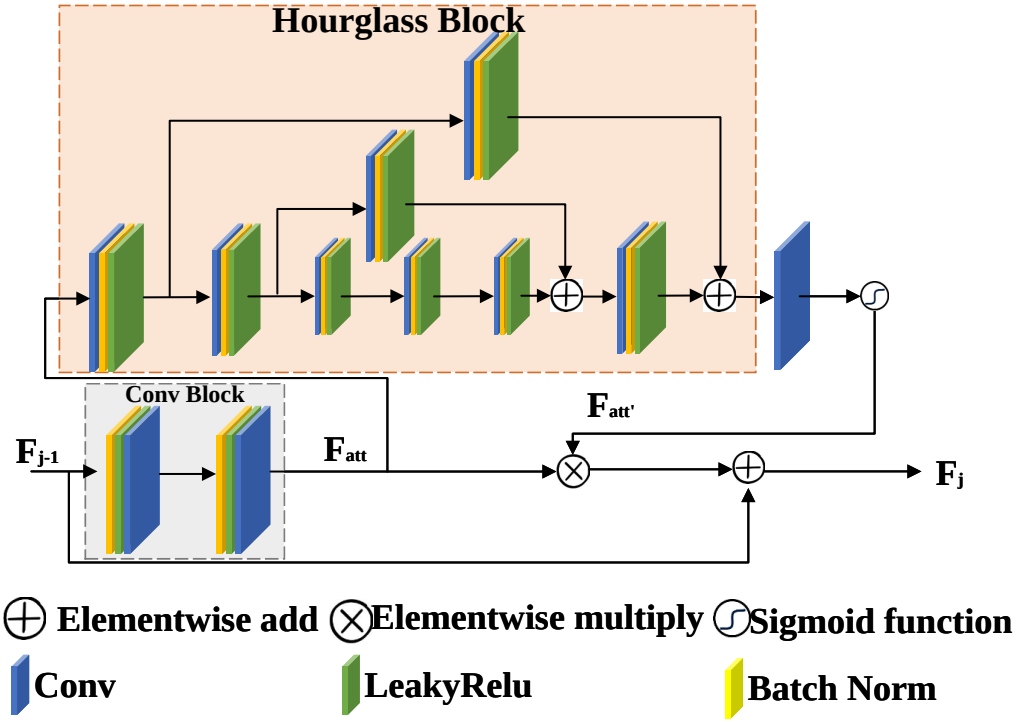

Figure 3. The architecture of the Face Attention Unit (FAU).

Taking the j-th FAU as an example, the input features and output features of the FAU are taken as $F_{j-1} \in R^{C_{j-1} \times H_{j-1} \times W_{j-1}}$ and $F_j \in R^{C_j \times H_j \times Wj}$. First, a convolutional block consisting of batch normalization layers, LeakyRelu activation layers, and $F_{att} \in R^{C_j \times H_j \times W_j}$ convolutional layers is used to extract the features containing higher-dimensional information from the input features.

$$F_{att} = H_{CB}(F_{j-1}) \tag{7}$$

where $H_{CB}()$ denotes a convolutional block consisting of batch normalization layers, LeakyRelu activation layers, and convolutional layers with a kernel size of 3 × 3. We then use F_{att} as the input of the hourglass block to extract face attention features $F_{att'} \in R^{1 \times H_j \times W_j}$ from the original features.

$$F_{att'} \in \mathbb{R}^{l \times H_j \times W_j} \tag{8}$$

where $H_{HB}()$ represents the hourglass block, $H_{Conv}^{3\times3}$ is a convolutional layer with a kernel size of 3 × 3, σ is the sigmoid function; then, use $F_{att'}$, F_{att}, and F_{j-1} to perform element-wise multiplication and element-wise addition to obtain the output feature F_j.

$$F_j = F_{j-1} + F_{att'} \otimes F_{att} \tag{9}$$

where $\otimes$ is element-wise multiplication. Similarly, we also use hourglass blocks to build the encoder and decoder (shown in Figure 4). The output feature F_j becomes

$$F_j = H_{scale}(F_{j-1}) + F_{att'} \otimes F_{att} \tag{10}$$

where $H_{scale}()$ represents the scale Conv layer. The downscale Conv in the encoder is a normal convolution layer with a step size of 2, and the upscale Conv in the decoder first performs nearest-neighbor upsampling, and then, performs the convolution operation, which helps to avoid checkerboard artifacts [33].

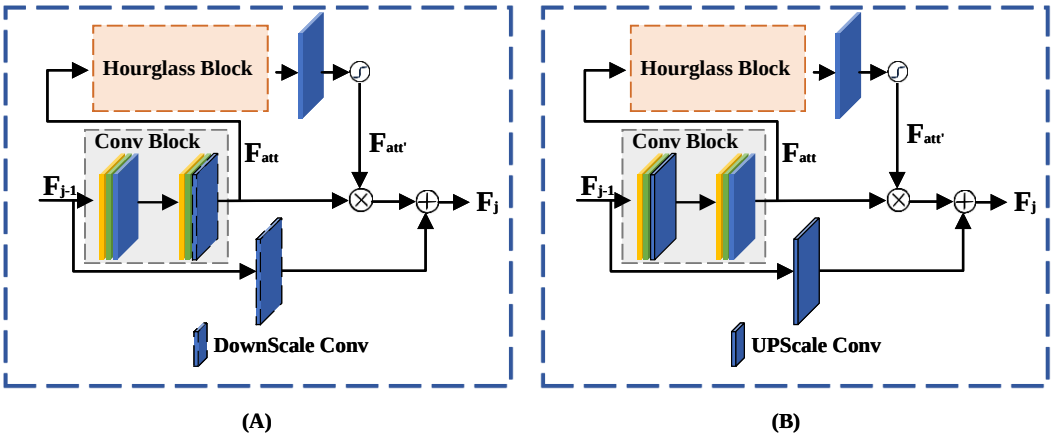

Figure 4. The architecture of encoder and decoder: (**A**) shows the detailed structure of encoder; (**B**) shows the detailed structure of decoder.

3.2.2. Global Representation Path

We use Swin Transformer to construct the global representation path, as shown in Figure 5A, RSTB is a residual block composed of Swin Transformer layers (STLs) and convolutional layers. Given the input feature $F_{i,0}$ of the i-th RSTB, then the intermediate features $F_{i,1}$, $F_{i,2}$, ... $F_{i,J}$, are expressed as

$$F_{i,j} = H_{STL_{i,j}}(F_{i,j-1}), j = 1, 2, \ldots J \tag{11}$$

where $H_{STL_{i,j}}()$ represents the j-th STL in the i-th RSTB, then the output feature vector $F_{i,out}$ of the i-th RSTB is expressed as

$$F_{i,out} = H^{3\times3}_{Conv}(F_{i,J}) + F_{i,0} \tag{12}$$

where $H^{3\times3}_{Conv}()$ represents a convolutional layer with a convolution kernel size of 3 × 3.

The Shifted Transformer Layer (STL) differs from the original Vision Transformer [24] by employing a shifted-window scheme, which enhances efficiency by confining self-attention computation to non-overlapping local windows while facilitating cross-window connections. This layered architecture offers scalability across different scales and maintains linear computational complexity relative to image size. The STL is shown in Figure 5B, and the propagation process of the input feature $X \in R^{H\times W\times C}$ is

$$\begin{aligned} X &= \text{MSA}(LN(X)) + X \\ X &= \text{MLP}(LN(X)) + X \end{aligned} \tag{13}$$

where $MSA()$ is the multi-head self-attention layer, and $MLP()$ is the multi-layer perceptron layer. A LayerNorm (LN) layer is added before the MSA and MLP, and both modules use residual connections.

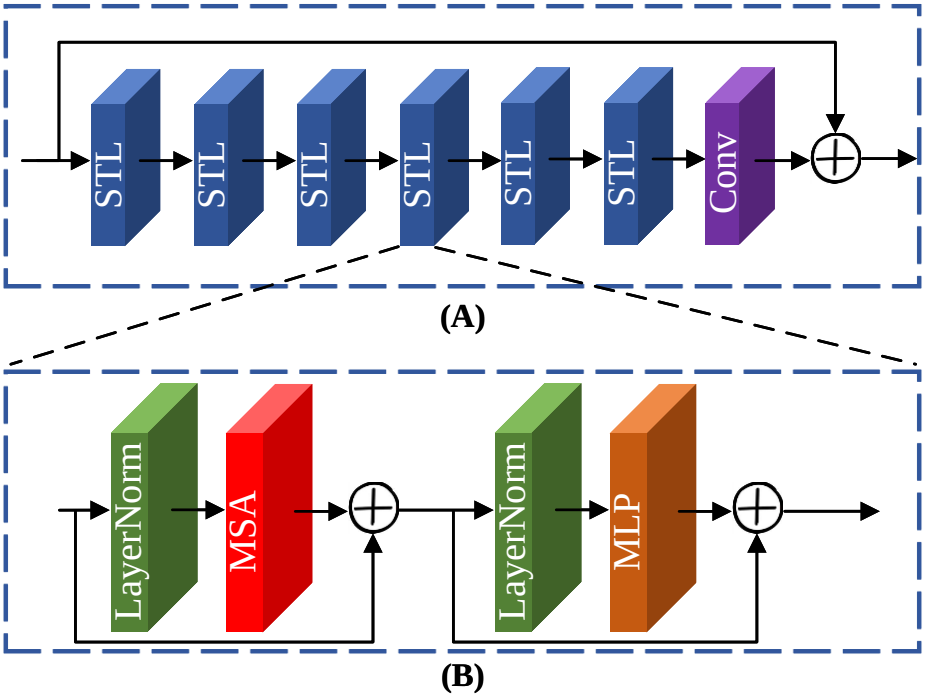

Figure 5. The architecture of RSTB. (**A**) Shows the overall structure of an RSTB, which contains six Swin Transformer layers (STLs) and one convolution layer. (**B**) Shows the detailed structure of an STL.

The multi-head self-attention layer first divides the input into $N \times N$ non-overlapping local windows, and reshapes the input feature size to $\frac{HW}{N^2} \times N^2 \times C$, where $\frac{HW}{N^2}$ is the total number of windows. Then, local self-attention is computed for each window separately. For feature $X \in R^{\frac{HW}{N^2} \times N^2 \times C}$, the query, key and value matrices Q, K, and V are computed as

$$Q = X\alpha_Q, K = X\alpha_K, V = X\alpha_V \tag{14}$$

where α_Q, α_K, and α_V represent the weight parameter matrix that needs to be trained and updated. $Q, K, V \in R^{\frac{HW}{N^2} \times M \times N^2 \times \frac{C}{M}}$, where M is the number of self-attention heads in the multi-head self-attention layer. Classical Transformer [21,22] uses either deterministic positional encoding or learnable positional encoding. Compared with absolute positional encoding, relative positional encoding [34] is able to learn stronger "relationships" between local content, bringing important performance improvements in the case of large-scale dataset training, and has been widely used [25,35]. We therefore add relative position encoding to the Transformer, and calculate the attention matrix through the self-attention mechanism in the local window. The attention matrix looks like this:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right) + E)V \tag{15}$$

where K^T is the transpose of the third and fourth dimensions of the K matrix, $d = \frac{C}{M}$, and E is a learnable relative positional encoding, which is added to the attention map as a bias item. To achieve the interaction between windows, we alternately use regular-window partitioning and shifted-window partitioning to achieve cross-window connections, where shifted-window partitioning shifts features by $(\frac{N}{2}, \frac{N}{2})$ pixels before partitioning. Next, a multi-layer perceptron is used for further feature transformation, which consists of two fully connected layers and the GELU nonlinear activation function. To sum up, the global representation path uses the global receptive field of Swin Transformer to model the global context. Such a design can improve the issue of the network only focusing on local

information, and make the network pay attention to the global information of the face at the same time.

3.2.3. ECA Module

In recent years, channel attention mechanisms have demonstrated significant potential in enhancing deep convolutional neural network performance. SENet, introduced by Hu et al. [36], utilizes fully connected layers to predict channel attention weights and alleviate attention to redundant channels. However, Wang et al. [37] noted that SENet's dimensionality reduction may introduce side effects to the channel attention mechanism, increasing network complexity by capturing dependencies among all channels. To strike a balance between performance and complexity, they proposed an Efficient Channel Attention (ECA) module. This module, with few parameters, achieves substantial performance gains by employing one-dimensional convolution for local cross-channel interaction without dimensionality reduction. Moreover, they devised a method to adaptively select the size of the 1D convolution kernel, determining the extent of local cross-channel interactions.

This paper uses the ECA module to focus on the difference in channel importance between the spatial attention module and the self-attention module, and also reduces the attention to redundant channels during feature fusion. The structure of the ECA module is shown in Figure 6, and GAP in the figure represents the global average pool layer. Assuming that the input feature of the ECA module is F_{input}, then the output feature F_{output} of the ECA module is expressed as

$$F_{output} = \phi(H_{Conv}(H_{GAP}(F_{input}))) \otimes F_{input} \tag{16}$$

In the formula, $\phi()$ denotes the sigmoid function. After F_{input} undergoes global average pooling via $H_{GAP()}$, it undergoes one-dimensional convolution $H_{Conv()}$ to establish inter-channel connections among neighboring local channels. The range of local cross-channel interaction is determined by the convolution kernel size of the one-dimensional convolution, which correlates positively with the channel dimension of the input feature F_{input} of the ECA module. Channel attention weights are derived by passing the output feature of the one-dimensional convolution through the sigmoid function; subsequently, the input features of the ECA module are element-wise multiplied by the channel attention weights to obtain the output features F_{output}.

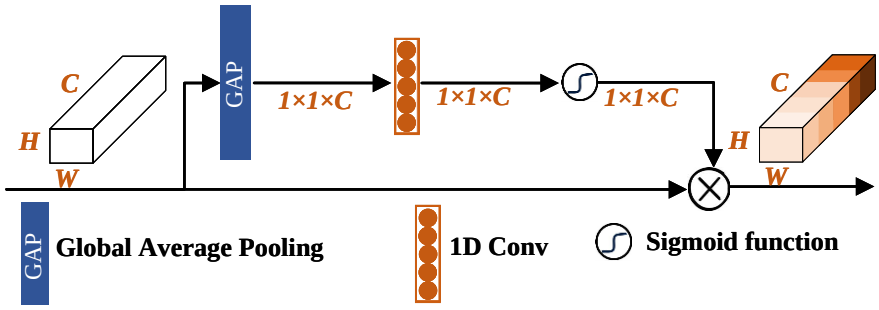

Figure 6. ECA module structure diagram.

3.3. Training and Loss Function

We train the network jointly with multiple loss functions, and the joint loss function is defined as

$$l = \alpha l_{pixel} + \beta l_{ssim} + \gamma l_{style} \tag{17}$$

In our experiments, we fixed $\alpha = 100$, $\beta = 10$, and $\gamma = 1$. In image conversion tasks, pixel loss serves as a measurement method relying on the disparity between the output and real images. This metric computes the average absolute error between corresponding

pixel pairs across both images, aiming to minimize discrepancies for greater similarity. It is expressed as:

$$l_{pixel}(I_{HR}, I_{SR}) = \frac{1}{N_M}\|I_{HR} - I_{SR}\| \tag{18}$$

Pixel loss, utilizing L1 loss (mean absolute error), constrains the SR image to closely match the HR image in pixel values. In practice, L1 loss demonstrates superior performance and convergence compared to L2 loss. Given that PSNR definition correlates closely with pixel-level differences, pixel loss directly maximizes PSNR, making it the most commonly used loss function. However, pixel loss neglects image perception quality and texture details, often resulting in perceptually unsatisfactory outcomes with diminished high-frequency details. Similar to pixel loss, SSIM loss aims to enhance the structural similarity of super-resolution images, operating on the principle that:

$$l_ssim(I_{HR}, I_{SR}) = \frac{1}{2}(1 - SSIM(I_{HR} - I_{SR})) \tag{19}$$

In face super-resolution, style loss is commonly employed to enhance facial details and visual quality, as seen in ASFFNet [38]. Both SR and HR images are passed through a pre-trained network (e.g., VGGFace [39]) to obtain their respective features F_{SR} and F_{HR}. Style loss, initially proposed in [40] for image style transfer, operates similarly to perceptual loss as both are feature-level loss functions. Subsequently, their Gram matrices are computed, and these matrices are utilized to calculate the loss, defined as:

$$l_{style}(I_{HR}, I_{SR}) = \|G(F_{HR}) - G(I_{SR})\|_2 \tag{20}$$

where G() represents the operation of obtaining the feature Gram matrix. We use the above three losses for joint training to accelerate the convergence of the network from multiple perspectives, and then, improve the network performance to a certain extent.

4. Experiments

4.1. Datasets

Training set: We use the CelebA dataset [41] to train SwinDPSR. We first use MTCNN [42] to detect faces in the original dataset (178 × 218), crop and align them, and then resize them to 128 × 128 with bicubic interpolation as the HR training set. The LR (16 × 16) training set is obtained by downsampling the corresponding HR images. This yields approximately 202k image pairs. To avoid overfitting, we perform data augmentation by random horizontal flipping and image scaling (between 1.0 and 1.3). Test set: We randomly select 200 and 100 images from the Helen dataset [43] and FFHQ dataset [44], respectively, as the test set, and evaluate the image quality of the test set. For identity similarity evaluation, we select low-resolution images of 130 people from the SCface surveillance scene face dataset [45] as the test set to verify the effectiveness of the algorithm in real scenes.

4.2. Implementation Details

The parameter settings of the SwinDPSR training are shown in Table 1. All codes are written and tested in PyTorch [46] and Python.

Table 1. Hyperparameter settings for SwinDPSR.

Hyperparameter	Value
Batch Size	16
Amplification Factor	8
Epoch	20

Table 1. *Cont.*

Hyperparameter	Value
Network Initialization	Xavier
Learning Rate	4×10^{-4}
Learning Rate Decay Strategy	Linear Decay
β_1 (Adam Optimizer)	0.9
β_2 (Adam Optimizer)	0.99
GPU	Tesla V100
Environment	PyTorch

4.3. Evaluation Metrics

The reconstruction results are assessed using four evaluation metrics: peak signal-to-noise ratio (PSNR), structural similarity (SSIM) [47], learned perceptual image patch similarity (LPIPS) [48], and mean perceptual score (MPS) [49]. PSNR and SSIM are conventional metrics extensively employed in vision tasks like image enhancement. LPIPS and MPS represent novel perceptual metrics for gauging image perceptual quality, with smaller LPIPS values indicating higher perceptual similarity. MPS is the average of SSIM and LPIPS. Its formula is as follows:

$$MPS = 0.5 \times (SSIM + (1 - LPIPS)) \tag{21}$$

Identity similarity measures how well identity information is preserved in super-resolution faces. We first use the pre-trained FaceNet model [50] to extract identity feature vectors for SR image and HR image faces, and then, calculate the cosine similarity between the two feature vectors as the identity similarity.

4.4. Ablation Experiments and Discussion

In the ablation experiments, we trained SwinDPSR on the CelebA dataset [41] and tested it on the Helen dataset [43]. During the test, a 16 $\times$ 16 low-resolution image was used as input, and face image super-resolution reconstruction with an enlargement factor of 8 was performed (SR $\times$ 8). Through the following experiments, it was finally determined to set the number of FAUs to 16, the number of RSTBs and STLs was set to 6, and the embedded channels of Transformer were set to 120, and the effectiveness of the two paths and the joint loss was verified.

In order to verify the influence of FAUs in the local representation path on the network performance, the experiment was carried out under the condition of removing the global representation path, and the results are shown in Figure 7. The results show that as the number of FAUs increases, although PSNR and SSIM gradually increase, the performance gain gradually saturates and reaches a peak when the number of FAUs is 16. Therefore, we chose 16 as the number of FAUs in the remaining experiments.

Figure 8A, Figure 8B, and Figure 8C, respectively, show the effects of the number of RSTBs, the number of STLs and the number of embedded channels in the RSTBs on the model performance under the premise that the number of FAUs is set to 16. The results show that the PSNR and SSIM gradually increase with the increase in the number of RSTBs and the number of STLs. The peak value is reached when the number of RSTBs is six. Likewise, the PSNR and SSIM reach their peak when the number of embedded channels is 120. Therefore, in the remaining experiments, we choose six as the number of RSTBs, 120 as the number of embedded channels, and in order to balance performance and model size, we choose six as the number of STLs.

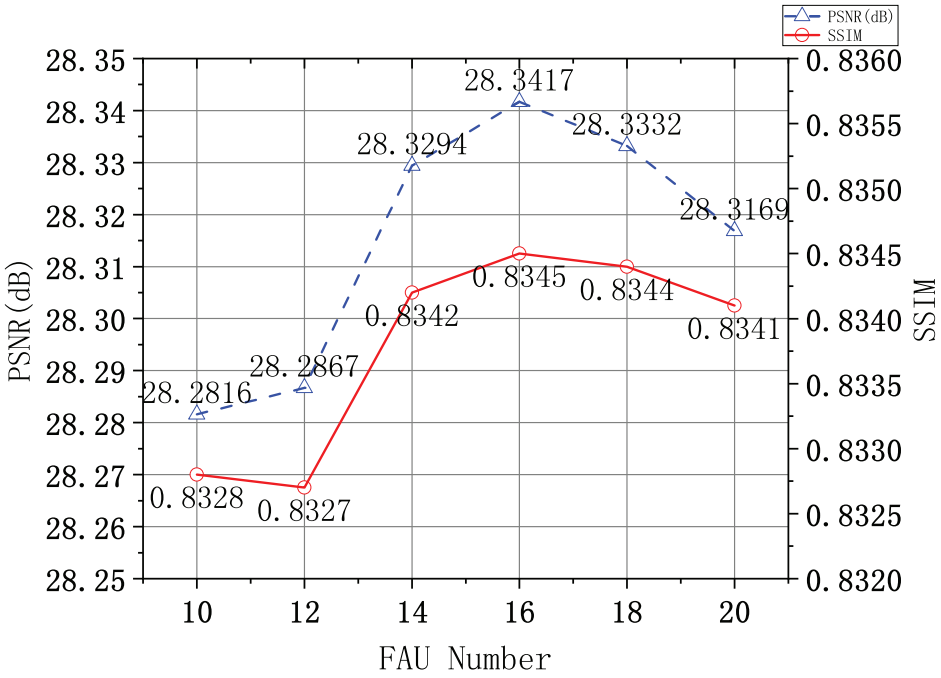

Figure 7. Ablation experiments with different numbers of FAUs.

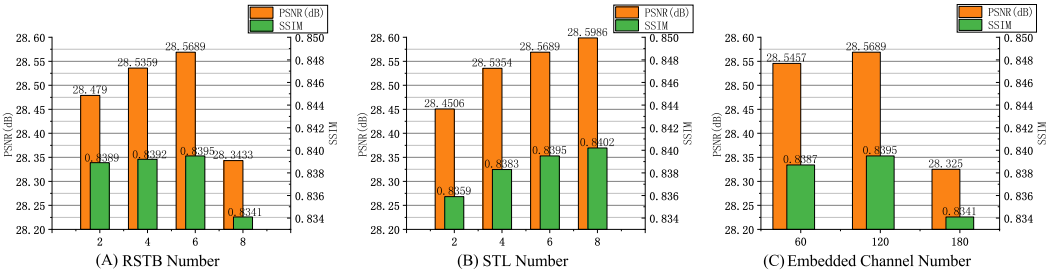

Figure 8. Ablation experiments with different numbers of RSTBs, STLs, and embedded channels.

We conducted three experiments to assess the impact of the two paths in SwinDPSR. The first experiment involved evaluating network performance after removing the global representation path, while the second experiment evaluated performance after removing the local representation path. The third experiment assessed the original SwinDPSR's performance. Results from these experiments are summarized in Table 1. Our findings indicate two key points: firstly, removal of the local representation path significantly degrades reconstruction performance, suggesting its importance in capturing facial local information; secondly, the Swin Transformer-based global representation path notably enhances network performance by capturing global facial structure information. In these experiments, the local representation path contained 16 FAUs, while the global representation path comprised six RSTBs and STLs each, with RSTBs featuring 120 embedded channels.

To demonstrate the impact of combining different loss functions on model performance, we provide Table 2 to show the gradual improvement in each loss function on the SR effect. It can be observed from Table 2 that although joint style loss training does not help much in improving the three indicators of PSNR, SSIM, and MPS, it can improve the LPIPS indicator to a certain extent. This is because style loss uses the Gram matrix instead of the covariance matrix, making the feature statistics of the generated image similar to the real image. Training with SSIM loss can improve the SSIM indicator to a certain extent, because SSIM loss always pays attention to the structural similarity difference between images.

Table 2. Verifying the effect of the local and global representation paths. "SwinDPSR w/o local" indicates that the local representation path is removed. "SwinDPSR w/o global" indicates that the global representation path is removed.

	PSNR	SSIM	LPIPS	MPS
SwinDPSR w/o global	28.3417	0.8345	0.2020	0.8162
SwinDPSR w/o local	26.0894	0.7636	0.3018	0.7308
SwinDPSR	28.5689	0.8395	0.1855	0.8270

We investigated the impact of different channel attention modules on reconstruction performance by integrating SE and ECA modules into the base network for feature fusion. As shown in Table 3, while the SE module marginally enhances PSNR and SSIM values, the ECA module outperforms it in enhancing network performance. This superiority stems from the ECA module's more effective extraction of channel features through one-dimensional convolution, reducing the influence of redundant features on network performance. Through these experiments, we conclude that SwinDPSR, when appropriately stacked with spatial attention units and residual Transformer blocks, can effectively enhance the reconstruction of structured images. The network primarily utilizes the spatial attention module, with the self-attention module serving as an auxiliary to establish LR-to-HR mapping. Furthermore, multi-loss joint training and the inclusion of an ECA channel attention module further elevate super-resolution reconstruction performance.

Table 3. Ablation experiments for training with different loss functions.

	PSNR	SSIM	LPIPS	MPS
Lpix	28.5689	0.8395	0.1855	0.8270
Lpix_Lstyle	28.5984	0.8396	0.1817	0.8289
Lpix_Lstyle_Lssim	28.6326	0.8415	0.1828	0.8293

4.5. Comparison with State-of-the-Art Methods

To further verify its practicality, we compared our proposed method with the current state-of-the-art methods, including SRGAN [51] and FSRGAN [7], based on generative adversarial networks; SPARNet [3], based on general methods; FSRNet [7] and EIPNet [11], based on prior information constraints; and AACNN [13], based on attribute information constraints. These methods are conditionally similar to our experiments, so their SR results are compared quantitatively and qualitatively. In addition, we also performed identity similarity comparisons. We verify the effectiveness of the method proposed in this paper on face super-resolution reconstruction through the following experiments.

4.5.1. Quantitative and Qualitative Comparison

As can be seen from Table 4, on the Helen test data set of 8 × SR, SwinDPSR is obviously superior to these latest technologies in three indexes, but lower than the FSRGAN method in LPIPS and MPS evaluation indices, because the FSRGAN method pays too much attention to the perceived distance between images. Thus, the mapping at the pixel level is ignored. Among these quantitative results, we can find that AACNN and FSRNet have not achieved satisfactory results. The main reason for this result is that the architecture based on attribute constraints and prior information constraints has high requirements on the accuracy of prior information, and inaccurate prior information will seriously affect the reconstruction effect. Although FSRGAN performs best in the LPIPS index, from the other three indicators, FSRGAN only focuses on the perceived distance between LR image and HR image, which can be proved by qualitative experiments, as shown in Figure 9.

Table 4. Effectiveness of SE module and ECA module on PSNR and SSIM indicators.

	PSNR	SSIM	LPIPS	MPS
BaseLine	28.6326	0.8415	0.1828	0.8293
SEModule	28.6517	0.8429	0.1819	0.8304
ECAModule	28.7688	0.8449	0.1799	0.8325

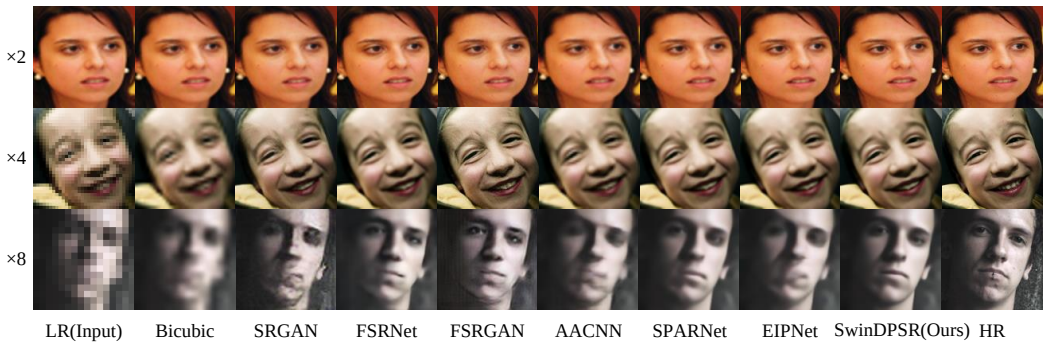

Figure 9. Qualitative results with state-of-the-art methods on Helen dataset. The resolution of the input is 16 × 16 and the upscale factor is 8.

Figure 9 shows the qualitative results of different methods for 8 × SR on the Helen test dataset. We can see that SRGAN does not take into account the special structure of the face, so the reconstruction effect is poor. The results reconstructed by AACNN, FSRNet, and EIPNet are indeed blurrier than the methods with better quantitative results. Compared with the SwinDPSR reconstruction results, FSRNet, AACNN, SPARNet, and EIPNet produce different degrees of distortion in the reconstruction of the glasses area in the first image and the earring area in the second image. Although the faces in the third and fourth images are not aligned, making face restoration difficult, SwinDPSR can reconstruct the eye area and mouth area better than other methods. Although the reconstruction perception effect of FSRGAN is good, the prior knowledge with low accuracy leads to serious loss of face identity information. The excellent reconstruction results of SwinDPSR strongly prove the advantage of Transformer in capturing the global facial structure. Compared with existing methods, SwinDPSR is able to maintain the consistency of the facial structure.

From the quantitative experiments on the FFHQ dataset in Table 5, it can be seen that SwinDPSR achieves overwhelming success in all evaluation indicators except the LPIPS indicator, which proves that the proposed method has a high generalization ability on different datasets. We can also prove this from the qualitative experiments in Figure 10.

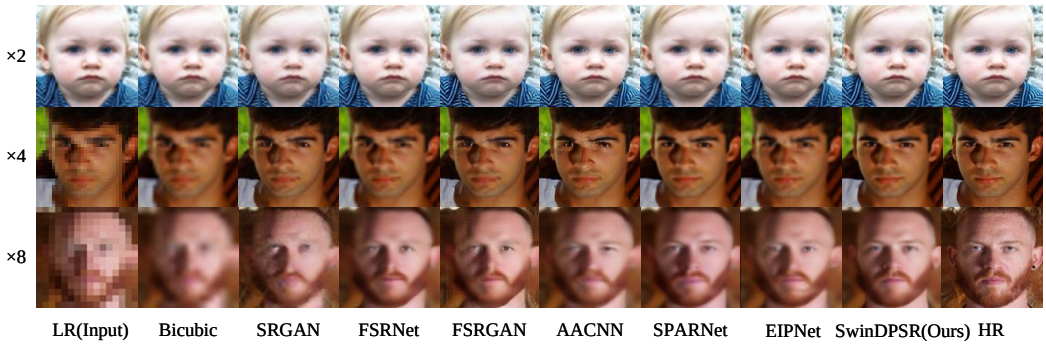

Figure 10. Qualitative comparison with state-of-the-art methods on FFHQ dataset. The resolution of the input is 16 × 16 and the upscale factor is 8.

4.5.2. Face Reconstruction and Recognition on Real-World Surveillance Scenarios

The ultimate goal of face super-resolution should be to serve reality, because the face images captured by surveillance cameras in real scenes often contain a lot of noise and serious texture distortion. Therefore, a real-world surveillance scenario is a challenging environment for face super-resolution. In order to verify the reconstruction performance of SwinDPSR in real surveillance scenarios, we selected 130 low-quality face images from the SCface dataset [45] for face reconstruction experiments, and these low-resolution images have no corresponding high-resolution images .

In this experiment, we use the identity similarity indicator to measure the preservation of identity information in SR images. Similar to [3], we first use the MTCNN network to perform face detection, alignment, and cropping on SR face images and HR face images, and then, use the pre-trained FaceNet [50] model to extract the identity eigenvectors of the preprocessed SR face images and HR face images, and finally, calculate the cosine similarity between two eigenvectors as the value of identity similarity. The visualization results of face reconstruction (8 × SR) on the SCface dataset by different methods are shown in Figure 11. We can conclude that SwinDPSR generates sharper face images and preserves more facial structure information than other methods. Table 6 lists the average identity similarity of 130 cases of face recognition in the real monitoring scene. We can conclude that the face reconstruction methods integrated with Swin Transformer can effectively improve the performance of face recognition by capturing the global face structure.

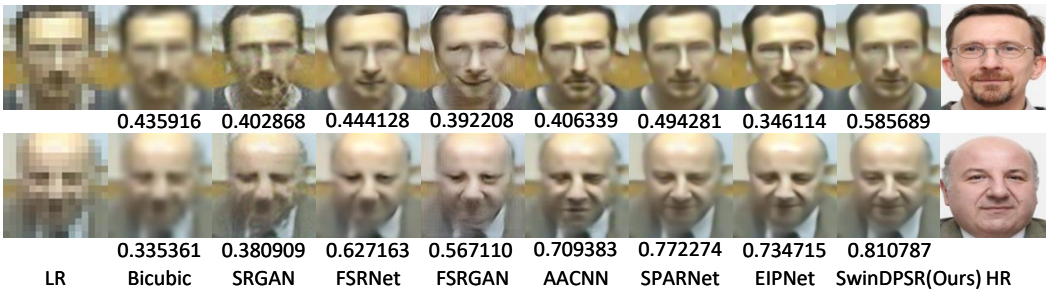

Figure 11. Visual reconstruction results on real-world surveillance scenarios for 8 × SR. The indicator below the SR images is the identity similarity between the SR image and the corresponding HR image.

Table 5. Quantitative results compared with state-of-the-art super-resolution methods. Best results are **bolded** and suboptimal results are underlined.

	Helen Dataset				FFHQ Dataset			
	PSNR	**SSIM**	**LPIPS**	**MPS**	**PSNR**	**SSIM**	**LPIPS**	**MPS**
Bicubic	24.5312	0.6981	0.5030	0.5975	24.2786	0.6609	0.5378	0.5615
SRGAN	25.2783	0.7171	0.1964	0.7603	24.6129	0.6735	0.2052	0.7341
FSRNet	26.9341	0.7950	0.2212	0.7869	26.4785	0.7673	0.2272	0.7700
FSRGAN	25.8452	0.7556	**0.1379**	0.8088	25.191	0.7191	**0.1380**	0.7905
AACNN	26.7893	0.7867	0.2369	0.7748	26.2496	0.7511	0.4811	0.6349
SPARNet	**28.2816**	0.8328	0.2037	0.8145	26.8418	0.7894	0.2245	0.7824
EIPNet	26.8985	0.7912	0.1913	0.7999	26.7129	0.7717	0.2192	0.7762
SwinDPSR	28.7688	**0.8449**	0.1799	**0.8325**	**27.9004**	**0.8099**	0.1886	**0.8106**

Table 6. Comparison results for matching average similarity of face images reconstructed by different methods. Best results are **bolded** and suboptimal results are underlined.

Method	**Average Identity Similarity**
Bicubic	0.228293
SRGAN	0.302244

Table 6. *Cont.*

Method	Average Identity Similarity
FSRNet	0.385995
FSRGAN	0.359804
AACNN	0.445988
SPARNet	0.506417
EIPNet	0.480793
SwinDPSR	**0.516619**

5. Conclusions

In this paper, leveraging the inherent symmetry of face images, we proposed a new dual-path FSR model fused with Swin Transformer. This model uses Swin Transformer to pay attention to global information, and performs feature fusion with the facial attention unit composed of a CNN. Maintaining the consistency of the global structure of the face while focusing on local details improves the fidelity of face reconstruction to a certain extent. Moreover, we additionally provide style loss and SSIM loss to constrain the network model training from the image perception level and image structure level, respectively, and use the channel attention mechanism of the ECA module to reduce the network's attention to redundant features. Extensive experiments and ablation studies demonstrate the effectiveness of SwinDPSR. However, SwinDPSR also has certain limitations. Although our method has achieved good performance in some evaluation indicators, it is still not the best from the perspective of network parameters and calculation. Therefore, how to optimize the network structure and reduce the amount of training parameters under the premise of ensuring the reconstruction performance will be the focus of future research.

Author Contributions: Conceptualization, M.G. and X.Z.; methodology, J.W. and X.L. (Xindong Lv); software, H.Z.; validation, X.L.(Xindong Lv); formal analysis, X.L. (Xing Liu) and M.G.; data curation, X.Z.; writing—original draft preparation, X.L.(Xing Liu) and Y.L.; writing—review and editing, H.D.; visualization, H.Z.; supervision, X.L. (Xing Liu) and Y.L.; project administration, H.D.; funding acquisition, H.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Central Guided Local Science and Technology Development Fund of Shanxi Province, project name: "Research on Key Techniques for Improving the Quality of Low-Quality Images" (YDZJSX2022A016), and the Key Research and Development Program of Shanxi Province: Project name: "Research and Development of Intelligent Monitoring and Harvesting Robots for Economic Orchards" (2022ZDYF128).

Data Availability Statement: The datasets used in this paper are publicly available at http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html (CelebA, accessed on 11 April 2024), https://github.com/zhfe99/helen (Helen, accessed on 11 April 2024), and https://github.com/NVlabs/ffhq-dataset (FFHQ, accessed on 11 April 2024).

Acknowledgments: We would like to acknowledge that there are no additional sources of support or contributions beyond those already mentioned in the author contribution and funding sections.

Conflicts of Interest: Authors Xing Liu, Yan Li, Miao Gu, Hailong Zhang, Xiaoguang Zhang were employed by the company China FAW Group Corporation. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Zhou, E.; Fan, H.; Cao, Z.; Jiang, Y.; Yin, Q. Learning face hallucination in the wild. In Proceedings of the AAAI Conference on Artificial Intelligence, Austin, TX, USA, 25–30 January 2015; No. 1.
2. Liu, H.; Han, Z.; Guo, J.; Ding, X. A noise robust face hallucination framework via cascaded model of deep convolutional networks and manifold learning. In Proceedings of the 2018 IEEE International Conference on Multimedia and Expo (ICME), San Diego, CA, USA, 23–27 July 2018; pp. 1–6.
3. Chen, C.; Gong, D.; Wang, H.; Li, Z.; Wong, K.-Y.K. Learning spatial attention for face super-resolution. *IEEE Trans. Image Process.* **2020**, *30*, 1219–1231. [CrossRef] [PubMed]

4. Liu, S.; Xiong, C.; Shi, X.; Gao, Z. Progressive face super-resolution with cascaded recurrent convolutional network. *Neurocomputing* **2021**, *449*, 357–367. [CrossRef]
5. Shi, J.; Wang, Y.; Yu, Z.; Li, G.; Hong, X.; Wang, F.; Gong, Y. Exploiting multi-scale parallel self-attention and local variation via dual-branch transformer-cnn structure for face super-resolution. *IEEE Trans. Multimed.* **2023**, *26*, 2608–2620 . [CrossRef]
6. Hou, H.; Xu, J.; Hou, Y.; Hu, X.; Wei, B.; Shen, D. Semi-cycled generative adversarial networks for real-world face super-resolution. *IEEE Trans. Image Process.* **2023**, *32*, 1184–1199. [CrossRef]
7. Chen, Y.; Tai, Y.; Liu, X.; Shen, C.; Yang, J. Fsrnet: End-to-end learning face super-resolution with facial priors. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2018, Salt Lake City, UT, USA, 18–23 June 2018; pp. 2492–2501.
8. Zhang, Y.; Wu, Y.; Chen, L. Msfsr: A multi-stage face super-resolution with accurate facial representation via enhanced facial boundaries. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops 2020, Seattle, WA, USA, 14–19 June 2020; pp. 504–505.
9. Yin, Y.; Robinson, J.; Zhang, Y.; Fu, Y. Joint super-resolution and alignment of tiny faces. In Proceedings of the AAAI Conference on Artificial Intelligence 2020; New York City, NY, USA, 7–12 February 2020; Volume 7, pp. 12693–12700.
10. Yu, X.; Fernando, B.; Ghanem, B.; Porikli, F.; Hartley, R. Face super-resolution guided by facial component heatmaps. In Proceedings of the European Conference on Computer Vision (ECCV) 2018, Munich, Germany, 8–14 September 2018; pp. 217–233.
11. Kim, J.; Li, G.; Yun, I.; Jung, C.; Kim, J. Edge and identity preserving network for face super-resolution. *Neurocomputing* **2021**, *446*, 11–22. [CrossRef]
12. Lu, Y.; Tai, Y.-W.; Tang, C.-K. Attribute-guided face generation using conditional cyclegan. In Proceedings of the European Conference on Computer Vision (ECCV) 2018, Munich, Germany, 8–14 September 2018; pp. 282–297.
13. Lee, C.-H.; Zhang, K.; Lee, H.-C.; Cheng, C.-W.; Hsu, W. Attribute augmented convolutional neural network for face hallucination. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops 2018, Salt Lake City, UT, USA, 18–23 June 2018; pp. 721–729.
14. Li, M.; Sun, Y.; Zhang, Z.; Xie, H.; Yu, J. Deep learning face hallucination via attributes transfer and enhancement. In Proceedings of the 2019 IEEE International Conference on Multimedia and Expo (ICME), Shanghai, China, 8–12 July 2019; pp. 604–609.
15. Li, M.; Zhang, Z.; Yu, J.; Chen, C.W. Learning face image super-resolution through facial semantic attribute transformation and self-attentive structure enhancement. *IEEE Trans. Multimed.* **2020**, *23*, 468–483. [CrossRef]
16. Jiang, K.; Wang, Z.; Yi, P.; Lu, T.; Jiang, J.; Xiong, Z. Dual-path deep fusion network for face image hallucination. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *33*, 378–391. [CrossRef]
17. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual attention network for scene segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2019, Long Beach, CA, USA, 16–20 June 2019; pp. 3146–3154.
18. Tao, A.; Sapra, K.; Catanzaro, B. Hierarchical multi-scale attention for semantic segmentation. *arXiv* **2020**, arXiv:2005.10821.
19. Zhong, Z.; Lin, Z.Q.; Bidart, R.; Hu, X.; Daya, I.B.; Li, Z.; Zheng, W.S.; Li, J.; Wong, A. Squeeze-and-attention networks for semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2020, Seattle, WA, USA, 14–19 June 2020; pp. 13065–13074.
20. Zhang, Y.; Li, K.; Li, K.; Wang, L.; Zhong, B.; Fu, Y. Image super-resolution using very deep residual channel attention networks. In Proceedings of the European Conference on Computer Vision (ECCV) 2018, Munich, Germany, 8–14 September 2018; pp. 286–301.
21. Qin, X.; Wang, Z.; Bai, Y.; Xie, X.; Jia, H. FFA-net: Feature fusion attention network for single image dehazing. In Proceedings of the AAAI Conference on Artificial Intelligence 2020, New York, NY, USA, 7–12 February 2020; pp. 11908–11915.
22. Tian, C.W.; Xu, Y.; Li, Z.Y.; Zuo, W.M.; Fei, L.K.; Liu, H. Attention-guided CNN for image denoising. *Neural Netw.* **2020**, *124*, 117–129. [CrossRef]
23. Vaswani, A.; Shazeer, N.M.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems 30 (NIPS 2017); Long Beach, CA, USA, 4–9 December 2017, Volume 30.
24. 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* **2020**, arXiv:2010.11929.
25. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021; pp. 10012–10022.
26. Liang, J.; Cao, J.; Sun, G.; Zhang, K.; Gool, L.V.; Timofte, R. SwinIR: Image restoration using Swin transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision 2021, Montreal, QC, Canada, 11–17 October 2021; pp. 1833–1844.
27. Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, X.; Tian, Q.; Wang, M. Swin-Unet: Unet-like pure transformer for medical image segmentation. In Proceedings of the European Conference on Computer Vision 2022, Tel-Aviv, Israel, 23–27 October 2022; pp. 205–218.
28. Song, Y.; He, Z.; Qian, H.; Du, X. Vision transformers for single image dehazing. *IEEE Trans. Image Process.* **2023**, *32*, 1927–1941. [CrossRef]
29. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In Proceedings of the European Conference on Computer Vision 2020, Glasgow, UK, 23-28 August 2020; pp. 213–229.
30. Yang, F.; Yang, H.; Fu, J.; Lu, H.; Guo, B. Learning texture transformer network for image super-resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2020, Seattle, WA, USA, 14–19 June 2020; pp. 5791–5800.

31. Newell, A.; Yang, K.; Deng, J. Stacked hourglass networks for human pose estimation. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Proceedings, Part VIII; pp. 483–499.
32. Bulat, A.; Tzimiropoulos, G. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In Proceedings of the IEEE International Conference on Computer Vision 2017, Venice, Italy, 22–29 October 2017; pp. 1021–1030.
33. Odena, A.; Dumoulin, V.; Olah, C. Deconvolution and checkerboard artifacts. *Distill* **2016**, *1*, e3. [CrossRef]
34. Shaw, P.; Uszkoreit, J.; Vaswani, A. Self-attention with relative position representations. *arXiv* **2018**, arXiv:1803.02155.
35. Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P.J. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* **2020**, *21*, 5485–5551.
36. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2018, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
37. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. Eca-net: Efficient channel attention for deep convolutional neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2020; Seattle, WA, USA, 14–19 June 2020; pp. 11534–11542.
38. Li, X.; Li, W.; Ren, D.; Zhang, H.; Wang, M.; Zuo, W. Enhanced blind face restoration with multi-exemplar images and adaptive spatial feature fusion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2020, Seattle, WA, USA, 14–19 June 2020; pp. 2706–2715.
39. Parkhi, O.; Vedaldi, A.; Zisserman, A. Deep face recognition. In Proceedings of the BMVC 2015, British Machine Vision Conference 2015, Swansea, Wales, United Kingdom, 7–10 September 2015.
40. Gatys, L.A.; Ecker, A.S.; Bethge, M. Image style transfer using convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2016, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 2414–2423.
41. Liu, Z.; Luo, P.; Wang, X.; Tang, X. Deep learning face attributes in the wild. In Proceedings of the IEEE International Conference on Computer Vision 2015, Santiago, Chile, 7–13 December 2015; pp. 3730–3738.
42. Zhang, K.; Zhang, Z.; Li, Z.; Qiao, Y. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Process. Lett.* **2016**, *23*, 1499–1503. [CrossRef]
43. Le V.; Br, t, J.; Lin, Z.; Bourdev, L.; Huang, T.S. Interactive facial feature localization. In Proceedings of the Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, 7–13 October 2012; Proceedings, Part III 12; pp. 679–692.
44. Karras, T.; Laine, S.; Aila, T. A style-based generator architecture for generative adversarial networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern recognition 2019, Long Beach, CA, USA, 16–20 June 2019; pp. 4401–4410.
45. Grgic, M.; Delac, K.; Grgic, S. Scface–surveillance cameras face database. *Multimed. Tools Appl.* **2011**, *51*, 863–879. [CrossRef]
46. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. Pytorch: An imperative style, high-performance deep learning library. In Proceedings of the Advances in Neural Information Processing Systems 32 (NeurIPS 2019), Vancouver, BC, Canada, 8–14 December 2019; Volume 32.
47. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [CrossRef] [PubMed]
48. Zhang, R.; Isola, P.; Efros, A.A.; Shechtman, E.; Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2018, Salt Lake City, UT, USA, 18–23 June 2018; pp. 586–595.
49. El Helou, M.; Zhou, R.; Süsstrunk, S.; Timofte, R.; Afifi, M.; Brown, M.S.; Xu, K.; Cai, H.; Liu, Y.; Wang, L.W.; et al. Aim 2020: Scene relighting and illumination estimation challenge. In Proceedings of the Computer Vision–ECCV 2020 Workshops, Glasgow, UK, 23–28 August 2020; Proceedings, Part III 16; pp. 499–518.
50. Schroff, F.; Kalenichenko, D.; Philbin, J. Facenet: A unified embedding for face recognition and clustering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2015, Boston, MA, USA, 7–12 June 2015; pp. 815–823.
51. Ledig, C.; Theis, L.; Huszár, F.; Caballero, J.; Aitken, A.P.; Tejani, A.; Totz, J.; Wang, Z.; Shi, W. Photo-realistic single image super-resolution using a generative adversarial network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2017, Honolulu, HI, USA, 21–26 July 2017; pp. 4681–4690.