Attention-Guided Multiscale Interaction Network for Face Super-Resolution

Xujie Wan, Wenjie Li, Guangwei Gao, *Senior Member, IEEE*, Huimin Lu, *Senior Member, IEEE*, Jian Yang, *Member, IEEE*, and Chia-Wen Lin, *Fellow, IEEE*

Abstract— Recently, CNN and Transformer hybrid networks demonstrated excellent performance in face super-resolution (FSR) tasks. Because of numerous features at different scales in hybrid networks, how to fuse these multiscale features and promote their complementarity is crucial for enhancing FSR. However, existing hybrid network-based FSR methods ignore this, only simply combining the Transformer and CNN. To address this issue, we propose an attention-guided multiscale interaction network (AMINet), which incorporates local and global feature interactions, as well as encoder–decoder phase feature interactions. Specifically, we propose a local and global feature interaction (LGFI) module to promote the fusion of global features and the local features extracted from different receptive fields by our residual depth feature extraction (RDFE) module. Additionally, we propose a selective kernel attention fusion (SKAF) module to adaptively select fusions of different features within the LGFI and encoder–decoder phases. Our above design allows the free flow of multiscale features from within modules and between the encoder and decoder, which can promote the complementarity of different scale features to enhance FSR. Comprehensive experiments confirm that our method consistently performs well with less computational consumption and faster inference.

Index Terms— Attention-guided, face super-resolution (FSR), hybrid networks, multiscale interaction.

I. INTRODUCTION

FACE super-resolution (FSR), also known as face hallucination, aims at restoring high-resolution (HR) face images from low-resolution (LR) face images [1]. In contrast to standard image super-resolution, the primary objective of FSR is to reconstruct as many facial structural features as possible (i.e., the shape and contour of facial components). In practical scenarios, a range of face-specific tasks, such as face detection [2] and face recognition [3], require HR face images. However, the quality of captured face images is frequently diminished due to variations in hardware configuration, positioning, and shooting angles of the imaging devices, seriously affecting the above downstream tasks. Therefore, FSR has garnered increasing attention in recent years.

Recently, because of the advantages exhibited by hybrid networks [5] of CNNs and Transformers in FSR, this type of method has gained increased attention. Specifically, CNN-based FSR methods [6] generally do not require large computational costs. Still, they specialize in extracting local details, such as the local texture of the face and color, and are unable to model long-range feature interaction, such as the global profile of the face. Transformer-based FSR methods [7] can simulate global modeling well, but their computational consumption is huge. Hybrid FSR methods leverage the strengths of both architectures, enabling efficient extraction of both local and global features while maintaining a reasonable computational overhead. The impressive performance of hybrid-based FSR methods comes from numerous features extracted inside their networks at different scales, such as global features from self-attention (SA), local features from convolution, and features from different stages of the encoder–decoder, which facilitates models to refine local details and global contours.

However, while existing hybrid-based FSR methods consider utilizing features from different scales to improve FSR, they ignore the problem of how to fuse these multiscale features to make their properties better complement each other. For example, FaceFormer [8] simply parallelizes the connected CNN modules and the window-based Transformer [9] modules. SCTANet [10] also only juxtaposes spatial attention-based residual blocks and multihead SA in designed modules. CTCNet [5] simply connects the CNN module in tandem with the Transformer module. These methods overlook the importance of blending multiscale features in a complementary manner and enabling smooth information flow across different scales to refine facial details effectively.

To address this problem, we propose an attention-guided multiscale interaction network (AMINet) for FSR in this work. Our AMINet fuses multiscale features in two main
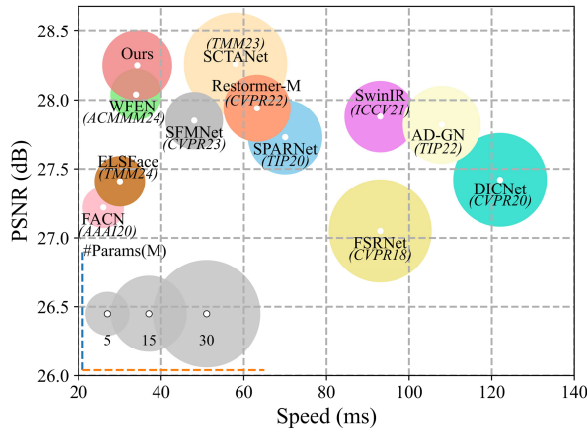
Fig. 1. Model complexity studies for $\times 8$ FSR on CelebA test set [4]. Our AMINet achieves an excellent balance among model size, model performance, and inference speed.

ways, including the fusion of features obtained from SA and convolution, and the fusion of features at different stages of the encoder–decoder. Specifically, we design a local and global feature interaction (LGFI) module to adaptively fuse global facial and local features with different receptive fields obtained by convolutions. In LGFI, SA is responsible for extracting global features, while our residual depth feature extraction (RDFE) module extracts local features at different scales using separable convolutional kernels of different sizes, and our selective kernel attention fusion (SKAF) module is responsible for weighted fusion of these two parts of features for our model to adaptively perform selective fusion during training. In addition, we also utilize our SKAF as a crucial fusion module in our encoder and decoder feature fusion (EDFF) module to further perform feature communications of our method by fusing features at different scales from the encoder–decoder processes.

Our above design greatly enhances the flow and exchange of features at different scales within the model and improves the representation of our model. As a result, our method can obtain a more powerful feature representation than existing FSR methods. As shown in Fig. 1, our method can achieve the best FSR performance with a smaller size and faster inference speed, demonstrating our method's effectiveness. In summary, the main contributions are as follows.

1) We design an LGFI to differ from the traditional Transformer by allowing free flow and selective fusion of local and global features within the module.
2) We design an RDFE, which enables better refinement of facial details by fusion and refinement of local features extracted by convolutional kernels of different sizes.
3) We design the SKAF to help selective fusions of different-scale features within LGFI and EDFF by selecting appropriate convolutional kernels.

II. RELATED WORK

A. Face Super-Resolution

Early deep learning approaches focused on leveraging facial priors as guidance to enhance FSR accuracy [11]. For instance,

Chen et al. [12] developed an end-to-end prior-based network that utilized facial landmarks and heatmaps to generate FSR images. Similarly, Kim et al. [13] employed a face alignment network for landmark extraction in conjunction with a progressive training technique to produce realistic face images. Ma et al. [14] introduced DICNet, which iteratively integrates facial landmark priors to enhance image quality at each step. Hu et al. [15] explored the use of 3-D shape priors to better capture and define sharp facial structures. While these methods have advanced FSR, they require additional labeling of training datasets. Moreover, inaccuracies in prior estimation can significantly diminish FSR performance, especially when dealing with highly blurred face images.

Attention-based FSR methods have been proposed to promote FSR to avoid the adverse effects of inaccurate prior estimates on FSR. Zhang and Ling [16] proposed a supervised pixel-by-pixel generation of the adversarial network to improve face recognition performance during FSR. Chen et al. [17] proposed the SPARNet, which can focus on important facial structure features adaptively by using spatial attention in residual blocks. Lu et al. [18] proposed a partial attention mechanism to enhance the consistency of the fidelity of facial detail and facial structure. Bao et al. [19] introduced the equalization texture enhancement module to enhance the facial texture detail through histogram equalization. Wang et al. [20] critically introduced the Fourier transform into FSR, fully exploring the correlation between spatial domain features and frequency domain features. Shi et al. [21] designed a two-branch network, which introduces a convolution based on local changes to enhance the ability of the convolution. Bao et al. [10] improves the interaction ability of regional and global features through designed hybrid attention modules. Li et al. [22] designed a wavelet-based network to reduce the loss of downsampling in the encoder–decoder. Although the above methods can reconstruct reasonable FSR images, they cannot promote the efficient fusion of local features with global features and different features at different stages of the encoder–decoder, affecting FSR's efficiency and accuracy.

B. Attention-Based Super-Resolution

The attention mechanism can improve the super-resolution accuracy of models due to its flexibility in focusing on key areas of facial features. In the super-resolution task, different variants of the attention mechanism include SA, spatial attention, channel attention, and hybrid attention [23].

Zhang et al. [24] inserted channel attention into residual blocks to enhance model representation. Xin et al. [25] utilized channel attention plus residual mechanisms to combine a multilevel information fusion strategy. Chen et al. [17] enhanced FSR by utilizing an improved facial spatial attention that cooperated with the hourglass structure. Gao et al. [26] performed shuffling to hybrid attention. Wang et al. [27] constructed a simplified feed-forward network (FFN) using spatial attention to reduce parameters and computational complexity. To model long-range feature interaction, the SA in Transformer [28] has been widely used in super-resolution. Gao et al. [29] reduced
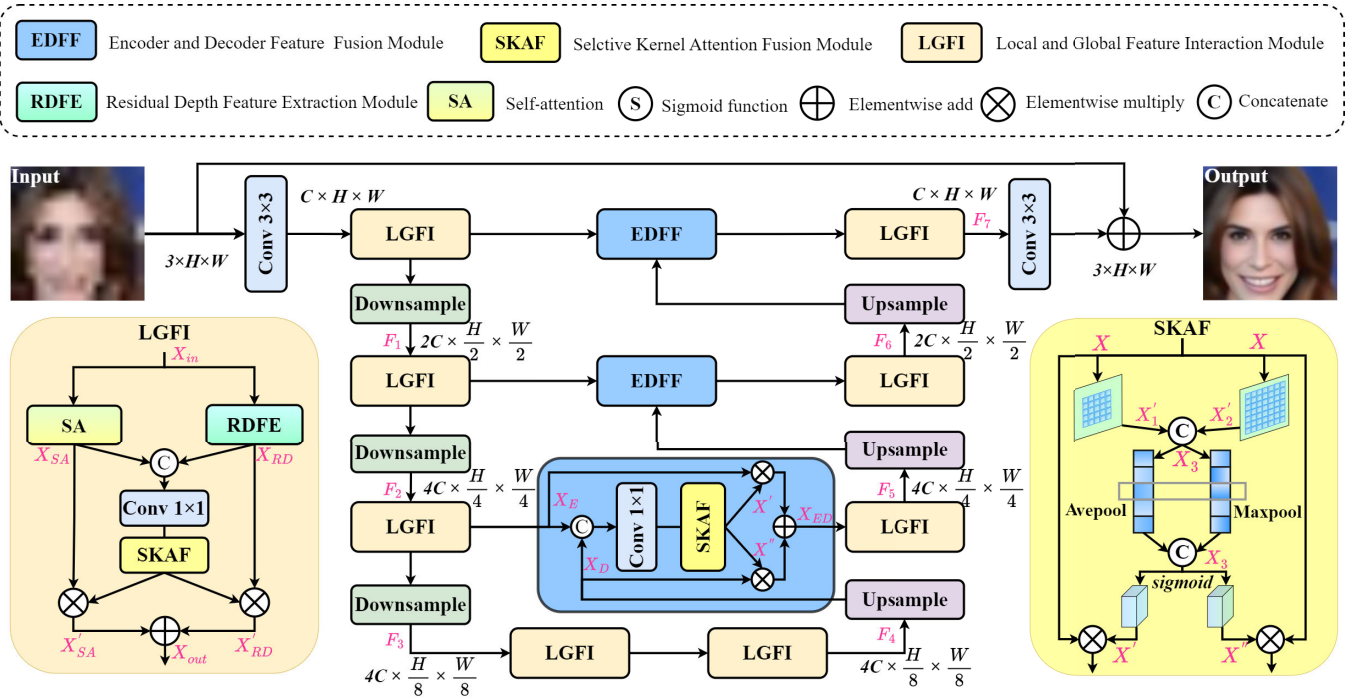
Fig. 2. Network structure of our AMINet, which is a U-shaped CNN-transformer hierarchical architecture with three distinct stages: encoding, bottleneck, and decoding.

costs by utilizing the recursive mechanism on SA. Li et al. [30] and [31] combined SA and convolutions to complement each other's required features. Zeng et al. [32] introduced an SA network that investigates the relationships among features at various levels. Shi et al. [21] enhanced FSR by mitigating the adverse effects of inaccurate prior estimates through a parallel SA mechanism, effectively capturing both local and nonlocal dependencies. To combine the advantages of different attentions, Yang and Qi [33] integrated channel attention with spatial attention to enhance feature acquisition and correlation modeling. Bao et al. [10] and Gao et al. [5] employed spatial attention and SA to capture facial structure and details. Zhang and Qi [34] employed a hybrid attention that combines SA, spatial attention, and channel attention to optimize fine-grained facial details and broad facial structure. Unlike the above attention-based methods that enhance model representation, we utilize attention to learn features from different receptive fields, allowing our network to adaptively select the appropriate convolutional kernel size to match the multiscale feature fusion. This design enables our network to perform multiscale feature extraction and improve the integration of features across various scales, leading to enhanced performance and greater adaptability.

III. PROPOSED METHOD

A. Overview of AMINet

As illustrated in Fig. 2, our proposed AMINet features a U-shaped CNN-Transformer hierarchical architecture with three distinct stages: encoding, bottleneck, and decoding. For an LR input face image $I_{LR} \in \mathbb{R}^{3 \times H \times W}$, in the encoding stage, our network aims to extract features at different scales and

capture multiscale feature representations of the input image to get the facial feature $F_3 \in \mathbb{R}^{C \times H \times W}$. Then, the bottleneck stage network continues to refine the feature F_3 and provides a more informative representation to get the refined feature F_4 for the subsequent reconstruction phase. In decoding, the network focuses on feature upsample and facial detail reconstruction. Meanwhile, an interactive connection is used between the encoding and decoding stages to ensure the features are fully integrated throughout the network. We can get the reconstructed face feature F_7 with rich facial details through the above operators. Finally, through a convolution with reduced channel dimensions plus a residual connection, we get the HR output face image $I_{HR} \in \mathbb{R}^{3 \times H \times W}$.

1) Encoding Stage: The encoding stage in our network aims to extract facial features of different scales. In this stage, given a degraded face image $I_{LR} \in \mathbb{R}^{3 \times H \times W}$, first, a 3×3 convolution is used to extract shallow facial features. Then, extracted facial features are further refined by three encoder stages. Each encoder comprises our designed LGFI module and a downsampling operator. After each encoder, the input face feature's channel count will be doubled, and the size of the image of the input face feature will be halved. As shown in Fig. 2, the features obtained after three encoders are as follows: $F_1 \in \mathbb{R}^{2C \times (H/2) \times (W/2)}$, $F_2 \in \mathbb{R}^{4C \times (H/4) \times (W/4)}$, and $F_3 \in \mathbb{R}^{8C \times (H/8) \times (W/8)}$.

2) Bottleneck Stage: In the bottleneck stage between the encoding and decoding stages, the obtained encoding features are designed to be fine-grained. $F_4 \in \mathbb{R}^{8C \times (H/8) \times (W/8)}$ is obtained through the bottleneck stage. In this stage, we continue to use two LGFIs to refine and enhance encoding features to ensure they are better utilized in the decoding stage. After this stage, our model can continuously enhance the information
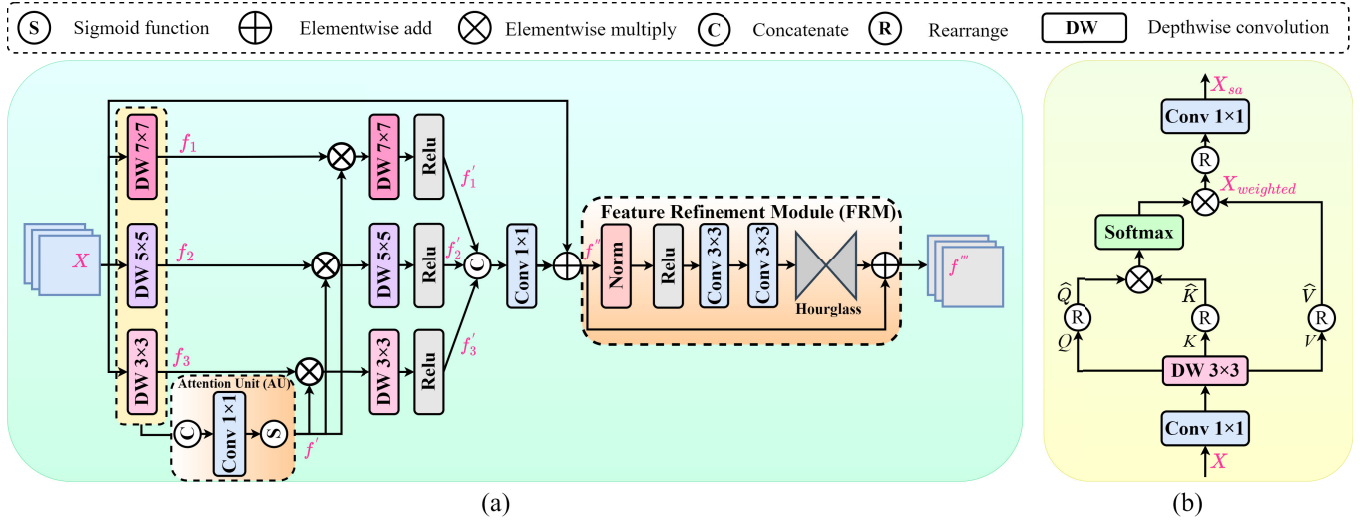
Fig. 3. Architecture of (a) RDFE module and (b) SA, respectively.

about the facial structure at different scales, thus improving the perception of facial details.

3) Decoding Stage: In the decoding stage, there are three decoders. We focus on multiscale feature fusion, aiming at reconstructing high-quality face images at this stage. As depicted in Fig. 2, each decoder includes an upsampling operation, an EDFF, and an LGFI. Each upsampling operator halves the input feature channel counts while doubling the width and weight of the input facial feature. Compared to encoding stages, decoding stages additionally use our proposed SKAF to adaptively and selectively fuse different scale features from the encoder and decoder stages. Through this design, different-scale features can interact to recover more detailed face features. The features obtained after three decoders are as follows: $F_5 \in \mathbb{R}^{4C \times (H/4) \times (W/4)}$, $F_7 \in \mathbb{R}^{C \times H \times W}$. Finally, a 3×3 convolution unit is utilized to transform our obtained deep facial feature into an output FSR image I_{SR}.

B. LGFI Module

In our AMINet, LGFI is mainly used for local and global facial feature extraction. As shown in Fig. 2, LGFI consists of SA, an RDFE module, and an SKAF module, used for local and global feature fusion and interaction, respectively. The SA is designed to extract global features. At the same time, RDEM is designed to extract local features at different scales and enrich local facial details through multiple convolutional kernels under numerous receptive fields. Specifically, the integration of local and global features is achieved through SKAF, which adaptively weights and fuses multiscale information. SKAF first extracts local and global features via convolutions with different receptive fields and then computes their importance using average pooling. These computed weights are applied to the corresponding features, facilitating an adaptive fusion process. By dynamically selecting and emphasizing key information, SKAF enhances the complementarity between local textures and global structural cues, ultimately improving

facial detail reconstruction and overall super-resolution performance.

1) Self-Attention: We utilize SA to extract global facial features, which can effectively model the relationships between distant features. Meanwhile, through the multihead mechanism in SR, features can be captured from different subspaces, improving the robustness and generalization ability of the model. As illustrated in Fig. 3(b), we start by applying a 1×1 convolutional layer followed by a 3×3 depthwise convolutional layer to combine pixel-level cross-channel information and extract channel-level spatial context. From this spatial context, we then generate $Q, K, V \in \mathbb{R}^{C \times H \times W}$. For an input facial feature $X \in \mathbb{R}^{C \times H \times W}$, the process of obtaining $Q, K, V \in \mathbb{R}^{C \times H \times W}$ can be described as follows:

$$Q = F_{\text{dw3}}\left(F_{\text{conv1}}\left(X\right)\right) \tag{1}$$

$$K = F_{\text{dw3}}\left(F_{\text{conv1}}\left(X\right)\right) \tag{2}$$

$$V = F_{\text{dw3}}\left(F_{\text{conv1}}\left(X\right)\right) \tag{3}$$

where $F_{\text{conv1}}(\cdot)$ is the 1×1 pointwise convolution and $F_{\text{dw3}}(\cdot)$ is the 3×3 depthwise convolution.

Next, we reshape Q, K, and V into $\hat{Q} \in \mathbb{R}^{C \times HW}$, $\hat{K} \in \mathbb{R}^{HW \times C}$, and $\hat{V} \in \mathbb{R}^{C \times HW}$, respectively. After that, the dot product is multiplied by V to obtain weights $X_w \in \mathbb{R}^{C \times HW}$, which facilitates the capturing of the important local context in SA. Finally, we rearrange X_w into $\hat{X_w} \in \mathbb{R}^{C \times H \times W}$. The above operations can be expressed as follows:

$$X_{\text{weighted}} = \text{Softmax}\left(\hat{Q} \cdot \hat{K} / \sqrt{d}\right) \cdot \hat{V} \tag{4}$$

$$X_{\text{sa}} = F_{\text{conv1}}\left(R\left(X_{\text{weighted}}\right)\right) \tag{5}$$

where X_{sa} is the attention map of SA, $(d)^{1/2}$ is a factor used to scale the dot product of $\hat{K}$ and $\hat{Q}$, $R(\cdot)$ stands for the rearrange operation, and X_{sa} denotes the output of SA.

2) RDFE Module: As shown in Fig. 3(a), we design RDFE to extract local facial features at different scales. Compared with the traditional FFN, our RDFE is beneficial for processing more complex features and multiscale features flexibly. Specifically, for the input feature $X \in \mathbb{R}^{C \times H \times W}$, we use

depthwise convolutions of 3×3, 5×5, and 7×7 to parallelly extract three scales of facial features, which depthwise convolution can reduce the computational complexity of the model, while convolution with different kernel sizes can effectively extract rich face details. The reason we employ depthwise separable convolution is to significantly reduce computational complexity while preserving local spatial features. In our model, depthwise convolutions at different scales enable the extraction of fine-grained facial details, such as textures and edge information, across varying receptive fields, enhancing the model's ability to capture local features efficiently. The above operations can be expressed as follows:

$$f_1, f_2, f_3 = F_{\text{dw3}}(X), F_{\text{dw5}}(X), F_{\text{dw7}}(X) \qquad (6)$$

where F_{dw3}, F_{dw5}, and F_{dw7} are 3×3, 5×5, and 7×7 depthwise convolution, respectively. However, relying solely on convolutions may lead to insufficient feature fusion across scales, potentially introducing redundant information or affecting the representation of key facial details. To address this issue, we use an attention unit (AU) F_{au} to calculate the feature weight of the fusion feature of three branches. Next, we use the calculated weights to modulate the three-branch features at different scales through elementwise multiplications. This mechanism enables the model to dynamically balance the contributions of multiscale features, effectively capturing key facial details across varying receptive fields, thereby enhancing detail restoration and overall generalization. To further refine and reconstruct the weighted fused facial features, we apply depthwise convolutions with kernel sizes of 3×3, 5×5, and 7×7, ensuring the joint optimization of local texture details and global structural consistency. This collaborative refinement ultimately leads to higher-quality facial reconstructions. The above operations can be expressed as follows:

$$f' = F_{\text{au}}(H_{\text{cat}}(f_1, f_2, f_3)) \qquad (7)$$
$$f_1', f_2', f_3' = f_1(X) \cdot f', f_2(X) \cdot f', f_3(X) \cdot f' \qquad (8)$$

where H_{cat} is a concat operator and $F_{\text{au}}(\cdot)$ indicates the attention unit. Then, we aggregate the three branches' features to combine facial detail information under different receptive fields. This process can be described as follows:

$$f'' = H_{\text{cat}}(F_{\text{conv1}}(f_1', f_2', f_3')) + X \qquad (9)$$

where $F_{\text{conv1}}(\cdot)$ represents 1×1 convolution. Finally, we utilize a feature refinement module (FRM) to refine features obtained from previous multiple branches. Specifically, we begin by applying normalization and multiple 3×3 convolutional layers to refine the local facial context. Afterward, the hourglass block further integrates multiscale information to capture global and local relationships

$$f''' = F_{\text{frm}}(f'') \qquad (10)$$

where $F_{\text{frm}}(\cdot)$ indicates the feature refinement module.

3) SKAF Module: Inspired by SKNet [35] and LSKNet [36], as shown in Fig. 2, we design an SKAF module to give our model the ability to select local and global features required for reconstruction for fusion interaction. Specifically, SKAF first extracts both local and global features using convolutions with different receptive fields. It then computes feature importance through an average pooling operation, which was previously not explicitly shown in the diagram. Finally, an adaptive weighting mechanism selects and fuses the most relevant local and global information, enabling SKAF to dynamically emphasize critical features under varying receptive fields. This design enhances the model's capacity for feature selection, improving both its performance and generalization in FSR. Given feature $X \in \mathbb{R}^{C \times H \times W}$ obtained by the SA and the RDFE, we first fuse the local and global features extracted by a 5×5 convolution and a 7×7 convolution to get a hybrid feature X. This operation can be expressed as follows:

$$X_1', X_2' = F_{\text{conv5}}(X), F_{\text{conv7}}(X) \qquad (11)$$
$$X_3 = H_{\text{cat}}(X_1', X_2') \qquad (12)$$

where $F_{\text{conv5}}(\cdot)$ represents 5×5 convolution, $F_{\text{conv7}}(\cdot)$ represents 7×7 convolution, and $H_{\text{cat}}(\cdot)$ indicates the concat operation along the channel dimension. Then, we impose pooling to learn the weight of the obtained hybrid features, where the weight reflects the importance of features under different receptive fields. The process of obtaining the weight for selecting required facial features is as follows:

$$X_3 = H_{\text{sig}}(H_{\text{cat}}(H_{\text{avep}}(X_3), H_{\text{maxp}}(X_3))) \qquad (13)$$

where $H_{\text{avep}}(\cdot)$, $H_{\text{maxp}}(\cdot)$, and $H_{\text{sig}}(\cdot)$ indicate the average and max pooling operation along the channel direction and sigmoid function, respectively. Finally, we multiply the weights obtained from the above calculations with the local and global features, respectively. Thus, our SKAF can adaptively select the important local and global information required for reconstruction. The process of obtaining local and global features X', X'' by adaptive weight selection is as follows:

$$X', X'' = H_{\text{cs}}(X_3) \qquad (14)$$

where $H_{\text{cs}}(\cdot)$ indicates the feature separation operation along the channel dimension. Through the above operators, we can get the adaptive selected local and global features.

C. EDFF Module

To fully utilize the multiscale features extracted from the encoding and decoding stage, we introduce an EDFF to fuse different features, enabling our AMINet with better feature propagation and representation capabilities. As shown in Fig. 2, our EDFF mainly utilizes our proposed SKAF to fuse and select different-scale features required for reconstruction. Given the feature $X_E \in \mathbb{R}^{C \times H \times W}$, the feature $X_D \in \mathbb{R}^{C \times H \times W}$ is from the decoding stage and the encoding stage, respectively. First, we concatenate features from the encoding and decoding stages along the channel dimension. Then, a 1×1 convolution is used to reduce the channel counts and reduce the process's computational costs to obtain two weights through our SKAF, which can be expressed as follows:

$$X', X'' = F_{\text{skaf}}(F_{\text{conv1}}(H_{\text{cat}}(X_E, X_D))) \qquad (15)$$

where $F_{\text{skaf}}(\cdot)$ represents the SKAF module, $F_{\text{conv1}}(\cdot)$ stands for the 1×1 convolutional layer, and $H_{\text{cat}}(\cdot)$ denotes the operation of concatenating features across the channel dimension. Next, we feed the two obtained weights into two branches for multiplication. Through this operator, we obtain the selected facial features from hybrid features obtained by fusing the encoding and decoding features. Finally, we add the features of the two branches

$$X_{ED} = X_E \cdot X' + X_D \cdot X''. \qquad (16)$$

Through the above operators, we can complete the process of the adaptive fusion of encoding and decoding features.

D. Loss Functions

As for the loss of our AMINet, given a dataset $\{I_{\text{LR}}^i, I_{\text{HR}}^i\}_{i=1}^N$, we optimize our AMINet by minimizing the pixel-level loss function

$$\mathcal{L}(\Theta) = \frac{1}{N} \sum_{i=1}^N \left\| F_{\text{AMINet}}\left(I_{\text{LR}}^i, \Theta\right) - I_{\text{HR}}^i \right\|_1 \qquad (17)$$

where N denotes the paired training face image counts. I_{LR}^i and I_{HR}^i are the face LR image and HR image of the ith pair, respectively. Meanwhile, $F_{\text{AMINet}}(\cdot)$ and Θ denote the AMINet and the number of parameters of AMINet, respectively.

Since the GAN-based methods [37], [38] can get better perceptual qualities, we expand our AMINet to AMIGAN to generate more high-quality SR results. The loss function used in training AMIGAN consists of the following three parts.

1) Pixel Loss: Pixel-level loss is used to reduce the pixel difference between the SR and HR images

$$\mathcal{L}_{\text{pix}} = \frac{1}{N} \sum_{i=1}^N \left\| G\left(I_{\text{LR}}^i\right) - I_{\text{HR}}^i \right\|_1 \qquad (18)$$

where G indicates the AMIGAN generator.

2) Perceptual Loss: To enhance the visual quality of super-resolution images, we apply perceptual loss. This involves using a pretrained VGG19 [39] model to extract facial features from both the HR images and our generated FSR images. Then, we compare the obtained perceptual features of HR and FSR images to constrain the generation of FSR features. Therefore, the perceptual loss can be described as follows:

$$\mathcal{L}_{\text{pcp}} = \frac{1}{N} \sum_{i=1}^N \sum_{l=1}^{L_{\text{VGG}}} \frac{1}{M_{\text{VGG}}^l} \left\| f_{\text{VGG}}^l\left(I_{\text{SR}}^i\right) - f_{\text{VGG}}^l\left(I_{\text{HR}}^i\right) \right\|_1 \qquad (19)$$

where f_{VGG}^l represents the feature map from the lth layer of the VGG network, L_{VGG} is the total number of layers in VGG, and M_{VGG}^l indicates the quantity of elements within that feature map.

3) Adversarial Loss: GANs have been shown to be effective in reconstructing photorealistic images [37], [38]. GAN generates FSR results through the generator while using the discriminator to distinguish between ground truth and FSR results, which ultimately enables the generator to generate realistic FSR results in the process of constant confrontation. This process is denoted as follows:

$$\mathcal{L}_{\text{dis}} = -\mathbb{E}\left[\log\left(D\left(I_{\text{HR}}\right)\right)\right] - \mathbb{E}\left[\log\left(1 - D\left(G\left(I_{\text{LR}}\right)\right)\right)\right]. \qquad (20)$$

Additionally, the generator tries to minimize

$$\mathcal{L}_{\text{adv}} = -\mathbb{E}\left[\log\left(D\left(G\left(I_{\text{LR}}\right)\right)\right)\right]. \qquad (21)$$

Thus, AMIGAN is refined by minimizing the following total objective function:

$$\mathcal{L} = \lambda_{\text{pix}}\mathcal{L}_{\text{pix}} + \lambda_{\text{pcp}}\mathcal{L}_{\text{pcp}} + \lambda_{\text{adv}}\mathcal{L}_{\text{adv}} \qquad (22)$$

where λ_{pix}, λ_{pcp}, and λ_{adv} represent the weighting factors for the corresponding pixel loss, perceptual loss, and adversarial loss, respectively.

IV. EXPERIMENTS

A. Datasets and Evaluation Metrics

We utilize the CelebA [4] dataset for training and evaluation on CelebA [4], Helen [40], and SCface [41] datasets, respectively. We center-crop the aligned faces and resize them to 128×128 pixels to obtain HR images. These HR images are then downsampled to 16×16 pixels using bicubic interpolation, producing the corresponding LR images. In our experiments, we randomly chose $18\,000$ CelebA images for training and 1000 for testing. In addition, we also utilize the SCface test set as a real evaluation dataset. To measure the quality of FSR results, we use PSNR [42], SSIM [42], LPIPS [43], VIF [44], and FID [45].

B. Implementation Details

We implement our model using PyTorch on an NVIDIA GeForce RTX 3090. The network is optimized using the Adam optimizer, with parameters set to $\beta_1 = 0.9$ and $\beta_2 = 0.99$. The initial learning rate is 2×10^{-4}, with separate learning rates for the generator and discriminator set at 1×10^{-4} and 4×10^{-4}, respectively. The loss function weights are configured as $\lambda_{\text{pix}} = 1$, $\lambda_{\text{pcp}} = 0.01$, and $\lambda_{\text{adv}} = 0.01$.

C. Ablation Studies

1) Study of LGFI: LGFI is proposed to extract local features and global relationships of images, which represents a new attempt to interact with local and global information. To verify the reasonableness of our design of LGFI, as shown in Table I, we design four ablation models. The first model removes the SA, labeled "LGFI w/o SA." The second model removes RDFE, labeled as "LGFI w/o RDFE." The third model removes SKAF, labeled as "LGFI w/o SKAF." We have the following observations.

1) Introducing SA and RDFE alone can improve model performance. This is because the above two modules can capture local and global features to promote facial feature reconstruction, including facial details and overall contours.
2) Model performance has been significantly increased by introducing the SKAF to capture the relationship

TABLE I
VERIFY THE EFFECTIVENESS OF LGFI (CelebA, ×8)

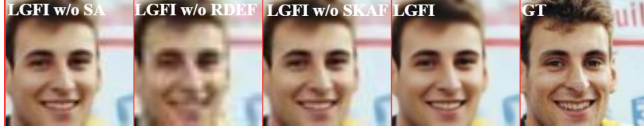| Methods	PSNR/SSIM↑	VIF↑	LPIPS↓	Speed↓	Params↓
LGFI w/o SA	27.75/0.7944	0.4652	0.1886	33ms	12.3M
LGFI w/o RDFE	27.51/0.7840	0.4495	0.2085	21ms	7.2M
LGFI w/o SKAF	27.74/0.7932	0.4611	0.1979	25ms	8.21M
LGFI	**27.83/0.7961**	**0.4725**	**0.1821**	34ms	12.62M

Fig. 4. Visual comparison of various outputs of LGFI.

TABLE II
QUANTITATIVE COMPARISON BETWEEN LGFI AND TRADITIONAL
TRANSFORMER AS SHOWN IN FIG. 6 (CelebA, ×8)

Methods	Parameters	PSNR↑	SSIM↑	VIF↑	LPIPS↓
Transformer	11.32M	27.73	0.7952	0.4511	0.1878
LGFI	12.62M	**27.83**	**0.7961**	**0.4725**	**0.1821**

TABLE III
PERFORMANCE AND COMPUTATIONAL COST COMPARISON
BETWEEN RDFE AND FFN (CelebA, ×8)

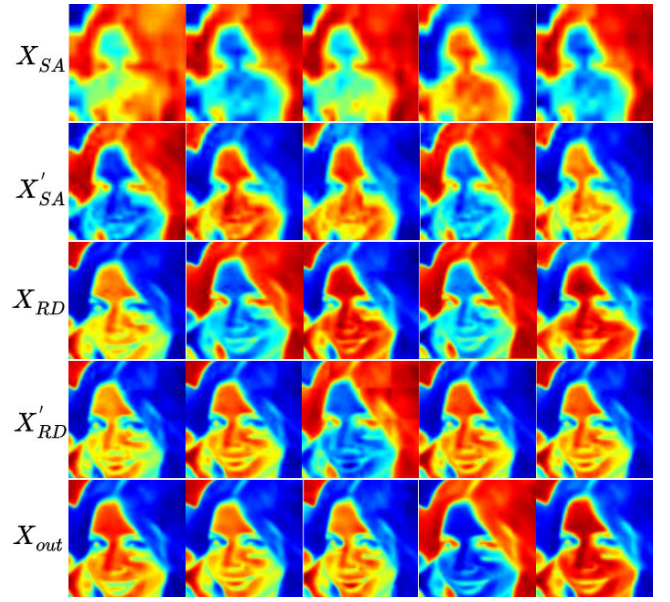| Methods	Parameters	PSNR↑	SSIM↑	VIF↑	LPIPS↓
FFN	12.11M	27.72	0.7931	0.4578	0.1922
RDFE	12.62M	**27.83**	**0.7961**	**0.4725**	**0.1821**

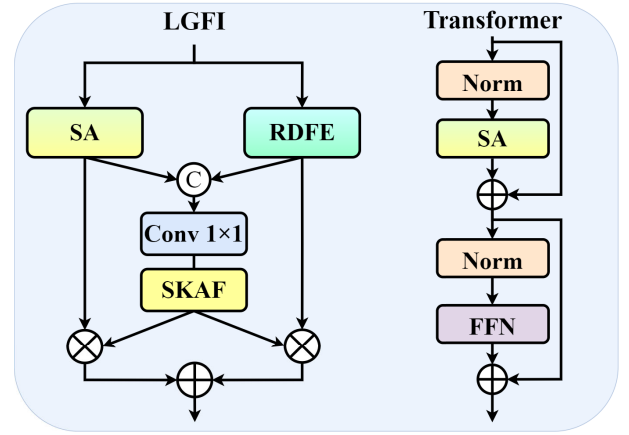Fig. 5. Heatmap outputs of different parts of LGFI.

Fig. 6. Comparison of LGFI and Transformer structures, where SA is self-attention, RDFE and FFN are CNN parts, and SKAF is our feature fusion module.

between local and global facial features. This is because our SKAF can promote interaction between our SA and RDFE, integrating richer information and providing supplementary information for the final FSR.

Furthermore, we provide a visual comparison in Fig. 4, illustrating the impact of removing certain components from LGFI. The reconstructed images exhibit noticeable blurring or artifacts, highlighting the importance of these components. Additionally, Fig. 5 presents heatmap visualizations of the outputs from different components within LGFI, with their corresponding locations in the network indicated in Fig. 2. Specifically, SKAF effectively integrates the global contours captured by SA with the regional features extracted by RDEF. This integration enables the model to focus more on essential facial structures and components while reducing emphasis on less critical details, such as hair. In addition, we quantitatively evaluate the computational efficiency of each module by comparing inference latency and parameter counts in Table I, where the RDEF module has the greatest impact on both inference time and parameter counts. This is attributed to its multibranch fusion strategy and deep refinement operations, which introduce additional computational complexity. However, RDEF delivers substantial performance improvements, which is overall worthwhile.

2) Comparison Between LGFI and Transformer: As shown in Fig. 6, LGFI uses a dual-branch structure to represent the local and global features. In contrast, the traditional Transformer in Restormer [50] uses a serial structure to link local and global features. To verify the effectiveness of LGFI, we replace all LGFIs with Transformers and conduct com-

parative experiments with similar parameters between the two models. From Table II, the network's performance using LGFI is better when the two networks maintain similar parameters. This is because LGFI utilizes the features of both local and global branches for interaction, facilitating the communication of multiscale facial information.

3) Comparison Between RDFE and FFN: The FFN performs independent nonlinear transformations of the inputs at each position to help the Transformer capture local features, but it cannot extract multiscale features, which is not favorable for accurate FSR. In contrast, our RDFE can extract multiscale local features well. To compare RDFE and FFN, we replace RDFE with FFN while keeping the parameters of the two models similar. As shown in Table III, since FFN's ability to capture feature interactions is limited compared to our RDFE, which utilizes multiple branches to capture different receptive field features, our RDFE performs better than FFN with similar computational cost.

TABLE IV
ABLATION STUDY OF OUR RDFE (CelebA, ×8)

Methods	PSNR↑	SSIM↑	VIF↑	LPIPS↓
Single path (3 × 3 dw)	27.73	0.7934	0.4619	0.1915
Single path (5 × 5 dw)	27.71	0.7912	0.4587	0.1944
Single path (7 × 7 dw)	27.72	0.7926	0.4602	0.1922
RDFE w/o AU	27.76	0.7951	0.4673	0.1846
RDFE w/o FRM	27.75	0.7941	0.4643	0.1928
RDFE	**27.83**	**0.7961**	**0.4725**	**0.1821**

TABLE V
ABLATION STUDY OF OUR SKAF (CelebA, ×8)

5 × 5 conv	7 × 7 conv	Avepool	Maxpool	PSNR↑	SSIM↑
×	×	×	×	27.69	0.7919
×	✓	✓	✓	27.76	0.7955
✓	×	✓	✓	27.73	0.7946
×	✓	×	×	27.74	0.7931
×	✓	✓	×	27.79	0.7951
×	✓	×	✓	27.78	0.7946
✓	✓	✓	✓	**27.83**	**0.7961**

4) Effectiveness of RDFE: In RDFE, a three-branch network guided by an attention mechanism is used for deep feature extraction, and the FRM is used to enrich feature representation. To verify the effectiveness of RDFE, we conduct multiple ablation experiments. We designed five improved models. The first model adopts a single branch structure of 3 × 3 depthwise convolution, labeled as "Single path (3 × 3 dw)." The second model adopts a single branch structure of 5 × 5 depthwise convolution, labeled as "Single path (5 × 5 dw)." The third model adopts a single branch structure of 7 × 7 depthwise convolution, labeled as "Single path (7 × 7 dw)." The fourth model removes AUs labeled as "w/o AU." The fifth model removes the FRM, labeled as "w/o FRM". From Table IV, we have the following observations.

1) By comparing the first three rows and the last row of the table, it can be seen that multiscale branching facilitates the model's performance due to its ability to extract face features at different levels.

2) From the comparison between the second and the last rows of the table and the last row, it can be seen that using AUs to guide three-branch feature extraction can enable the model to adaptively allocate weights, enhance the representation of important facial information, and thus improve model performance.

3) From the last two rows of the table, we can conclude that the FRM can further integrate multiscale information, refine multiscale fusion features, and thus improve performance.

5) Effectiveness of SKAF: SKAF is an important component of LGFI, facilitating information exchange between local and global branches. We perform ablation experiments to validate our SKAF module's impact and assess the combined approach's practicality. Since SKAF consists of dual-branch convolutional layers, maximum pooling layers, and average pooling layers, we verify the effectiveness of module components in SKAF. From Table V, we have the following observations.

1) From the last three rows of the table, we find that using a single pooling branch results in reduced performance,

TABLE VI
ABLATION STUDY OF OUR EDFF (CelebA, ×8)

Methods	Baseline			+ Our EDFF		
	Parameters	PSNR↑	SSIM↑	Parameters	PSNR↑	SSIM↑
SPARNet [17]	10.6M	27.73	0.7949	12.1M	**27.85**	**0.7964**
SFMNet [20]	8.6M	27.96	0.7996	10.7M	**28.07**	**0.8011**

while using average pooling alone results in lower performance than using maximum pooling alone. This is because the salient features of the face are the key to facial recovery, with maximum pooling focusing on salient facial feature information. In contrast, average pooling focuses on the overall information of the face. Compared to the third and fifth rows, it can be concluded that using both 5 × 5 and 7 × 7 simultaneously can improve performance and fully utilize key facial information under different receptive fields.

6) Study of EDFF: This section presents a set of experiments to validate the effectiveness of our EDFF, a module tailored for fusing multiscale features. We add EDFF to SPARNet [17], which uses EDFF to connect the encoding and decoding stages in SPARNet and send them to the next decoding stage. Additionally, we add EDFF to SFMNet [20], and the specific operation is the same as in SPARNet. From the results of Table VI, we can see that although the parameters of both models increase slightly, the performance of the models improves, which precisely proves that EDFF is helpful for feature fusion in the encoding and decoding stages.

D. Comparisons With Other Methods

We compare AMINet with existing FSR methods, including SAN [46], RCAN [24], HAN [47], SwinIR [9], FSRNet [12], DICNet [14], FACN [48], SPARNet [17], SISN [18], AD-GNN [49], Restormer-M [50], LAAT [51], ELSFace [52], SFMNet [20], and SPADNet [53].

1) Comparisons on CelebA Dataset: We conduct a quantitative comparison of AMINet against existing FSR methods on the CelebA test set, as detailed in Table VII. Our AMINet outperforms all evaluation metrics, including PSNR, SSIM, LPIPS, and VIF, which fully demonstrates its efficiency. This strongly validates the effectiveness of AMINet. Additionally, the visual comparison in Fig. 7 reveals that previous FSR methods struggled to reproduce facial features like the eyes and mouth accurately. In contrast, AMINet excels at preserving the facial structure and producing more precise results.

2) Comparisons on Helen Dataset: We evaluate our method on the Helen test set to further assess AMINet's versatility. Table VII provides a quantitative comparison of ×8 FSR results about it, where AMINet achieves the better performance. Visual comparisons in Fig. 8 indicate that existing FSR methods struggle to maintain accuracy, leading to blurred shapes and a loss of facial details. In contrast, AMINet successfully preserves facial contours and details, reinforcing its effectiveness and adaptability across different datasets. In addition, we provide a visual comparison of face parsing maps for recovered face images, as shown in Fig. 9, which

TABLE VII
QUANTITATIVE COMPARISONS OF OURS AND EXISTING FSR METHODS FOR ×8 FSR ON CelebA AND Helen TEST SETS

| Methods	CelebA [4]				Helen [40]			
	PSNR↑	SSIM↑	VIF↑	LPIPS↓	PSNR↑	SSIM↑	VIF↑	LPIPS↓
Bicubic	23.61	0.6779	0.1821	0.4899	22.95	0.6762	0.1745	0.4912
SAN [46]	27.43	0.7826	0.4553	0.2080	25.46	0.7360	0.4029	0.3260
RCAN [24]	27.45	0.7824	0.4618	0.2205	25.50	0.7383	0.4049	0.3437
HAN [47]	27.47	0.7838	0.4673	0.2087	25.40	0.7347	0.4074	0.3274
SwinIR [9]	27.88	0.7967	0.4590	0.2001	26.53	0.7856	0.4398	0.2644
FSRNet [12]	27.05	0.7714	0.3852	0.2127	25.45	0.7364	0.3482	0.3090
DICNet [14]	27.42	0.7840	0.4234	0.2129	26.15	0.7717	0.4085	_0.2158_
FACN [48]	27.22	0.7802	0.4366	_0.1828_	25.06	0.7189	0.3702	0.3113
SPARNet [17]	27.73	0.7949	0.4505	0.1995	26.43	0.7839	0.4262	0.2674
SISN [18]	27.91	0.7971	_0.4785_	0.2005	26.64	0.7908	0.4623	0.2571
AD-GNN [49]	27.82	0.7962	0.4470	0.1937	26.57	0.7886	0.4363	0.2432
Restormer-M [50]	27.94	_0.8027_	0.4624	0.1933	_26.91_	_0.8013_	_0.4595_	0.2258
LAAT [51]	27.91	0.7994	0.4624	0.1879	26.89	0.8005	0.4569	0.2255
ELSFace [52]	27.41	0.7922	0.4451	0.1867	26.04	0.7873	0.4193	0.2811
SFMNet [20]	_27.96_	0.7996	0.4644	0.1937	26.86	0.7987	0.4573	0.2322
SPADNet [53]	27.82	0.7966	0.4589	0.1987	26.47	0.7857	0.4295	0.2654
AMINet	**28.26**	**0.8091**	**0.4893**	**0.1755**	**27.01**	**0.8042**	**0.4694**	**0.2067**
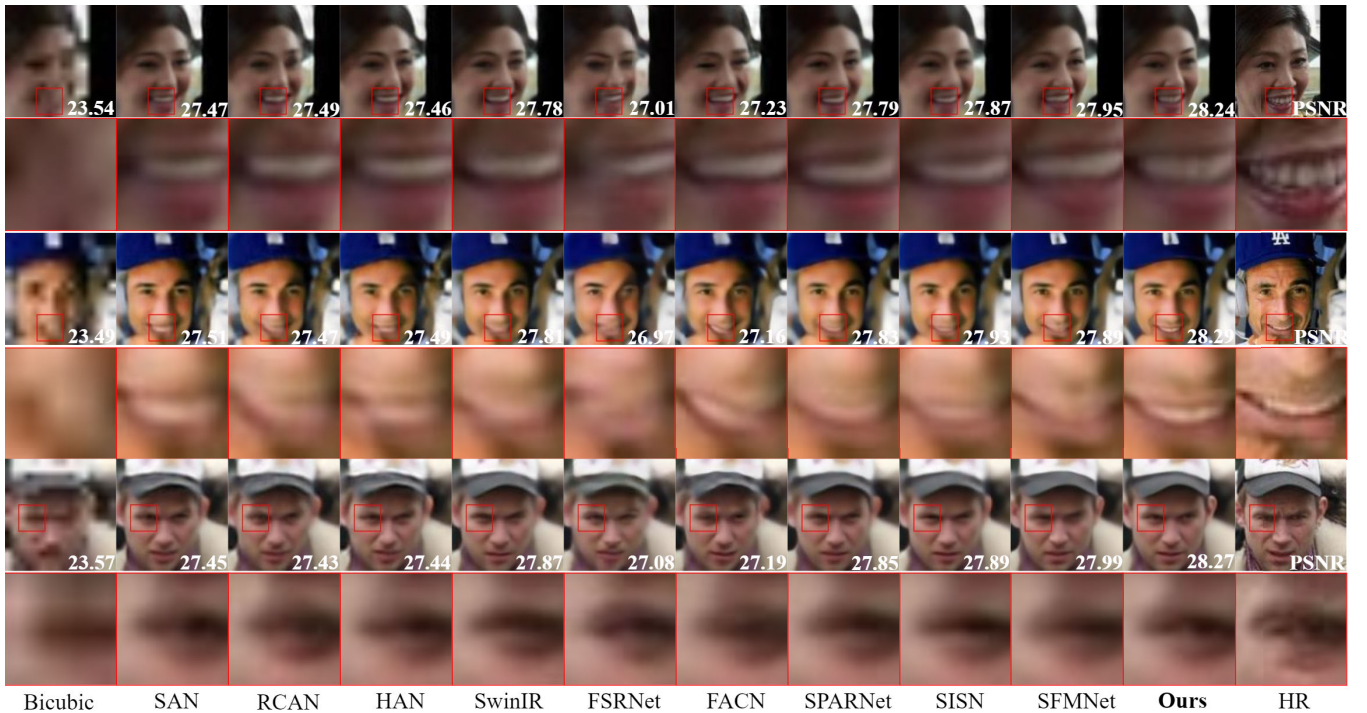
Fig. 7. Visual comparisons for ×8 FSR on CelebA [4] test set. Our method can recover accurate face images.

clearly shows that our AMINet facilitates downstream tasks such as face parsing maps segmentation.

3) Comparisons With GAN-Based Methods: We present AMIGAN as a novel approach to enhance the visual fidelity of image restoration. To demonstrate its effectiveness, we compare AMIGAN with existing GAN-based methods, including FSRGAN [12], DICGAN [14], SPARGAN [17], and SFM-GAN [20]. In addition to conventional metrics, we adopt FID [45] for quantitative evaluation. Results on the Helen dataset (Table VIII) show that AMIGAN consistently outperforms prior methods. Visual comparisons in Fig. 10 further highlight AMIGAN's superior ability to restore fine facial structures and texture details, particularly around the mouth and nose, delivering clearer and more realistic reconstructions with fewer artifacts.

4) Comparisons on Real-World Surveillance Faces: All the above comparisons are tested on synthetic test sets, which fail to simulate real-world scenarios accurately. To further evaluate our model's performance in real-world conditions, we also conduct experiments using low-quality face images from the SCface dataset [41]. As shown in Fig. 11, we compare the reconstruction results. From this figure, we find that the reconstruction results of face prior-based methods are not satisfactory. The challenge lies in accurately estimating priors

Fig. 8. Visual comparisons for ×8 FSR on Helen [40] test set. Our method can recover accurate face images.
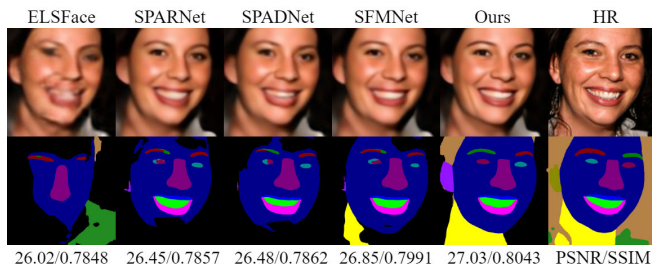

Fig. 9. Comparisons of face parsing on the Helen test set.

Fig. 10. Visual comparisons of existing GAN-based FSR methods on the Helen test set. Our AMIGAN can reconstruct high-quality face images with clear facial components.

TABLE VIII
QUANTITATIVE COMPARISON OF OURS WITH OTHER GAN-BASED METHODS (HELEN, ×8)

Methods	PSNR↑	SSIM↑	VIF↑	FID↓
FSRGAN [12]	25.02	0.7279	0.3400	146.55
DICGAN [14]	25.59	0.7398	0.3925	144.25
SPARGAN [17]	25.86	0.7518	0.3932	149.54
SFMGAN [20]	25.96	0.7618	0.4019	141.23
AMIGAN (Ours)	**26.35**	**0.7769**	**0.4101**	**122.43**

TABLE IX
COMPARISONS OF COSINE SIMILARITY ON ×8 SCFACE

| Methods | Cosine Similarity↑ | | | |
	Case 1	Case 2	Case 3	Case 4
SAN [46]	0.8133	0.8145	0.8244	0.8192
RCAN [24]	0.8196	0.8214	0.8199	0.8201
FSRNet [12]	0.8032	0.7982	0.8105	0.8087
FACN [48]	0.8002	0.7989	0.8115	0.8014
SPARNet [17]	0.8215	0.8209	0.8244	0.8195
SISN [18]	0.8345	0.8378	0.8373	0.8391
LAAT [51]	0.8501	0.8497	0.8456	0.8472
ELSFace [52]	0.8165	0.8124	0.8136	0.8148
SFMNet [20]	0.8411	0.8456	0.8429	0.8402
AMINet	**0.8533**	**0.8577**	**0.8601**	**0.8501**

details and faithful face structures. As shown in Table IX, we also provide a quantitative comparison of cosine similarity using the above methods. This result fully demonstrates our method's effectiveness in real scenarios.

E. Model Complexity and Convergence Analysis

In addition to the performance indicators mentioned earlier, the number of model parameters, inference time, and

from real-world LR facial images. Incorrect prior information can lead to misleading guidance during the reconstruction process. In contrast, our AMINet can restore clearer face

Fig. 11. Visual comparisons for ×8 FSR on SCface [41] test set. Our method recovers clearer face images.

TABLE X
COMPARISONS OF MODEL COMPLEXITY ON ×8 CelebA

Methods	PSNR↑	Params↓	MACs↓	Speed↓
FSRNet [12]	27.05	27.5M	40.7G	89ms
SPARNet [17]	27.73	16.6M	7.1G	40ms
DICNet [14]	27.42	22.8M	35.5G	121ms
AD-GNN [49]	27.82	15.8M	15.0G	108ms
CTCNet [5]	28.37	22.4M	47.2G	106ms
SCTANet [10]	28.26	27.7M	10.4G	58ms
LAAT [51]	27.91	22.4M	8.9G	36ms
ELSFace [52]	27.41	6.8M	10.6G	32ms
SFMNet [20]	27.96	8.6M	30.6G	48ms
AMINet-S	27.93	8.2M	9.4G	25ms
AMINet	28.26	12.62M	15.6G	35ms

Fig. 12. Model convergence and visualization analysis.

computational complexity are crucial factors in evaluating performance. As shown in Table X, we have selected some models for comparison. Meanwhile, as shown in Fig. 1, we compare our model with existing ones in terms of parameters, PSNR values, and inference speed. We can see that AMINet can have faster inference time, smaller parameter count, and computational complexity with similar performance to SCTANet. In addition, we provide a small parameter version called "AMINet-S." AMINet-S performs similarly to SFMNet in terms of parameter quantity, while AMINet has more advantages in computational complexity and inference time. This is thanks to AMINet achieving high efficiency by promoting extensive multiscale feature exchange within the network. This design enables features from different receptive fields to interact effectively, allowing the model to select the most relevant information for facial reconstruction adaptively. Therefore, AMINet maintains strong FSR performance while balancing parameters and computational costs.

We also provide the training loss curve in Fig. 12, which shows a consistent decrease as iterations progress, indicating stable optimization and convergence. For clarity, the loss values are scaled by a factor of 100. We also visualize the intermediate results at different training stages, illustrating the clear improvement from early to late iterations as the model progressively learns more discriminative features. These observations collectively confirm the effectiveness of the training strategy and its stable convergence properties.

V. DISCUSSION AND FUTURE WORKS

Although our AMINet performs well in FSR, it still has certain limitations. The model's robustness to extreme poses at very low resolutions, such as side profiles, and its ability to handle challenging lighting conditions, including low-light and overexposed environments, require further improvement. Moreover, while our method significantly reduces computational cost compared to existing methods, it remains insufficient for deployment on mobile devices.

Future work will focus on enhancing AMINet's robustness to extreme poses and challenging lighting conditions while accelerating inference. This includes integrating pose-invariant feature learning through attention-based mechanisms or 3D-aware priors for better side-profile restoration and developing adaptive illumination correction using physics-based relighting models or low-light enhancement. Additionally, optimizing the model with lightweight architectures, quantization, and efficient inference strategies will enable faster inference while maintaining high-quality restoration.

VI. CONCLUSION

We propose an AMINet for FSR. The core component, LGFI, facilitates effective interaction between global features from SA and local features extracted by the proposed RDFE module. To enrich local representations, RDFE employs multiscale depthwise separable convolutions combined with attention for feature extraction and refinement. Moreover, an adaptive kernel selection mechanism further promotes multiscale feature fusion. Extensive experiments on synthetic and real-world datasets demonstrate that our design substantially enhances cross-scale feature interaction, enabling our method to surpass existing approaches in reconstruction quality, model size, and inference efficiency.

REFERENCES

[1] L. Liu, R. Lan, and Y. Wang, "Discriminative face hallucination via locality-constrained and category embedding representation," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 51, no. 12, pp. 7314–7325, Dec. 2021.

[2] D. Mamieva, A. B. Abdusalomov, M. Mukhiddinov, and T. K. Whangbo, "Improved face detection method via learning small faces on hard images based on a deep learning approach," *Sensors*, vol. 23, no. 1, p. 502, Jan. 2023.

[3] G. Hu et al., "When face recognition meets with deep learning: An evaluation of convolutional neural networks for face recognition," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop (ICCVW)*, Dec. 2015, pp. 384–392.

[4] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3730–3738.

[5] G. Gao, Z. Xu, J. Li, J. Yang, T. Zeng, and G.-J. Qi, "CTCNet: A CNN-transformer cooperation network for face image super-resolution," *IEEE Trans. Image Process.*, vol. 32, pp. 1978–1991, 2023.

[6] E. Zhou, H. Fan, Z. Cao, Y. Jiang, and Q. Yin, "Learning face hallucination in the wild," in *Proc. AAAI*, Jan. 2015, pp. 3871–3877.

[7] J. Shi et al., "IDPT: Interconnected dual pyramid transformer for face super-resolution," in *Proc. 31st Int. Joint Conf. Artif. Intell.*, Jul. 2022, pp. 1306–1312.

[8] Y. Wang, T. Lu, Y. Zhang, Z. Wang, J. Jiang, and Z. Xiong, "FaceFormer: Aggregating global and local representation for face hallucination," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 6, pp. 2533–2545, Jun. 2023.

[9] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "SwinIR: Image restoration using Swin transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 1833–1844.

[10] Q. Bao, Y. Liu, B. Gang, W. Yang, and Q. Liao, "SCTANet: A spatial attention-guided CNN-transformer aggregation network for deep face image super-resolution," *IEEE Trans. Multimedia*, vol. 25, pp. 8554–8565, 2023.

[11] W. Li et al., "Survey on deep face restoration: From non-blind to blind and beyond," 2023, *arXiv:2309.15490*.

[12] Y. Chen, Y. Tai, X. Liu, C. Shen, and J. Yang, "FSRNet: End-to-end learning face super-resolution with facial priors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2492–2501.

[13] D. Kim, M. Kim, G. Kwon, and D.-S. Kim, "Progressive face super-resolution via attention to facial landmark," 2019, *arXiv:1908.08239*.

[14] C. Ma, Z. Jiang, Y. Rao, J. Lu, and J. Zhou, "Deep face super-resolution with iterative collaboration between attentive recovery and landmark estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5569–5578.

[15] X. Hu et al., "Face super-resolution guided by 3D facial priors," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 763–780.

[16] M. Zhang and Q. Ling, "Supervised pixel-wise GAN for face super-resolution," *IEEE Trans. Multimedia*, vol. 23, pp. 1938–1950, 2021.

[17] C. Chen, D. Gong, H. Wang, Z. Li, and K.-Y.-K. Wong, "Learning spatial attention for face super-resolution," *IEEE Trans. Image Process.*, vol. 30, pp. 1219–1231, 2021.

[18] T. Lu et al., "Face Hallucination via split-attention in split-attention network," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 5501–5509.

[19] Q. Bao, R. Zhu, B. Gang, P. Zhao, W. Yang, and Q. Liao, "Distilling resolution-robust identity knowledge for texture-enhanced face hallucination," in *Proc. 30th ACM Int. Conf. Multimedia*, Oct. 2022, pp. 6727–6736.

[20] C. Wang, J. Jiang, Z. Zhong, and X. Liu, "Spatial-frequency mutual learning for face super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 22356–22366.

[21] J. Shi et al., "Exploiting multi-scale parallel self-attention and local variation via dual-branch transformer-CNN structure for face super-resolution," *IEEE Trans. Multimedia*, vol. 26, pp. 2608–2620, 2024.

[22] W. Li et al., "Efficient face super-resolution via wavelet-based feature enhancement network," in *Proc. 32nd ACM Int. Conf. Multimedia*, Oct. 2024, pp. 4515–4523.

[23] J. Li et al., "A systematic survey of deep learning-based single-image super-resolution," *ACM Comput. Surv.*, vol. 56, no. 10, pp. 1–40, Oct. 2024.

[24] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 286–301.

[25] J. Xin, N. Wang, X. Gao, and J. Li, "Residual attribute attention network for face image super-resolution," in *Proc. AAAI*, 2019, vol. 33, no. 1, pp. 9054–9061.

[26] G. Gao, W. Li, J. Li, F. Wu, H. Lu, and Y. Yu, "Feature distillation interaction weighting network for lightweight image super-resolution," in *Proc. AAAI Conf. Artif. Intell.*, 2022, vol. 36, no. 1, pp. 661–669.

[27] Y. Wang, Y. Li, G. Wang, and X. Liu, "Multi-scale attention network for single image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2024, pp. 5950–5960.

[28] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[29] G. Gao, Z. Wang, J. Li, W. Li, Y. Yu, and T. Zeng, "Lightweight bimodal network for single-image super-resolution via symmetric CNN and recursive transformer," in *Proc. 31st Int. Joint Conf. Artif. Intell.*, Jul. 2022, pp. 913–919.

[30] W. Li et al., "Cross-receptive focused inference network for lightweight image super-resolution," *IEEE Trans. Multimedia*, vol. 26, pp. 864–877, 2024.

[31] W. Li et al., "Efficient image super-resolution with feature interaction weighted hybrid network," *IEEE Trans. Multimedia*, vol. 27, pp. 2256–2267, 2025.

[32] K. Zeng, Z. Wang, T. Lu, J. Chen, J. Wang, and Z. Xiong, "Self-attention learning network for face super-resolution," *Neural Netw.*, vol. 160, pp. 164–174, Mar. 2023.

[33] Y. Yang and Y. Qi, "Image super-resolution via channel attention and spatial graph convolutional network," *Pattern Recognit.*, vol. 112, Apr. 2021, Art. no. 107798.

[34] Z. Zhang and C. Qi, "Feature maps need more attention: A spatial-channel mutual attention-guided transformer network for face super-resolution," *Appl. Sci.*, vol. 14, no. 10, p. 4066, May 2024.

[35] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 510–519.

[36] Y. Li, Q. Hou, Z. Zheng, M.-M. Cheng, J. Yang, and X. Li, "Large selective kernel network for remote sensing object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 16794–16805.

[37] C. Ledig et al., "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4681–4690.

[38] X. Wang et al., "ESRGAN: Enhanced super-resolution generative adversarial networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV) Workshops*, 2018, pp. 1–16.

[39] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[40] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang, "Interactive facial feature localization," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 679–692.

[41] M. Grgic, K. Delac, and S. Grgic, "SCface—Surveillance cameras face database," *Multimedia Tools Appl.*, vol. 51, no. 3, pp. 863–879, Feb. 2011.

[42] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[43] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 586–595.

[44] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 430–444, Feb. 2006.

[45] A. Obukhov and M. Krasnyanskiy, "Quality assessment method for GAN based on modified metrics inception score and Fréchet inception distance," in *Proc. CoMeSySo*, 2020, pp. 102–114.

[46] T. Dai, J. Cai, Y. Zhang, S.-T. Xia, and L. Zhang, "Second-order attention network for single image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11065–11074.

[47] B. Niu et al., "Single image super-resolution via a holistic attention network," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 191–207.

[48] J. Xin, N. Wang, X. Jiang, J. Li, X. Gao, and Z. Li, "Facial attribute capsules for noise face super resolution," in *Proc. AAAI*, 2020, vol. 34, no. 7, pp. 12476–12483.

[49] Q. Bao, B. Gang, W. Yang, J. Zhou, and Q. Liao, "Attention-driven graph neural network for deep face super-resolution," *IEEE Trans. Image Process.*, vol. 31, pp. 6455–6470, 2022.

[50] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Restormer: Efficient transformer for high-resolution image restoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5728–5739.

[51] G. Li, J. Shi, Y. Zong, F. Wang, T. Wang, and Y. Gong, "Learning attention from attention: Efficient self-refinement transformer for face super-resolution," in *Proc. 32nd Int. Joint Conf. Artif. Intell.*, Aug. 2023, pp. 1035–1043.

[52] H. Qi, Y. Qiu, X. Luo, and Z. Jin, "An efficient latent style guided transformer-CNN framework for face super-resolution," *IEEE Trans. Multimedia*, vol. 26, pp. 1589–1599, 2024.

[53] C. Wang, J. Jiang, K. Jiang, and X. Liu, "SPADNet: Structure prior-aware dynamic network for face super-resolution," *IEEE Trans. Biometrics, Behav., Identity Sci.*, vol. 6, no. 3, pp. 326–340, Jul. 2024.
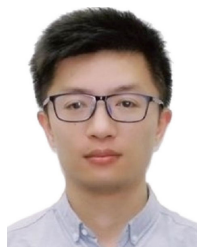
Guangwei Gao (Senior Member, IEEE) received the Ph.D. degree in pattern recognition and intelligent systems from Nanjing University of Science and Technology (NJUST), Nanjing, China, in 2014.

He was a Visiting Student with the Department of Computing, The Hong Kong Polytechnic University, Hong Kong, in 2011 and 2013, respectively. From 2019 to 2021, he was a Project Researcher with the National Institute of Informatics, Chiyoda, Japan. He is currently a Professor at Nanjing University of Posts and Telecommunications, Nanjing. His research interests include pattern recognition and computer vision.

Huimin Lu (Senior Member, IEEE) received the Ph.D. degree in electrical engineering from the Kyushu Institute of Technology, Kitakyushu, Japan, in 2014.

From 2013 to 2016, he was a JSPS Research Fellow with the Kyushu Institute of Technology. From 2016 to 2024, he was an Associate Professor with the Kyushu Institute of Technology and an Excellent Young Researcher of Ministry of Education, Culture, Sports, Science and Technology. He is currently a Professor with Southeast University, Nanjing, China. His research interests include artificial intelligence, computer vision, and robotics.

Xujie Wan received the B.S. degree in automation from the School of Microelectronics and Control Engineering, Changzhou University, Changzhou, China, in 2022, and the M.S. degree in control science and engineering from the College of Automation, Nanjing University of Posts and Telecommunications, Nanjing, China, in 2025.

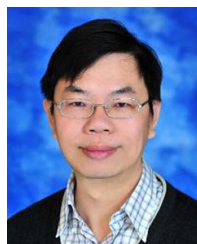His research interest includes image super-resolution.

Jian Yang (Member, IEEE) received the Ph.D. degree in pattern recognition and intelligent systems from Nanjing University of Science and Technology (NJUST), Nanjing, China, in 2002.

From 2006 to 2007, he was a Post-Doctoral Fellow with the Department of Computer Science, New Jersey Institute of Technology, Newark, NJ, USA. He is currently a Professor with the School of Computer Science and Technology, NJUST. He is the author of more than 400 scientific papers in pattern recognition and computer vision. His research interests include pattern recognition, computer vision, and machine learning.

Dr. Yang is a fellow of IAPR. He is/was an Associate Editor of *Pattern Recognition* and IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS.

Wenjie Li received the M.S. degree in control science and engineering from the College of Automation, Nanjing University of Posts and Telecommunications, Nanjing, China, in 2023. He is currently pursuing the Ph.D. degree in artificial intelligence with the School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing, China.

His research interests include image restoration.

Chia-Wen Lin (Fellow, IEEE) received the Ph.D. degree in electrical engineering from National Tsing Hua University (NTHU), Hsinchu, Taiwan, in 2000.

He was with the Department of Computer Science and Information Engineering, National Chung Cheng University, Chiayi, Taiwan, from 2000 to 2007. He is currently a Distinguished Professor with the Department of Electrical Engineering and the Institute of Communications Engineering, NTHU. He is also the Deputy Director of the AI Research Center, NTHU. His research interests include image and video processing, computer vision, and video networking.

Dr. Lin is currently serving as an Associate Editor-in-Chief of IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY.