

CTCNet: A CNN-Transformer Cooperation Network for Face Image Super-Resolution

Guangwei Gao, *Member, IEEE*, Zixiang Xu, Juncheng Li, Jian Yang, *Member, IEEE*, Tieyong Zeng, *Member, IEEE*, and Guo-Jun Qi, *Fellow, IEEE*

Abstract—Recently, deep convolution neural networks (CNNs) steered face super-resolution methods have achieved great progress in restoring degraded facial details by jointly training with facial priors. However, these methods have some obvious limitations. On the one hand, multi-task joint learning requires additional marking on the dataset, and the introduced prior network will significantly increase the computational cost of the model. On the other hand, the limited receptive field of CNN will reduce the fidelity and naturalness of the reconstructed facial images, resulting in suboptimal reconstructed images. In this work, we propose an efficient CNN-Transformer Cooperation Network (CTCNet) for face super-resolution tasks, which uses the multi-scale connected encoder-decoder architecture as the backbone. Specifically, we first devise a novel Local-Global Feature Cooperation Module (LGCM), which is composed of a Facial Structure Attention Unit (FSAU) and a Transformer block, to promote the consistency of local facial detail and global facial structure restoration simultaneously. Then, we design an efficient Local Feature Refinement Module (LFRM) to enhance the local facial structure information. Finally, to further improve the restoration of fine facial details, we present a Multi-scale Feature Fusion Unit (MFFU) to adaptively fuse the features from different stages in the encoder procedure. Comprehensive evaluations on various datasets have assessed that the proposed CTCNet can outperform other state-of-the-art methods significantly.

Index Terms—Face super-resolution, Transformer, CNN-Transformer cooperation, GAN.

I. INTRODUCTION

FACE super-resolution (FSR), a.k.a. face hallucination, refers to a technology for obtaining high-resolution (HR) face images from input low-resolution (LR) face images. In practical application scenarios, due to the inherent differences in the hardware configuration, placement position, and shooting angle of the image capture device, the quality of the face images obtained by shooting is inevitably poor. Lower quality images seriously affect the downstream tasks such as face analysis and face recognition. Different from general image SR, the core goal of FSR is to reconstruct as much as

G. Gao is with the Institute of Advanced Technology, Nanjing University of Posts and Telecommunications, Nanjing, China (e-mail: csggao@gmail.com).

Z. Xu is with the College of Automation, Nanjing University of Posts and Telecommunications, Nanjing, China (e-mail: wszixiangxu@gmail.com).

J. Li and T. Zeng are with the Center for Mathematical Artificial Intelligence, Department of Mathematics, The Chinese University of Hong Kong, Hong Kong, China (e-mail: cvjunchengli@gmail.com, zeng@math.cuhk.edu.hk).

J. Yang is with the School of Computer Science and Technology, Nanjing University of Science and Technology, Nanjing, China (e-mail: csjyang@njjust.edu.cn).

G.-J. Qi is with the Department of Computer Science, University of Central Florida, Orlando, USA (e-mail: guojunq@gmail.com).

possible the facial structure information (i.e., shapes of face components and face outline) that is missing in the degraded observation. Although these structures only occupy a small part of the face, they are the key to distinguishing different faces. Compared with other areas in a face image, the facial feature and contours of a person are usually more difficult to restore since they often span a large area and require more global information.

Most of the previous FSR algorithms [1]–[3] mainly adopted the strategy of successive multi-task training. These methods used the facial landmark heatmaps or parsing maps to participate in the formal training to constrain the performance of the FSR reconstruction network. However, they also need extra labeled data to achieve the goal. Besides, in the previous FSR methods [4], [5], the encoding and decoding parts are connected in series. This kind of connection cannot fully utilize the low-level features, and the low-level features also cannot thoroughly guide the learning of the high-level features, resulting in the unsatisfied performance in FSR task. In addition, many FSR networks [6]–[10] have been built using Convolution Neural Networks (CNNs) due to the powerful local modeling capabilities of CNN to predict fine-grained facial details. However, the human face usually has a fixed geometric features structure [11]–[13]. Therefore, if only focusing on the extraction of the local information while ignoring the relationship between them (global information), it will inevitably affect the restoration of the global facial structure, leading to blurry effects in the generated faces.

As we know, local methods (such as CNN-based methods) mainly focus on the local facial details, while global methods (such as Transformer-based methods) usually capture the global facial structures. How to collaboratively make full use of the local and global features, and how to efficiently aggregate the multi-scale abundant features is important. To achieve this, in this work, we propose an efficient CNN-Transformer Cooperation Network (CTCNet) for FSR. Like most of previous FSR models, our CTCNet also uses an encoder-decoder structure. Specifically, in the encoder and decoder branches, the specially designed Local-Global Feature Cooperation Module (LGCM) are used for feature extraction. LGCM is composed of a Facial Structure Attention Unit (FSAU) and a Transformer block. Among them, FSAU is specially designed to extract key face components information and Transformer blocks are introduced to explore the long-distance visual relation modeling. The combination of FASU and Transformer block can simultaneously capture local facial texture details and global facial structures. Meanwhile, instead

of using the successive connections, we design a Multi-scale Feature Fusion Unit (MFFU) to flexibly fuse the features from different stages of the network. In addition, we use the Feature Refinement Module (FRMs) between the encoder and decoder branches to further enhance the extracted features, thus further improving the performance of CTCNet. In summary, the main contributions of this work are as follows

- We propose an efficient Local-Global Feature Cooperation Module (LGCM). The combination of CNN and Transformer structure make it can simultaneously capture local facial texture details and global facial structures, which benefit for high-quality face super-resolution image reconstruction.
- We propose a elaborately designed Multi-scale Feature Fusion Unit (MFFU) to fuse the dense features from different scales and depths of the network. This operation ensures that our model can obtain rich features to better reconstruct high-quality images.
- We propose a Feature Refinement Module (FRM) to strengthen the different face structure information and further enhance the extracted features.
- We devise a novel CNN-Transformer Cooperation Network (CTCNet) for face super-resolution. Without relying on any prior information, our CTCNet gains state-of-the-art performance in terms of various kinds of metrics. Meanwhile, the extended model, CNN-Transformer Cooperation Generative Adversarial Network (CTCGAN) can generate more realistic face images.

II. RELATED WORK

A. Face Super-Resolution

Recently, due to the powerful feature representation capabilities of deep convolution neural networks (CNNs), significant progress has been made in image super-resolution [14], [15], which also greatly promoted the progress of Face Super-Resolution (FSR). For example, Yu et al. [6] introduced an ultra-resolution by the discriminative generative network to solve the problem of blurry face image reconstruction. Zhang et al. [7] proposed a super-identity CNN, which introduced super-identity loss to assist the network in generating super-resolution face images with more accurate identity information. Huang et al. [16] turned the research core to the wavelet domain and proposed a WaveletSRNet capable of predicting the wavelet coefficients of HR images. Lu et al. [17] devised a split-attention in split-attention network based on their designed external-internal split attention group for clear facial images reconstruction.

In addition, some scholars have considered the particularity of the FSR task and proposed some FSR models guided by facial priors (e.g., face parsing maps and landmarks). For instance, Yu et al. [18] directly aggregate the extracted features with facial component heatmaps in the middle of the network and achieve good results. Chen et al. [4] proposed the first end-to-end face super-resolution convolution network, which utilized the facial parsing maps and landmark heatmaps to guide the super-resolution process. Kim et al. [9] also used face key point maps and face heatmaps to construct facial

attention loss and used them to train a progressive generator. To tackle face images that exhibit large pose variations, Hu et al. [2] introduced the 3D facial priors to better capture the sharp facial structures. Ma et al. [1] designed an iterative collaboration method that focuses on facial recovery and landmark estimation respectively. Li et al. [19] incorporated face attributes and face boundaries in a successive manner together with self-attentive structure enhancement to super-resolved tiny LR face images. Although these models have achieved promising results, they requires additional marking on the dataset, and the accuracy of priors will greatly limit the accuracy of the reconstruction results.

B. Attention Mechanism

In the past few decades, the attention mechanism has made prominent breakthroughs in various visual image understanding tasks, such as image classification [20]–[22], image restoration [10], [23]–[25], etc. The attention mechanism can give more attention to key features, which is beneficial to feature learning and model training. Zhang et al. [23] proved that by considering the interdependence between channels and adjusting the channel attention mechanism, high-quality images could be reconstructed. Chen et al. [10] presented a facial spatial attention mechanism, which uses the hourglass structure to form an attention mechanism, thus the convolutional layers can adaptively extract local features related to critical facial structures.

Recently, Transformer [26], [27] are also widely used in computer vision tasks, such as image recognition [28], [29], object detection [30], [31], and image restoration [32]–[35]. The key idea of Transformer is the self-attention mechanism that can capture the long-range correlation between words/pixels. Although pure Transformers have great advantages in distilling the global representation of images, only depending on image-level self-attention will still cause the loss of local fine-grained details. Therefore, how to effectively combine the global information and local features of the image is important for high-quality image reconstruction, which is also the goal of this work.

III. CNN-TRANSFORMER COOPERATION NETWORK

In this section, we first depict the overall architecture of the proposed CNN-Transformer Cooperation Network (CTCNet). Then, we introduce each module in the network in detail. Finally, we introduce related loss functions for supervised CTCGAN training.

A. Overview of CTCNet

As shown in Fig. 1, the proposed CTCNet is a U-shaped symmetrical hierarchical network with three stages: encoding stage, bottleneck stage, and decoding stage. Among them, the encoding stage is designed to extract local and global features with different scales, and the decoding stage is designed for feature fusion and image reconstruction. Meanwhile, the multi-scale connections are used between the encoding stage and the decoding stage to achieve feature aggregation. To better

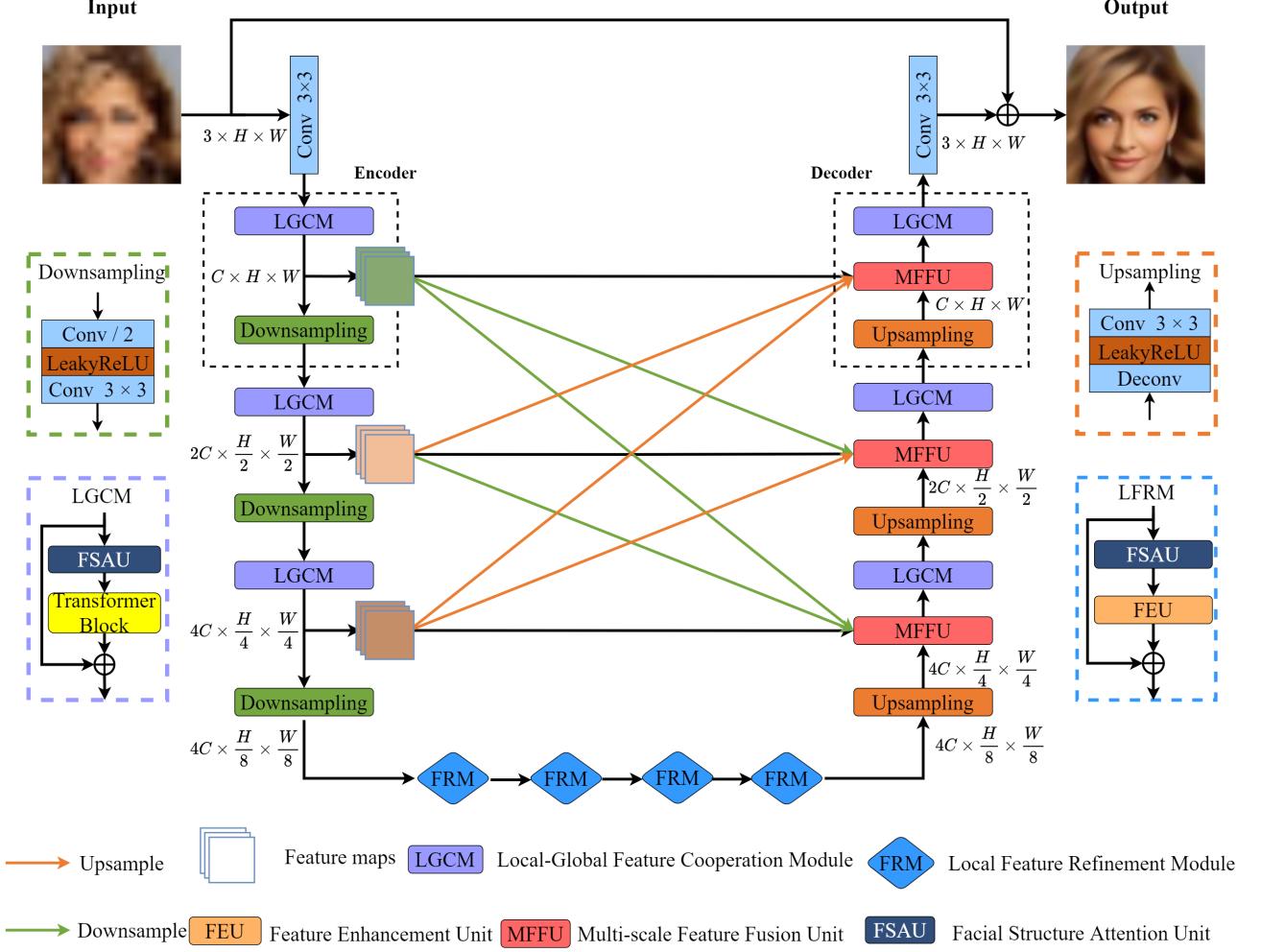


Fig. 1: The complete structure of the proposed CNN-Transformer Cooperation Network (CTCNet).

demonstrate the model, we define I_{LR} , I_{SR} , and I_{HR} as the LR input image, the recovered SR image, and the ground-truth HR image, respectively.

1) *Encoding Stage:* As we mentioned above, the encoding stage is designed for feature extraction. Therefore, give a degraded image I_{LR} as the input, we first apply a 3×3 convolution layer to extract the shallow features. After that, the extract features are passed through 3 encoding stages. Each encoding stage includes one specially designed Local-Global Feature Cooperation Module (LGCM) and one downsampling block. Among them, LGCM consists of a Facial Structure Attention Unit (FSAU) and a Transformer block. The downsampling block consists of a 3×3 convolutional layer with stride 2, a LeakyReLU activation function, and a 3×3 convolution with stride 1, in which the first convolution uses stride 2 to extract feature information and reduce the size simultaneously. Therefore, after each encoding stage, the size of the output feature maps will be halved, while the number of output channels will be doubled. For instance, given the input feature maps $I_{LR} \in \mathbb{R}^{C \times H \times W}$, the i -th stage of the encoder produces the feature maps $I_{en}^i \in \mathbb{R}^{2^i C \times \frac{H}{2^i} \times \frac{W}{2^i}}$.

2) *Bottleneck Stage:* There exist a bottleneck stage among the encoding and decoding stages. At this stage, all encoded

features will be converged here. In order to make these features better utilized in the decode stage, we introduce Feature Refinement Module (FRM) to further refine and enhance the encoded features. With the help of FRMs, our model can focus on more facial structures and continuously strengthen different face structure information.

3) *Decoding Stage:* In the decoding stage, we focus on feature utilization and aim to reconstruct high-quality face images. To achieve this, we introduced a novel module, called Multi-scale Feature Fusion Unit (MFFU). Specifically, the decoder takes the latent features of LR image as inputs and progressively fuse them through MFFUs to reconstruct the SR representations. As shown in Fig. 1, each decoder consists of an upsampling block, a MFFU, and a LGCM. Among them, the upsampling block consists of a 6×6 transposed convolutional layer with stride 2, a LeakyReLU activation function, and a 3×3 convolution with stride 1, in which the transposed convolutional layer uses stride 2 to extract feature information and increase the size of features simultaneously. Therefore, each decoder halves the number of the output feature channels while doubles the size of the output feature maps. It is worth mentioning that in MFFU, it will simultaneously fuses features with different scales extracted

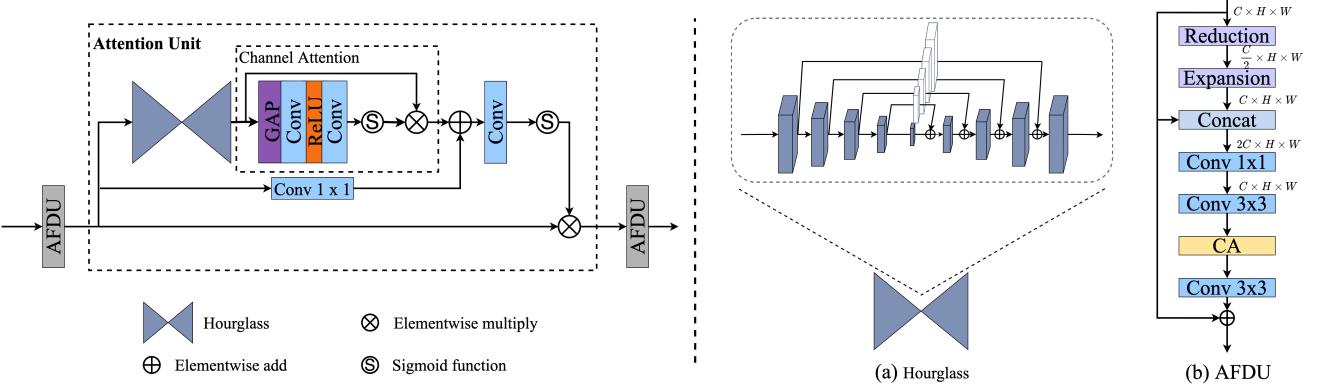


Fig. 2: The architecture of the proposed Facial Structure Attention Unit (FSAU).

in the encoding stage. Therefore, all local and global features with different scale can be fully used to reconstruct high-quality face images. At the end of the decoding stage, we use a 3×3 convolutional layer to convert the learned features into the final SR features I_{Out} .

Finally, the high-quality SR face image is obtained by $I_{SR} = I_{LR} + I_{Out}$. Given a training dataset $\{I_{LR}^i, I_{HR}^i\}_{i=1}^N$, we optimize our CTCNet by minimizing the following pixel-level loss function:

$$\mathcal{L}(\Theta) = \frac{1}{N} \sum_{i=1}^N \|F_{CTCNet}(I_{LR}^i, \Theta) - I_{HR}^i\|_1, \quad (1)$$

where N denotes the number of the training images. I_{LR}^i and I_{HR}^i are the LR image and the ground-truth HR image of the i -th image, respectively. Meanwhile, $F_{CTCNet}(\cdot)$ and Θ denote the CTCNet and its network parameters, respectively.

B. Local-Global Feature Cooperation Module (LGCM)

As one of the most important module in CTCNet, LGCM is designed for local and global feature extraction. As shown in Fig. 1, LGCM consists of a Facial Structure Attention Unit (FSAU) and a Transformer Block, which are used for local and global feature extraction, respectively.

1) *Facial Structure Attention Unit (FSAU):* In FSR, the main challenge is how to extract the key facial features (such as eyes, eyebrows, and mouth), and make the network pay more attention to these features. To achieve this, we propose the Facial Structure Attention Unit (FSAU) to make our model extract as much as possible useful information for better detail restoration. As shown in Fig. 2, FSAU mainly consists of one Attention Unit and two Adaptive Feature Distillation Units (AFDUs). In the Attention Unit, we use channel attention nested in spatial attention to better extract spatial features and promote channel information interaction. This is because combining the two attention mechanisms can promote the representation power of the extracted features. Specifically, we first adopt the hourglass structure to capture facial landmark features at multiple scales since the hourglass structure has been successfully used in human pose estimation and FSR tasks [36], [37]. After that, in order to make the module focus on the features of the critical facial components, we

introduce the channel attention (CA) mechanism [23] to pay more attention to the channels containing landmark features. Then, we use an additional 3×3 convolutional layer and Sigmoid function to generate the spatial attention maps of the key components of the face. Finally, to alleviate the problem of vanishing gradients, we also add the residual connection between the input of the hourglass and the output of CA.

In addition, we also introduce Adaptive Feature Distillation Units (AFDUs) at the beginning and end of the attention unit for local feature extraction. As shown in Fig. 2 (b), to save memory and the number of parameters, we first use Reduction operation to halve the number of the the feature maps and then restore it by the Expansion operation. Among them, Reduction and Expansion operations are both composed of a 3×3 convolutional layer. Meanwhile, we apply the concatenation operation to aggregate the input of Reduction and the output of Expansion along the channel dimension, followed by a 1×1 convolutional layer and a 3×3 convolutional layer. The 1×1 convolution is used to fully utilize the hierarchical features, while the 3×3 convolution dedicated to reducing the number of feature maps. After that, a CA module is employed to highlight the channels with high activated values, and a 3×3 convolutional layer is used to refine the extract features. Finally, the residual learning mechanism [38] is also introduced to learn the residual information from the input and stabilize the training.

2) *Transformer Block:* As we mentioned above, FSAU is mainly designed for local features extraction. However, this is far from enough to restore high-quality face images since the global facial structure (such as face contour) will be ignored due to the limited receptive field of CNN. To solve this problem, we introduce a Transformer Block to collaboratively learn the long-term dependence of images. Motivated by [35], in the multi-head self-attention part, we use the Multi-Dconv Head Transposed Attention (MDTA) to alleviate the time and memory complexity issues. Specifically, to make up for the limitations of the Transformer in capturing local dependencies, deep-wise convolution is introduced to enhance the local features to generate the global attention map. As depicted in Fig. 3 (c), different from the original Transformer block directly achieved $query(Q)$, $key(K)$, and $value(V)$ by a linear layer, a 1×1 convolutional layer is

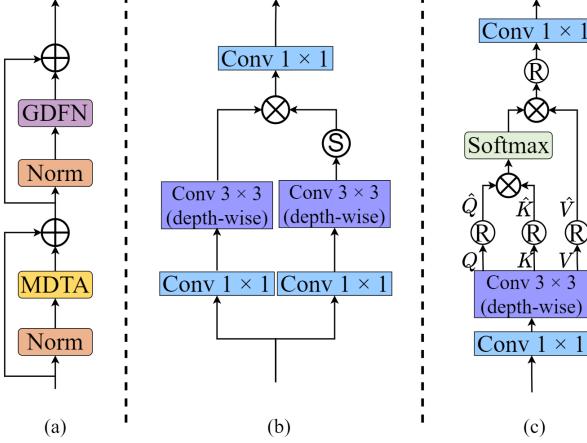


Fig. 3: The architecture of (a) Transformer Block (b) GDFN, and (c) MDTA, respectively.

used to aggregate pixel-level cross-channel context and a 3×3 depth convolutional layer is utilized to encode channel-level spatial context and generate $Q, K, V \in \mathbb{R}^{C \times H \times W}$. Given the input feature $X \in \mathbb{R}^{C \times H \times W}$ and the layer normalized tensor $X' \in \mathbb{R}^{C \times H \times W}$, we have

$$Q = H_{pconv}^{1 \times 1}(H_{dconv}^{3 \times 3}(X')), \quad (2)$$

$$K = H_{pconv}^{1 \times 1}(H_{dconv}^{3 \times 3}(X')), \quad (3)$$

$$V = H_{pconv}^{1 \times 1}(H_{dconv}^{3 \times 3}(X')), \quad (4)$$

where $H_{pconv}^{1 \times 1}(\cdot)$ is the 1×1 point-wise convolutional layer and $H_{dconv}^{3 \times 3}(\cdot)$ is the 3×3 depth-wise convolutional layer.

By calculating the correlation between Q and K , we can obtain global attention weights from different locations, thereby capturing the global information. 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. Thus the dot-product interaction of \hat{Q} and \hat{K} will generate a transposed-attention map with size $\mathbb{R}^{C \times C}$, rather than the huge size of $\mathbb{R}^{HW \times HW}$. After that, the global attention weights are subsequently multiplied with V to get the weighted integrated features $X_w \in \mathbb{R}^{C \times HW}$. This can help the module to capture valuable local context. Finally, we reshape X_w into $\hat{X}_w \in \mathbb{R}^{C \times H \times W}$ and use a 1×1 convolutional layer to realize feature communication. The above procedure can be formulated as follows:

$$X_{weighted} = \text{Softmax}(\hat{Q} \cdot \hat{K} / \sqrt{d}) \cdot \hat{V}, \quad (5)$$

$$Y_M = H_{pconv}^{1 \times 1}(R(X_{weighted})), \quad (6)$$

where Y_M denotes the output of MDTA, $R(\cdot)$ stands for the reshaping operation. Here, \sqrt{d} is a temperature parameter to control the magnitude of the dot product of \hat{K} and \hat{Q} before applying the Softmax function.

At the same time, we also introduce depth-wise convolutions into Gated-Dconv Feed-Forward Network (GDFN) to encode information from spatially neighboring pixel positions, responsible for learning local image structures for effective restoration. Given the input x , we have

$$x' = H_{dconv}^{3 \times 3}(H_{pconv}^{1 \times 1}(x)), \quad (7)$$

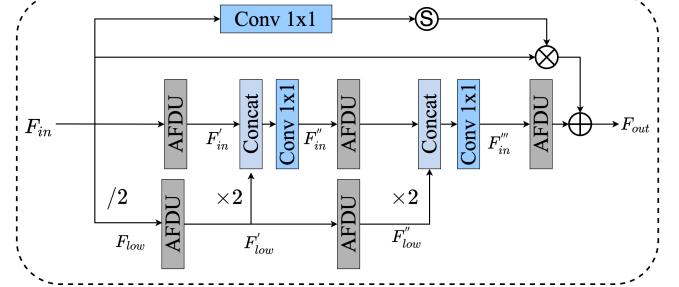


Fig. 4: The architecture of proposed FEU

$$Y_G = H_{pconv}^{1 \times 1}(x' \cdot \sigma(x')), \quad (8)$$

where σ denotes the GELU non-linearity operation [39] and Y_G denotes the output of GDFN.

With the help of FSAU and Transformer Block, LGCM is able to capture both local features and global relationships of faces, which is benefit for high-quality images reconstruction.

C. Feature Refinement Module (FRM)

In the bottleneck stage, we introduce the well-designed Feature Refinement Modules (FRMs) to continuously refine and enhance the important encoded features of the face. As shown in Fig. 1, each FRM encompasses an FSAU and a Feature Enhancement Unit (FEU). To reduce the computational burden and feature redundancy of the network, we use a double-branch structure in FEU. As shown in Fig. 4, the first branch mainly uses AFDUs to extract the information in the original scale, while the second branch extracts features from the down-sampled feature maps, which are then up-sampled to fuse with the outputs of the first branch. In comparison with the general residual learning, we also add a feature self-calibration path to the residual connection to fully mine the hierarchical features and stabilize the training simultaneously. The above operations can be expressed as

$$F'_{in} = f_a(F_{in}), F'_{low} = f_a(\downarrow F_{in}), F''_{low} = f_a(F'_{low}), \quad (9)$$

$$F''_{in} = H_{conv}^{1 \times 1}(H_{cat}(f_a(F'_{in}), \uparrow f_a(F'_{low}))), \quad (10)$$

$$F'''_{in} = H_{cat}(f_a(F''_{in}), \uparrow f_a(F''_{low})), \quad (11)$$

$$F_{out} = f_a(F'''_{in}) + F_{in} \cdot \sigma(H_{conv}^{1 \times 1}(F_{in})), \quad (12)$$

where $f_a(\cdot)$ denotes the operation of AFDU, $H_{cat}(\cdot)$ indicates the feature concatenating operation along the channel dimension, $H_{conv}^{1 \times 1}(\cdot)$ stands for the 1×1 convolutional layer, and σ denotes the Sigmoid function.

D. Multi-scale Feature Fusion Unit (MFFU)

In order to make full use of the multi-scale features extracted in the encoding stage, we introduce the multi-scale feature fusion scheme in the decoding stage to enable the network to have better feature propagation and representation capabilities. Specifically, our main goal is to explore and exploit the features from the encoding stage during the decoding process. However, the size of these features are different, how to integrate these features more effectively

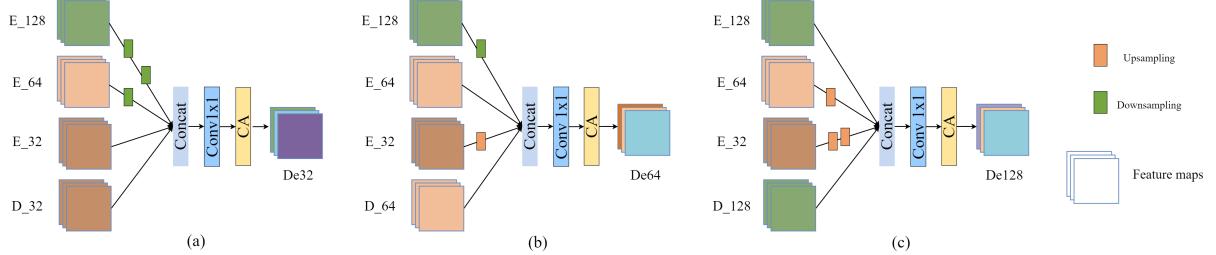


Fig. 5: Schematic diagram of how Multi-scale Feature Fusion Unit (MFFU) aggregates features from different scales.

is critical important. Take the size of the input image as 128×128 as an example, the size of the feature maps we obtained in the encoding stages is 128×128 , 64×64 , and 32×32 , respectively. However, the size of the feature maps in the decoding stage is 32×32 , 64×64 , and 128×128 , successively. To solve this problem, we design a Multi-scale Feature Fusion Unit (MFFU). The details of MFFU are given in Fig 5. According to the figure, we can observe that we first use upsampling and downsampling operations to scale the image feature maps with inconsistent sizes. After unifying the size of all feature maps, we concatenate the four types of feature maps along the channel dimension. Then, we use a 1×1 convolutional layer to generate the preliminary fusion result. Finally, we assign a channel direction attention weight to each channel through the CA mechanism.

Based on the size of the feature maps, the fusion scheme can be divided into three situations. The schematic diagram of how MFFU aggregates features from different scales is shown in Fig 5. For the sake of simplicity, we only give the formulation of Fig 5 (b). The formulation of Fig 5 (b) can be defined as:

$$E_{128_64} = H_{conv}^{k3s2}(E_{128}), \quad (13)$$

$$E_{32_64} = H_{deconv}^{k6s2p2}(E_{32}), \quad (14)$$

$$De'_{64} = H_{conv}^{k1s1}(H_{cat}(E_{128_64}, E_{32_64}, E_{64}, D_{64})), \quad (15)$$

$$De_{64} = CA(De'_{64}), \quad (16)$$

where E_k ($k = 32, 64, 128$) represents the feature maps from the previous three encoding stages with the size of $k \times k$, and D_{64} represents the original feature maps of the current decoder with the size of 64×64 . E_{m_n} indicates that the size of the feature maps has changed from $m \times m$ to $n \times n$. $H_{conv}^{k3s2}(\cdot)$ denotes the 3×3 convolution operation with the stride to be 2, while $H_{deconv}^{k6s2p2}(\cdot)$ denotes the 6×6 transposed convolution operation with stride and padding to be 2. $H_{cat}(\cdot)$ denotes the concatenating operation along the channel dimension. De'_{64} represents the preliminary fusion result and De_{64} means the final fusion result.

E. Model Extension

As we know, Generative Adversarial Network (GAN) has been proven to be effective in recovering photo-realistic images [40], [41]. Therefore, we also extended our model with GAN and propose a extended model in this work, named CNN-Transformer Cooperation Generative Adversarial Network (CTCGAN). In CTCGAN, we use our CTCNet as

the generative model, and utilize the discriminative model in the conditional manner [42]. The new loss functions adopted in training the CTCGAN consists of three parts:

1) *Pixel Loss*: The same as CTCNet, we use the pixel-level loss to constrain the low-level information between the SR image and HR image. It is can be defined as

$$\mathcal{L}_{pix} = \frac{1}{N} \sum_{i=1}^N \|G(I_{LR}^i) - I_{HR}^i\|_1, \quad (17)$$

where $G(\cdot)$ indicates the CTCGAN generator.

2) *Perceptual Loss*: The perceptual loss is mainly used to promote the perceptual quality of the reconstructed SR images. Specifically, we use a pre-trained face recognition VGG19 [43] to extract the facial features. Therefore, we can calculate the feature-level similarity of the two images. The perceptual loss can be defined as

$$\mathcal{L}_{pcp} = \frac{1}{N} \sum_{i=1}^N \sum_{l=1}^{L_{VGG}} \frac{1}{M_{VGG}^l} \|f_{VGG}^l(I_{SR}^i) - f_{VGG}^l(I_{HR}^i)\|_1, \quad (18)$$

where $f_{VGG}^l(\cdot)$ is the l -th layer in VGG, L_{VGG} denotes the total number of layers in VGG, and M_{VGG}^l indicates the number of elements in f_{VGG}^l .

3) *Adversarial Loss*: The principle of GAN is that generator G strives to create fake images, while discriminator D tries to distinguish fake pictures. In other words, the discriminator D aims to distinguish the super-resolved SR image and the HR image by minimizing

$$\mathcal{L}_{dis} = -\mathbb{E}[\log(D(I_{HR}))] - \mathbb{E}[\log(1 - D(G(I_{LR})))]. \quad (19)$$

In addition, the generator tries to minimize

$$\mathcal{L}_{adv} = -\mathbb{E}[\log(D(G(I_{LR})))]. \quad (20)$$

Therefore, CTCGAN is optimized by minimizing the following overall objective function:

$$\mathcal{L} = \lambda_{pix}\mathcal{L}_{pix} + \lambda_{pcp}\mathcal{L}_{pcp} + \lambda_{adv}\mathcal{L}_{adv}, \quad (21)$$

where λ_{pix} , λ_{pcp} , and λ_{adv} indicate the trade-off parameters for the pixle loss, the perceptual loss, and the adversarial loss, respectively.

IV. EXPERIMENTS

A. Datasets

In our experiments, we use CelebA [44] for training and evaluate the model validity on Helen [45] and SCface [46]

TABLE I: Verify the effectiveness of LGCM.

Methods	PSNR↑	SSIM↑	VIF↑	LPIPS↓
w/o LGCM	27.56	0.7867	0.4487	0.2051
LGCM w/o TB	27.82	0.7964	0.4707	0.1833
LGCM w/o FSAU	27.83	0.7972	0.4637	0.1845
LGCM	27.90	0.7980	0.4721	0.1797

datasets. The height and width of the face pictures in CelebA are inconsistent. Therefore, we crop the image according to the center point, and the size is adjusted to 128×128 pixels, which is used as the HR image. Then we down-sample these HR images into 16×16 pixels with bicubic operation and treat them as the LR inputs. We use 18,000 samples of the CelebA dataset for training, 200 samples for validating, and 1,000 samples for testing. Furthermore, we also directly test our model on Helen and SCface datasets using the model trained on CelebA.

B. Implementation Details

We implement our model using the PyTorch framework. Meanwhile, we optimize our model by Adam and set $\beta_1 = 0.9$ and $\beta_2 = 0.99$. The initial learning rate is set to 2×10^{-4} . For CTCGAN, we empirically set $\lambda_{pix} = 1$, $\lambda_{pcp} = 0.01$, and $\lambda_{adv} = 0.01$. We also use Adam to optimize both G and D with $\beta_1 = 0.9$ and $\beta_2 = 0.99$. The learning rates of G and D are set to 1×10^{-4} and 4×10^{-4} , respectively.

To assess the quality of the SR results, we employ four objective image quality assessment metrics: Peak Signal to Noise Ratio (PSNR), Structural Similarity (SSIM) [47], Learned Perceptual Image Patch Similarity (LPIPS) [48], and Visual Information Fidelity (VIF) [49].

C. Ablation Studies

1) *Effectiveness of LGCM*: LGCM is the most important module in CTCNet, which is designed to extract local features and global relationships of the image. At the same time, this is a new attempt to combine CNN and Transformer structures. To verify the effectiveness of LGCM and the feasibility of this combined method, we carried out a series of ablation studies in this part. As we know, LGCM contains an FSAU and a Transformer Block (TB). Therefore, design three modified models. The first model removes all LGCMs in the encoding and decoding stages, marked as “w/o LGCM”. The second model removes all FSAUs while retaining the Transformer Block, marked as “LGCM w/o FSAU”. The third model removes all Transformer Blocks while retaining the FSAU in LGCM, marked as “LGCM w/o TB”. In Table I, we show the results of these modified networks. According the table, we have the following observations: (a) By comparing the first and the last lines in Table I, we can observe that the introduced LGCM can significantly improve the performance of the model. This fully verifies the effectiveness of LGCM; (b) By comparing the first three lines, we can see that the performance of the model can also be improved by introducing FSAU or TB alone. This is because both local features and global relationships of the image are helpful for image reconstruction;

TABLE II: Performance of different numbers of FRM.

Methods	PSNR/SSIM↑	VIF↑	LPIPS↓	Parameters↓
CTCNet-V0	27.77/0.7954	0.4683	0.1856	10.416M
CTCNet-V2	27.83/0.7965	0.4692	0.1858	16.014M
CTCNet-V4	27.87/0.7979	0.4728	0.1834	21.613M
CTCNet-V6	27.85/0.7967	0.4691	0.1872	27.212M

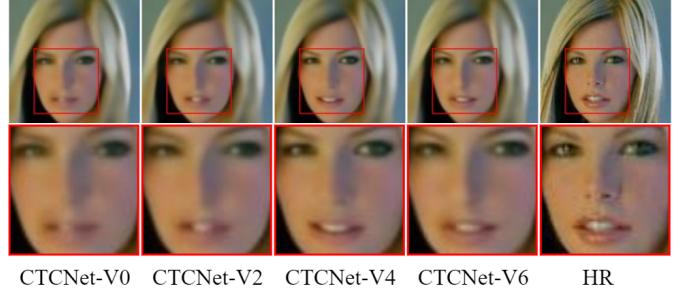


Fig. 6: Visual comparisons of different numbers of FRM on CelebA dataset for $\times 8$ SR.

(c) By comparing the last three lines, we can clearly observe that both FASU and TB play a unique role in FSR tasks. This is because FSAU can capture the local details while TB can capture the global facial structures simultaneously, which provide complementary information for the final SR image reconstruction. Only using one of them cannot achieve the best results. This further verifies the effectiveness of LGCM and the feasibility of combining CNN with Transformer.

2) *Effectiveness of FRM*: To evaluate the effectiveness of FRM, we change the number of FRM in the bottleneck stage. In this part, we gradually increase the numbers of FRM and denote the model with N FRMs as CTCNet-VN, where $N \in \{0, 2, 4, 6\}$. From Table II, we can observe that the model achieve the worst results when all FRMs are removed (CTCNet-V0). This illustrates the necessity of the existence of FRM in CTCNet. Meanwhile, it can be observed that the model performance can be improved with the increase of FRM within a certain range. However, we also notice that when the number of FRM exceeds to 4, the model performance will decreases and the model size will become larger. Therefore, we set $N = 4$ to achieve a good balance between model performance and size. Meanwhile, from Fig. 6, we can intuitively see that as the number of FRM gradually increases from 0 to 4, the facial contours gradually become clear, which fully demonstrates the effectiveness of stacking multiple FRMs.

3) *Effectiveness of MFFU*: MFFU is specially designed for multi-scale feature fusion. In this part, we conducts a series of experiments to demonstrate the effects of Multi-Scale Connections (MSC) and various feature fusion methods in MFFU. The first experiment is used to verify the necessity of MSC. The second and third experiments preserve the MSC but only use the concatenate or add operation to achieve multi-scale features fusion. The last two experiments use channel attention to reweigh the channels after the concatenate or add operation. From Table III, it can be observed that (a) Using multi-scale feature fusion strategy can effectively

TABLE III: Performance of different feature fusion method in MFFU. The last line is the strategy used in our final model.

MSC	Concat	Add	CA	PSNR↑	SSIM↑
×	×	×	×	27.76	0.7961
✓	✓	×	×	27.84	0.7969
✓	×	✓	×	27.82	0.7955
✓	×	✓	✓	27.83	0.7960
✓	✓	×	✓	27.87	0.7979

TABLE IV: Study of each component in FSAU.

CA	SA	PSNR	SSIM↑	VIF↑	LPIPS↓
×	×	27.80	0.7989	0.4701	0.1874
✓	×	27.83	0.7966	0.4673	0.1881
×	✓	27.82	0.7964	0.4676	0.1908
✓	✓	27.87	0.7979	0.4728	0.1834

improve model performance, which proves the importance of multi-scale features for image reconstruction; (b) Using Channel Attention (CA) mechanism has positive effects on improving the model performance; (c) The effect of combining the concatenate operation and CA is apparent. This further verifies that adopting a suitable feature fusion strategy can well provide help for the subsequent reconstruction process.

4) *Study of FSAU*: In FSAU, we use the structure of the nested channel attention mechanism in the spatial attention mechanism to better extract spatial features and promote channel information interaction. To prove the effectiveness of using this nested structure, we remove channel attention and spatial attention respectively to perform ablation studies. From Table IV, we can see the effectiveness enlightened by the channel and spatial attention mechanisms. Adding channel attention or spatial attention alone can only slightly improve the PSNR value by 0.03dB and 0.02dB, respectively. However, when using the nested structure, the PSNR values increase from 27.80dB to 27.87dB. Therefore, we can draw a conclusion that we can gain better performance by applying the channel and spatial attention mechanisms simultaneously.

5) *Study of FEU*: FEU is an essential part of FRM, which uses a double-branch structure to enhance feature extraction. As mentioned earlier, FEU mainly includes several AFDUs and a feature self-calibration path. In this part, we conducted three ablation experiments to verify the effectiveness of AFDU, dual-branch structure, and feature self-calibration path in FEU. From Table V, we can see that (a) If we do not use AFDU in FEU, the performance will drop sharply, and the usage of AFDU increases the PSNR value by 0.1dB; (b) Compared with a simple single-branch structure (without the downsampling and upsampling operations), using the dual-branch structure promotes the PSNR value by 0.06dB. It further verifies that multi-scale feature extraction often has better feature representation; (c) The usage of the feature self-calibration path increases the PSNR value by 0.07dB, since this path can highlight the helpful features with higher activation values.

D. Comparison with Other Methods

In this part, we compare our CTCNet with other state-of-the-art (SOTA) methods, including general image SR methods

TABLE V: Study of each component in FEU.

Methods	PSNR↑	SSIM↑	VIF↑	LPIPS↓
FEU w/o AFDU	27.77	0.7947	0.4628	0.1952
FEU w/o path	27.80	0.7959	0.4659	0.1907
FEU w/o dual	27.81	0.7951	0.4679	0.1933
FEU	27.87	0.7979	0.4728	0.1834

TABLE VI: Comparison results of GAN-based methods for 8× SR on the CelebA and Helen test sets.

Methods	DataSet	PSNR↑	SSIM↑	FID↓	VIF↑
FSRGAN	CelebA	26.49	0.7719	30.60	0.3857
SPARNetHD		27.08	0.7661	29.07	0.4202
CTCGAN (Ours)		27.78	0.7898	25.96	0.4367
FSRGAN	Helen	25.02	0.7279	146.55	0.3400
DICGAN		25.59	0.7398	144.25	0.3925
SPARNetHD		25.86	0.7518	149.54	0.3932
CTCGAN (Ours)		26.41	0.7776	118.05	0.4112

SAN [24], RCAN [23], HAN [25], novel FSR methods FSRNet [4], DICNet [1], FACN [5], SPARNet [10], SISN [17], and pioneer Transformer based image restoration method SwinIR [32]. For a fair comparison, all models are trained using the same CelebA dataset.

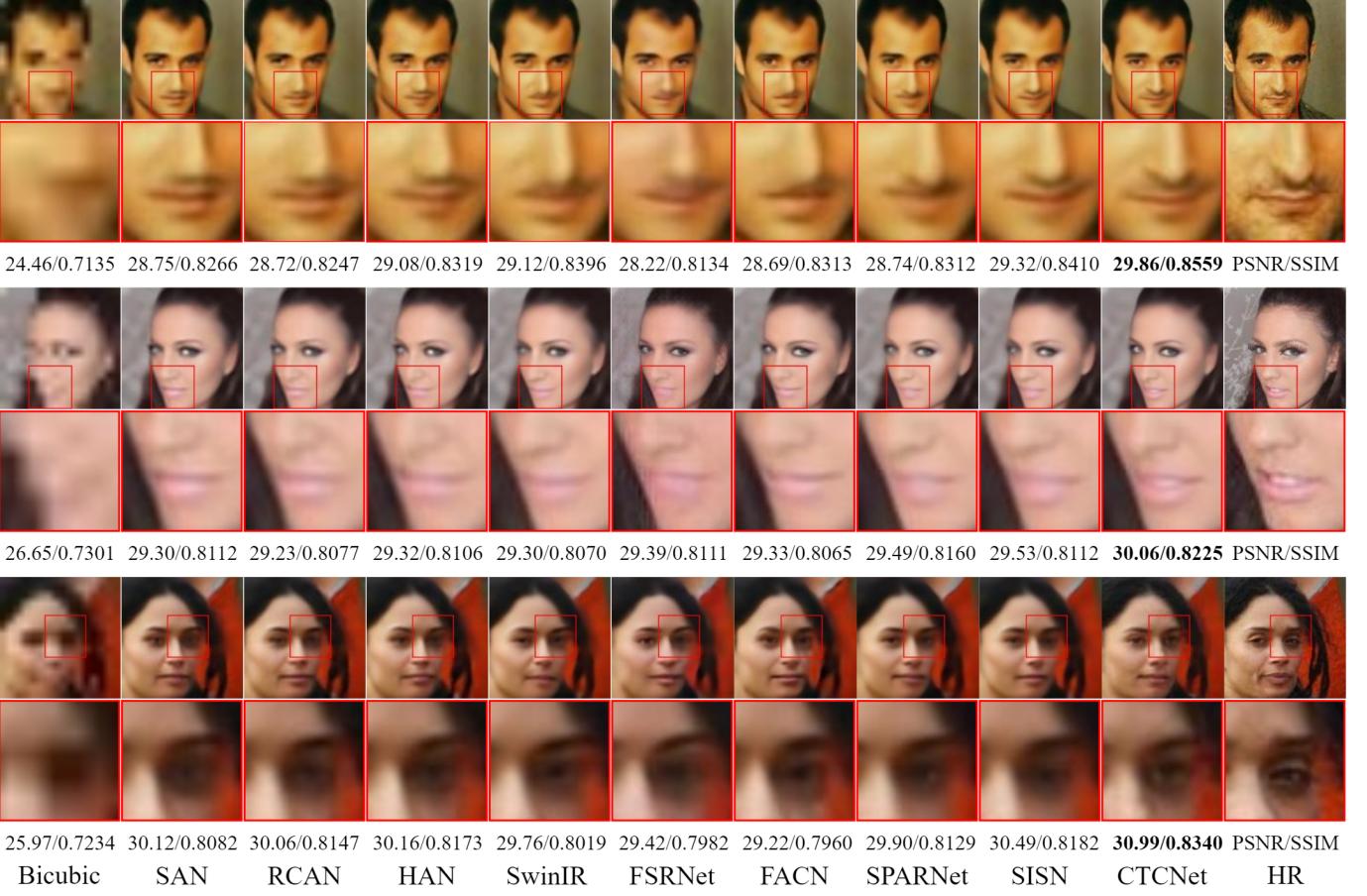
1) *Comparison on CelebA dataset*: The quantitative comparisons with other SOTA method on the CelebA test set are provided in Table VII. According to the table, we can see that CTCNet significantly outperforms other competitive methods in terms of PSNR, VIP, LPIPS, and SSIM. This fully verifies the effectiveness of CTCNet. Meanwhile, from the visual comparisons in Fig. 7 we can see that most of the previous methods cannot clearly restore the eyes and nose in the face, while our CTCNet can better restore face structures and generate more precise results. The reconstructed face images are closer to the real HR images, which further proves the effectiveness and excellence of CTCNet.

2) *Comparison on Helen dataset*: In this part, we directly use the model trained on the CelebA dataset to test the model performance on the Helen test set to study the generality of CTCNet. Table VII lists the quantitative experimental results on the Helen test set for ×8 SR. According to the table, we can clearly see that our CTCNet still achieves the best results on the Helen data set. From Fig. 8, we can also observe that the performance of most competitive methods degrades sharply, they cannot restore faithful facial details, and the shape is blurred. On the contrary, our CTCNet can still restore realistic facial contours and facial details. This further verifies the effectiveness and generality of CTCNet.

3) *Comparison with GAN-based methods*: As we mentioned above, we also propose an extended model named CTCGAN. In this part, we compare our CTCGAN with three popular GAN-based FSR models: FSRGAN [4], DICGAN [1], and SPARNetHD [10]. As we all know, GAN-based SR methods usually have superior visual qualities but lower quantitative values (such as PSNR and SSIM). Therefore, we also introduce Frechet Inception Distance score (FID) [50] as a new metric to evaluate the performance of GAN-based SR methods. In Table VI, we provide the quantitative comparisons of these model on CelebA and Helen test sets. Obviously,

TABLE VII: Quantitative comparisons for $\times 8$ SR on the CelebA and Helen test sets.

Methods	<i>CelebA</i>				<i>Helen</i>			
	PSRN↑	SSIM↑	VIF↑	LPIPS↓	PSRN↑	SSIM↑	VIF↑	LPIPS↓
Bicubic	23.61	0.6779	0.1821	0.4899	22.95	0.6762	0.1745	0.4912
SAN [24]	27.43	0.7826	0.4553	0.2080	25.46	0.7360	0.4029	0.3260
RCAN [23]	27.45	0.7824	0.4618	0.2205	25.50	0.7383	0.4049	0.3437
HAN [25]	27.47	0.7838	0.4673	0.2087	25.40	0.7347	0.4074	0.3274
SwinIR [32]	27.88	0.7967	0.4590	0.2001	26.53	0.7856	0.4398	0.2644
FSRNet [4]	27.05	0.7714	0.3852	0.2127	25.45	0.7364	0.3482	0.3090
DICNet [1]	-	-	-	-	26.15	0.7717	0.4085	0.2158
FACN [5]	27.22	0.7802	0.4366	0.1828	25.06	0.7189	0.3702	0.3113
SPARNet [10]	27.73	0.7949	0.4505	0.1995	26.43	0.7839	0.4262	0.2674
SISN [17]	27.91	0.7971	0.4785	0.2005	26.64	0.7908	0.4623	0.2571
CTCNet (Ours)	28.37	0.8115	0.4927	0.1702	27.08	0.8077	0.4732	0.2094

Fig. 7: Visual comparisons for $\times 8$ SR on the CelebA test set. Obviously, our CTCNet can reconstruct clearer face images.

our CTCGAN gains much better performance than other methods in terms of PSNR, SSIM, FID, and VIF. Meanwhile, the qualitative comparisons on the Helen test set are also provided in Fig. 9. According to the figure, we can see that those competitive methods cannot generate realistic faces and have undesirable artifacts and noise. In contrast, our CTCGAN can restore key facial components and the texture details in the mouth and eyes. This fully demonstrates the effectiveness and excellence of our CTCGAN.

4) *Comparison on real-world surveillance faces:* As we know, restoring face images from real-world surveillance scenarios is still a huge challenge. All the above experiments are in the simulation cases, which can not simulate the real-world

scenarios well. To further verify the effectiveness of our CTCNet, we also conduct experiments on real-world low-quality face images, which are selected from the SCface dataset [46]. The images in SCface are captured by surveillance cameras, which inherently have lower resolutions hence no manual downsampling operation is required.

In this part, we try to restore the SR face images with more texture details and good facial structures. Visual comparison of reconstruction performance is given in Fig. 10. We can see that the face priors based methods reconstruct unsatisfactory results. The reason may be that estimating accurate priors from real-world LR face images is a difficult problem. Meanwhile, inaccurate prior information will bring misleading guidance

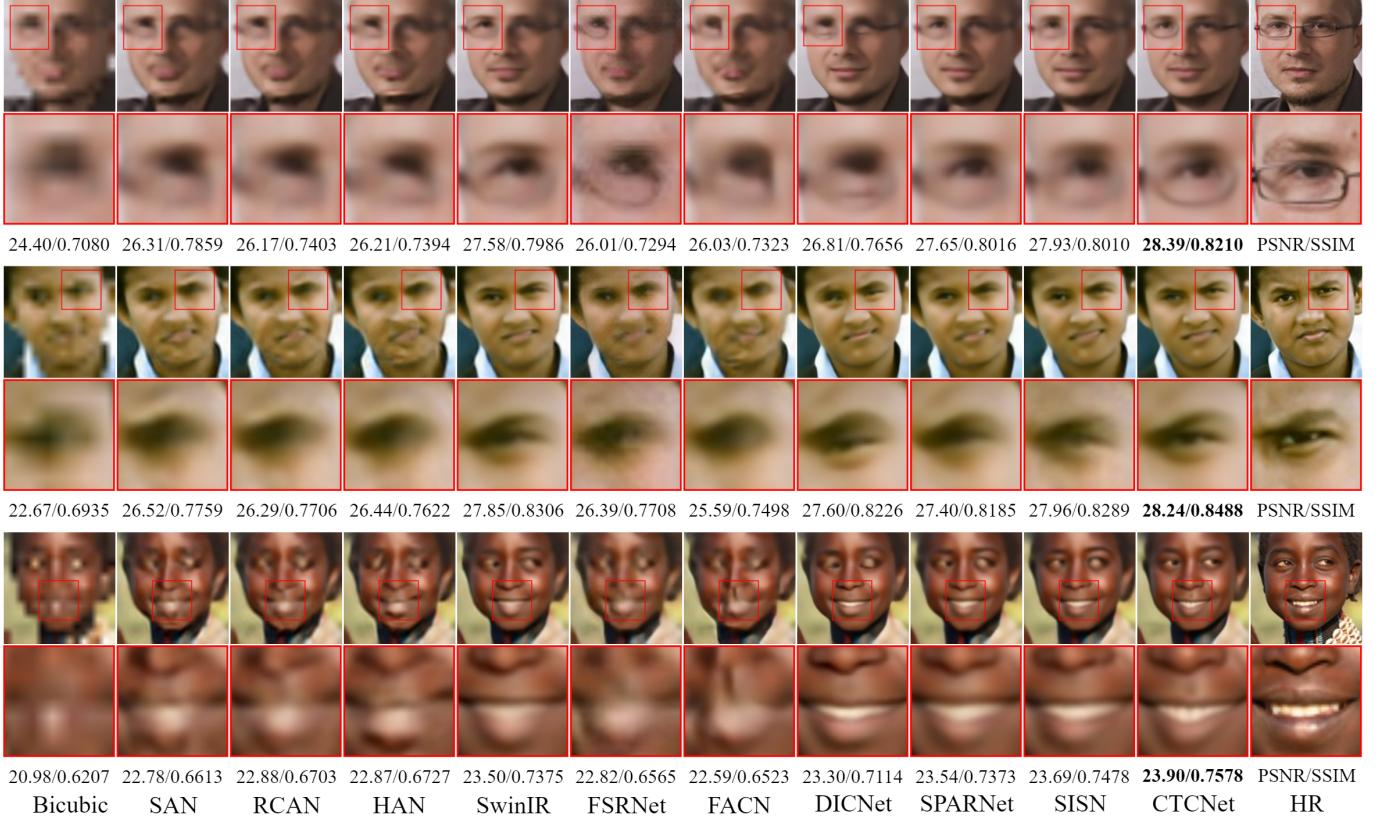


Fig. 8: Visual comparisons for $\times 8$ SR on the Helen test set. Obviously, our CTCNet can reconstruct clearer face images.

TABLE VIII: Comparison results for average similarity of face images super-resolved by different methods.

Methods	Average Similarity									
	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6	Case 7	Case 8	Case 9	Case 10
SAN [24]	0.8897	0.9061	0.9029	0.8940	0.8889	0.9061	0.9042	0.8844	0.9026	0.9107
RCAN [23]	0.8927	0.9000	0.9038	0.8957	0.8963	0.9090	0.9028	0.8807	0.9045	0.9064
HAN [25]	0.8909	0.9096	0.8977	0.9074	0.8914	0.9020	0.9061	0.8740	0.8950	0.9121
SwinIR [32]	0.9087	0.9196	0.8991	0.9079	0.9105	0.9040	0.9119	0.8939	0.9080	0.9093
FSRNet [4]	0.8996	0.8844	0.9017	0.8971	0.8927	0.9061	0.8908	0.8977	0.9040	0.9064
DICNet [1]	0.8859	0.8814	0.8692	0.8760	0.8736	0.8755	0.8837	0.8743	0.8687	0.8914
FACN [5]	0.9048	0.9009	0.9040	0.9017	0.9058	0.8985	0.8970	0.8906	0.8687	0.9007
SPARNet [10]	0.9089	0.9188	0.8995	0.9015	0.9075	0.8980	0.9077	0.9067	0.9025	0.9142
SISN [17]	0.9127	0.9206	0.9086	0.9049	0.9080	0.8999	0.9175	0.9098	0.9060	0.9227
CTCNet	0.9278	0.9219	0.9129	0.9165	0.9243	0.9194	0.9228	0.9136	0.9106	0.9280

to the reconstruction process. In comparison, benefit by the CNN-Transformer Cooperation mechanism, which is the prominent difference between CTCNet and other methods, our CTCNet can recover cleaner facial details and faithful facial structures. We also verify the superiority of our CTCNet over the performance of downstream tasks such as face matching. The high-definition frontal face images of the test candidates are selected as the source samples while the corresponding LR face images captured by the surveillance camera are treated as the target samples. To make the experiments more convincing, we conducted 10 cases. In each case, we randomly select five pairs of candidate samples and calculate the average similarity. The quantitative results can be seen in Table VIII. We can see that our method can achieve higher similarity in each case, which further indicates that our CTCNet can also produce more faithful HR faces in real-world surveillance scenarios,

making it highly practical and applicable.

V. CONCLUSIONS

In this work, we proposed a novel CNN-Transformer Cooperation Network (CTCNet) for face super-resolution. CTCNet uses the multi-scale connected encoder-decoder architecture as the backbone and exhibits extraordinary results. Specifically, we designed an efficient Local-Global Feature Cooperation Module (LGCM), which consists of a Facial Structure Attention Unit (FSAU) and a Transformer block, to simultaneously focus on local facial details and global facial structures. Meanwhile, to further improve the restoration results, we presented a Multi-scale Feature Fusion Unit (MFFU) to adaptively and elaborately fuse the features from different scales and depths. Extensive experiments on both simulated and real-world datasets have demonstrated the superiority of our CTCNet over other competitive methods in terms of quantitative

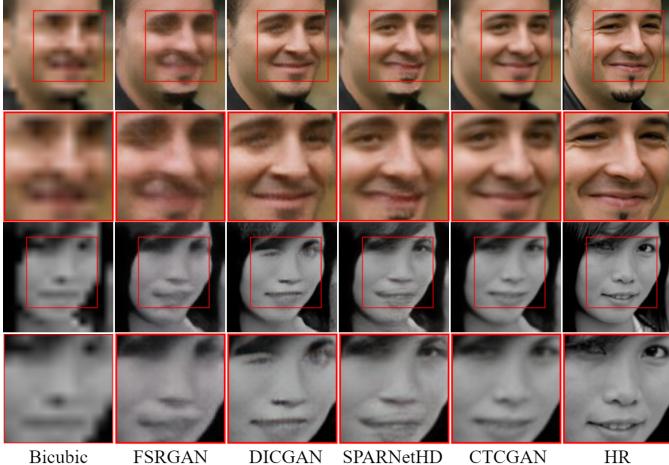


Fig. 9: Visual comparison of different GAN-based methods on the Helen test set. Obviously, our CTCGAN can reconstruct high-quality face images with clear facial components.

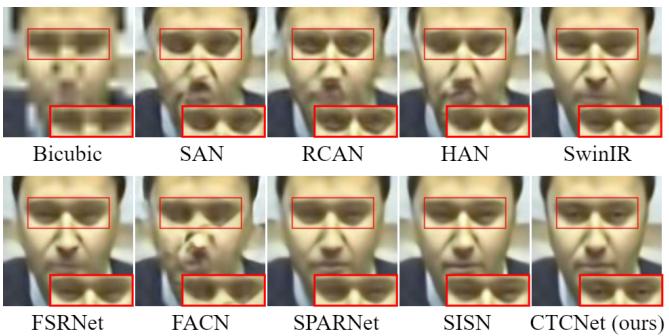


Fig. 10: Visual comparison of respective methods on real-world surveillance scenarios for $\times 8$ SR. Obviously, our CTCNet can reconstruct more clear and accurate eyes.

and qualitative comparisons. Furthermore, its reconstructed images show excellent results in downstream tasks such as face matching, which fully demonstrates its practicality and applicability.

REFERENCES

- [1] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5569–5578.
- [2] X. Hu, W. Ren, J. LaMaster, X. Cao, X. Li, Z. Li, B. Menze, and W. Liu, “Face super-resolution guided by 3d facial priors,” in *Proceedings of the European Conference on Computer Vision*, 2020, pp. 763–780.
- [3] J. Cai, H. Han, S. Shan, and X. Chen, “Fcsr-gan: Joint face completion and super-resolution via multi-task learning,” *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 2, no. 2, pp. 109–121, 2019.
- [4] Y. Chen, Y. Tai, X. Liu, C. Shen, and J. Yang, “Fsrnet: End-to-end learning face super-resolution with facial priors,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2492–2501.
- [5] J. Xin, N. Wang, X. Jiang, J. Li, X. Gao, and Z. Li, “Facial attribute capsules for noise face super resolution,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, pp. 12476–12483.
- [6] X. Yu and F. Porikli, “Ultra-resolving face images by discriminative generative networks,” in *Proceedings of the European Conference on Computer Vision*, 2016, pp. 318–333.
- [7] K. Zhang, Z. Zhang, C.-W. Cheng, W. H. Hsu, Y. Qiao, W. Liu, and T. Zhang, “Super-identity convolutional neural network for face hallucination,” in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 183–198.
- [8] B. Dogan, S. Gu, and R. Timofte, “Exemplar guided face image super-resolution without facial landmarks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019.
- [9] D. Kim, M. Kim, G. Kwon, and D.-S. Kim, “Progressive face super-resolution via attention to facial landmark,” *arXiv preprint arXiv:1908.08239*, 2019.
- [10] C. Chen, D. Gong, H. Wang, Z. Li, and K.-Y. K. Wong, “Learning spatial attention for face super-resolution,” *IEEE Transactions on Image Processing*, vol. 30, pp. 1219–1231, 2021.
- [11] J. Jiang, R. Hu, Z. Wang, and Z. Han, “Face super-resolution via multi-layer locality-constrained iterative neighbor embedding and intermediate dictionary learning,” *IEEE Transactions on Image Processing*, vol. 23, no. 10, pp. 4220–4231, 2014.
- [12] L. Chen, J. Pan, J. Jiang, J. Zhang, and Y. Wu, “Robust face super-resolution via position relation model based on global face context,” *IEEE Transactions on Image Processing*, vol. 29, pp. 9002–9016, 2020.
- [13] G. Gao, Y. Yu, J. Xie, J. Yang, M. Yang, and J. Zhang, “Constructing multilayer locality-constrained matrix regression framework for noise robust face super-resolution,” *Pattern Recognition*, vol. 110, p. 107539, 2021.
- [14] Z. Wang, J. Chen, and S. C. Hoi, “Deep learning for image super-resolution: A survey,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 10, pp. 3365–3387, 2020.
- [15] J. Li, Z. Pei, and T. Zeng, “From beginner to master: A survey for deep learning-based single-image super-resolution,” *arXiv preprint arXiv:2109.14335*, 2021.
- [16] H. Huang, R. He, Z. Sun, and T. Tan, “Wavelet-srnet: A wavelet-based cnn for multi-scale face super resolution,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1689–1697.
- [17] T. Lu, Y. Wang, Y. Zhang, Y. Wang, L. Wei, Z. Wang, and J. Jiang, “Face hallucination via split-attention in split-attention network,” in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 5501–5509.
- [18] X. Yu, B. Fernando, B. Ghanem, F. Porikli, and R. Hartley, “Face super-resolution guided by facial component heatmaps,” in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 217–233.
- [19] M. Li, Z. Zhang, J. Yu, and C. W. Chen, “Learning face image super-resolution through facial semantic attribute transformation and self-attentive structure enhancement,” *IEEE Transactions on Multimedia*, vol. 23, pp. 468–483, 2021.
- [20] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, “Cbam: Convolutional block attention module,” in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 3–19.
- [21] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141.
- [22] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, “Eca-net: Efficient channel attention for deep convolutional neural networks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11534–11542.
- [23] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, “Image super-resolution using very deep residual channel attention networks,” in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 286–301.
- [24] T. Dai, J. Cai, Y. Zhang, S.-T. Xia, and L. Zhang, “Second-order attention network for single image super-resolution,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11065–11074.
- [25] B. Niu, W. Wen, W. Ren, X. Zhang, L. Yang, S. Wang, K. Zhang, X. Cao, and H. Shen, “Single image super-resolution via a holistic attention network,” in *Proceedings of the European Conference on Computer Vision*, 2020, pp. 191–207.
- [26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Proceedings of the Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.
- [27] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018.
- [28] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*,

- “An image is worth 16x16 words: Transformers for image recognition at scale,” *arXiv preprint arXiv:2010.11929*, 2020.
- [29] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, “Training data-efficient image transformers & distillation through attention,” in *International Conference on Machine Learning*, 2021, pp. 10347–10357.
- [30] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, “End-to-end object detection with transformers,” in *Proceedings of the European Conference on Computer Vision*, 2020, pp. 213–229.
- [31] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, “Deformable detr: Deformable transformers for end-to-end object detection,” *arXiv preprint arXiv:2010.04159*, 2020.
- [32] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, “Swinir: Image restoration using swin transformer,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1833–1844.
- [33] Z. Lu, H. Liu, J. Li, and L. Zhang, “Efficient transformer for single image super-resolution,” *arXiv preprint arXiv:2108.11084*, 2021.
- [34] Z. Wang, X. Cun, J. Bao, and J. Liu, “Uformer: A general u-shaped transformer for image restoration,” *arXiv preprint arXiv:2106.03106*, 2021.
- [35] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, “Restormer: Efficient transformer for high-resolution image restoration,” *arXiv preprint arXiv:2111.09881*, 2021.
- [36] Y. Chen, C. Shen, X.-S. Wei, L. Liu, and J. Yang, “Adversarial posenet: A structure-aware convolutional network for human pose estimation,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1212–1221.
- [37] A. Newell, K. Yang, and J. Deng, “Stacked hourglass networks for human pose estimation,” in *Proceedings of the European Conference on Computer Vision*, 2016, pp. 483–499.
- [38] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.
- [39] D. Hendrycks and K. Gimpel, “Gaussian error linear units (gelus),” *arXiv preprint arXiv:1606.08415*, 2016.
- [40] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang *et al.*, “Photo-realistic single image super-resolution using a generative adversarial network,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4681–4690.
- [41] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. Change Loy, “Esrgan: Enhanced super-resolution generative adversarial networks,” in *Proceedings of the European Conference on Computer Vision (ECCV) workshops*, 2018, pp. 1–16.
- [42] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-image translation with conditional adversarial networks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1125–1134.
- [43] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.
- [44] Z. Liu, P. Luo, X. Wang, and X. Tang, “Deep learning face attributes in the wild,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3730–3738.
- [45] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang, “Interactive facial feature localization,” in *Proceedings of the European Conference on Computer Vision*, 2012, pp. 679–692.
- [46] M. Grgic, K. Delac, and S. Grgic, “Scface—surveillance cameras face database,” *Multimedia Tools and Applications*, vol. 51, no. 3, pp. 863–879, 2011.
- [47] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: from error visibility to structural similarity,” *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.
- [48] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, “The unreasonable effectiveness of deep features as a perceptual metric,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 586–595.
- [49] H. R. Sheikh and A. C. Bovik, “Image information and visual quality,” *IEEE Transactions on Image Processing*, vol. 15, no. 2, pp. 430–444, 2006.
- [50] A. Obukhov and M. Krasnyanskiy, “Quality assessment method for gan based on modified metrics inception score and fréchet inception distance,” in *Proceedings of the Computational Methods in Systems and Software*, 2020, pp. 102–114.