IDPT: Interconnected Dual Pyramid Transformer for Face Super-Resolution

Jingang Shi[1] , **Yusi Wang**[1] , **Songlin Dong**[1] , **Xiaopeng Hong**[1] , **Zitong Yu**[2*] , **Fei Wang**[1] , **Changxin Wang**[1] and **Yihong Gong**[1]

[1]School of Software Engineering, Xi'an Jiaotong University
[2]School of Electrical and Electronic Engineering, Nanyang Technological University
{jingang, yusiwang, dsl972731417, feynmanw, wcxhello, ygong}@xjtu.edu.cn,
hongxiaopeng@ieee.org, zitong.yu@ntu.edu.sg

Abstract

Face Super-resolution (FSR) task works for generating high-resolution (HR) face images from the corresponding low-resolution (LR) inputs, which has received a lot of attentions because of the wide application prospects. However, due to the diversity of facial texture and the difficulty of reconstructing detailed content from degraded images, FSR technology is still far away from being solved. In this paper, we propose a novel and effective face super-resolution framework based on Transformer, namely Interconnected Dual Pyramid Transformer (IDPT). Instead of straightly stacking cascaded feature reconstruction blocks, the proposed IDPT designs the pyramid encoder/decoder Transformer architecture to extract coarse and detailed facial textures respectively, while the relationship between the dual pyramid Transformers is further explored by a bottom pyramid feature extractor. The pyramid encoder/decoder structure is devised to adapt various characteristics of textures in different spatial spaces hierarchically. A novel fusing modulation module is inserted in each spatial layer to guide the refinement of detailed texture by the corresponding coarse texture, while fusing the shallow-layer coarse feature and corresponding deep-layer detailed feature simultaneously. Extensive experiments and visualizations on various datasets demonstrate the superiority of the proposed method for face super-resolution tasks.

1 Introduction

Face Super-resolution (FSR), also known as face hallucination, works for generating high-resolution (HR) face images from the corresponding low-resolution (LR) inputs [Jiang *et al.*, 2021; Shi *et al.*, 2018; Shi and Zhao, 2019]. Since most existing facial analysis techniques (*e.g.*, face detection, face alignment, face recognition) would suffer from worse performance when encounter the degradation of very low resolution problem, the research on FSR has far-reaching significance.

FSR is a challenging task with promising application prospects, which can be considered as a specific task of single image super resolution (SISR). Different from SISR, the target of FSR concentrates on the recovery of important facial structures (*e.g.*, facial components, facial contour) instead of arbitrary details in natural images. These structures are crucial for displaying the characteristics of human faces, but the recovery is difficult when the input low-resolution faces need a large magnification in the reconstruction. Recently, most FSR algorithms have been developed by employing deep Convolutional Neural Network (CNN) [Chen *et al.*, 2018; Yu *et al.*, 2018; Ma *et al.*, 2020] to reconstruct facial textures. Although these CNN-based methods have obtained impressive performance, they still suffer from several barriers which would hinder FSR algorithms to achieve perfect results. As we know, convolution is a local operation which aims to extract features from local regions in the image. Thus, it has a limitation in describing the interactions between different regions and capturing long-range dependencies. To alleviate the problem, some recent FSR approaches usually construct deeper network structure and stack manually designed feature reconstruction blocks (FRBs) to explore cross-region detailed features. However, several disadvantages exist in these network architectures that need to be further solved. First, though it could achieve better performance by stacking more FRBs and employing deeper network, the improvement is not effective compared to the increase of computational cost. Second, since the input images suffer from low-resolution and other degraded factors, some inaccurate details can be produced in feature extraction layers. The artifacts from shallow layers will be accumulated and become serious when the network architecture goes deeper. Third, various local contents may have different intrinsic characteristics (*e.g.*, width, height, complexity), so it is better to design hierarchical structure for adapting the diversity of textures.

Recently, Transformer takes the advantage of self-attention mechanism to explore global interactions and has achieved outstanding performance in the vision fields [Cao *et al.*, 2021; Liang *et al.*, 2021; Yu *et al.*, 2022]. The pioneering work ViT [Dosovitskiy *et al.*, 2021] directly divides the image into patches with fixed size to adapt the Transformer module, which presents its capability in classification task. After the success of ViT, kinds of Transformers have been proposed recently [Khan *et al.*, 2021]. One representative work is Swin

*Corresponding author

Transformer [Liu *et al.*, 2021] which designs the shifted window based self-attention and presents superior performance on vision tasks. However, current Transformer-based methods have few researches on the FSR task, so how to efficiently use Transformer to solve it is still a meaningful challenge.

In this paper, we propose the Interconnected Dual Pyramid Transformer (IDPT) framework for face super-resolution task. There are three core innovations for IDPT. First, we design a dual pyramid encoder/decoder Transformer architecture which could extract hierarchical features on various spatial dimensions from shallow to deep layers. Concretely, we increase the depth of Transformer blocks in the encoder pathway gradually while decrease it in the decoder pathway. Compared with the methods that simply stack network depth, the proposed architecture is more adaptive to learn high-quality textural details for reconstruction.

Second, inspired by the human visual cognitive mechanism that human brain modulates the ventral pathway through the subcortical pathway to improve the recognition ability [Liu *et al.*, 2017], we design a novel fusing modulation module (FMM) to modulate the deep-layer in decoder by the corresponding shallow-layer in encoder. As we know, the shallow-layer extracts the coarse feature while the deep-layer could produce the detailed feature. FMM is proposed to refine the detailed feature by the guidance of shallow-layer for eliminating the deep-layer artifacts, while also fusing the coarse feature and detailed feature for improving the reconstruction capability of the network.

Finally, we propose the bottom pyramid feature extractor (BPFE) to construct the relationship between the bottom layers of dual pyramid structures. It serves as a coarse-to-fine feature extraction procedure, which explores the detailed feature from the output of the encoder and further feeds it into the decoder pyramid. Compared with prevalent face super-resolution methods, the proposed IDPT framework could achieve superior results on multiple datasets. The visualizations could also verify that our method is very effective to recover complex textural details on facial images.

Our main contributions are summarized as follows:

- We propose a novel dual pyramid Transformer architecture to deal with FSR task, which could hierarchically extract the shallow-layer coarse feature and deep-layer detailed feature from different spatial spaces.
- Inspired by the research of the human visual cognitive mechanism, we propose the fusing modulation module to refine the deep-layer detailed feature and fuse it with corresponding shallow-layer coarse feature.
- A coarse-to-fine feature extraction module is constructed across the bottom of dual pyramids, which is employed to explore the underlying relationship between encoder and decoder.
- Extensive experiments and visualizations demonstrate the superiority of the proposed IDPT when compared with the state-of-the-art methods.

2 Related Works

Face Super-Resolution. In the past years, deep learning based methods have great promoted the development of FSR field. URDGN [Yu and Porikli, 2016] firstly proposes a discriminative generative network structure which is utilized to restore details from very low resolution face images. Considering the identity information of face image, SICNN [Zhang *et al.*, 2018] defines the super-identity loss to maintain the consistency in the identity metric space for achieving better performance. SPARNet [Chen *et al.*, 2020] introduces a spatial attention mechanism that could focus on the recovery of feature-rich regions to generate high-quality SR images. SISN [Lu *et al.*, 2021] proposes the external-internal split attention group to fuse structural information and textural details of facial images in the reconstruction. In some latest FSR methods, facial prior is also utilized as an additional constraint to supervise the training phase. FSRNet [Chen *et al.*, 2018] makes use of landmark heatmaps and parsing maps in the training process, which induces impressive results. In [Kim *et al.*, 2019], it proposes a facial attention loss which encourages the network to focus on the alignment of facial landmarks. DIC [Ma *et al.*, 2020] uses a deep iterative collaboration network for FSR to enable face images recovery and landmark estimation in a mutually reinforcing mechanism. GPEN [Yang *et al.*, 2021] designs the neural network based on GAN prior to solve blind face restoration problem.

Visual Transformer. Transformer model has gained great success in natural language processing (NLP) domain. Inspired by its significant performance in NLP, Transformer is utilized to solve computer vision tasks recently. ViT [Dosovitskiy *et al.*, 2021] firstly proves that a pure Transformer structure could achieve state-of-the-art performance on classification task. With pre-training on large-scale datasets, it reshapes the inputs into a set of medium-size flattened patches and surpasses the capability of CNN-based classification layers. T2T [Yuan *et al.*, 2021] further aggregates nearby tokens into a single token recursively, instead of the simple image split in ViT. Another variant called Swin Transformer is also proposed in [Liu *et al.*, 2021]. Due to its shifted window-based multi-head attentions, it achieves state-of-the-art performance in various fields, such as image classification, object detection, and semantic segmentation. Recently, Transformer-based structure has also been utilized for image restoration. In [Cao *et al.*, 2021], the authors introduce VSR-Transformer with self-attention mechanism to deal with video super-resolution (SR). SwinIR [Liang *et al.*, 2021] is also proposed recently, where the network architecture is constructed based on [Liu *et al.*, 2021] for image restoration.

3 Approach

3.1 Overview

In the FSR method, the task is to restore the high-frequency facial details from the LR input face images and generate super-resolved results. Here I^{LR}, I^{HR}, and I^{SR} denote the LR images, the ground truth HR images, and the generated super-resolved images respectively. Given an LR face image I^{LR}, we first upsample it to the same spatial dimension of

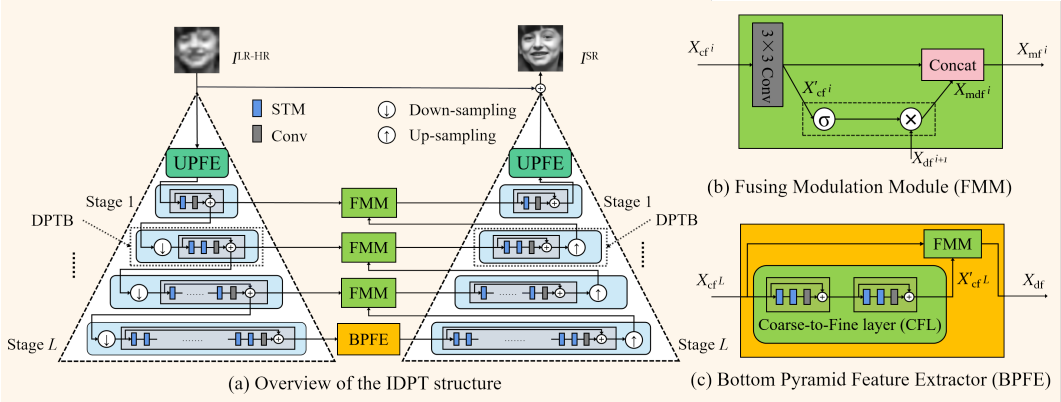

Figure 1: (a) Framework of the IDPT. "UPFE", "DPTB" and "⊕" denote a 3×3 convolution with LeakyReLU, dual pyramid Transformer blocks and element-wise addition operator, respectively. (b) The architecture of the Fusing Modulation Module (FMM). (c) The architecture of the Bottom Pyramid Feature Extractor (BPFE).

I^{HR} by bicubic interpolation. The interpolated image is represented as $I^{\mathrm{LR-HR}}$ and fed to IDPT to generate I^{SR}.

As shown in Figure 1, we propose the Interconnected Dual Pyramid Transformer (IDPT) to extract feature progressively, which consists of the Dual Pyramid Transformer Blocks (DPTB), the Fusing Modulation Modules (FMM) and the Bottom Pyramid Feature Extractor (BPFE). DPTB is employed to hierarchically extract the coarse feature and detailed feature from various spatial dimensions. FMM is devised to refine the deep-layer detailed feature by coarse feature and further fuses the shallow-layer and deep-layer for reconstructing better results. BPFE is designed to establish the underlying coarse-to-fine connection between encoder and decoder.

In the proposed algorithm, we first utilize the upper pyramid feature extractor (UPFE) which is a 3×3 convolutional layer with LeakyReLU to extract the low-frequency texture feature X_{lf}. Then X_{lf} is fed into L encoder stages to hierarchically extract the coarse feature at various spatial dimensions.

The coarse feature X_{cf^i} extracted by the i-th encoder can be formulated as:

$$X_{\mathrm{cf}^i} = \mathrm{F}_{\mathrm{ES}^i}(X_{\mathrm{cf}^{i-1}}), \tag{1}$$

where $\mathrm{F}_{\mathrm{ES}^i}(\cdot)$ denotes the i-th encoder stage.

At the end of the encoder, we propose the BPFE to extract the detailed feature X_{df} for the bottom layer of pyramid structure, which can be formulated as:

$$X_{\mathrm{df}} = \mathrm{F}_{\mathrm{BPFE}}(X_{\mathrm{cf}^L}), \tag{2}$$

Then X_{df} is fed into L decoder stages to extract detailed feature. In each decoder stage, both the features from the $(i+1)$-th decoder stage and the i-th encoder stage are fed into the FMM to get the modulated feature X_{mf^i}, which is the input as the i-th decoder stage ($i < L$).

The i-th decoder stage extracts the detailed feature X_{df^i} as:

$$X_{\mathrm{df}^i} = \mathrm{F}_{\mathrm{DS}^i}(X_{\mathrm{mf}^i}), \tag{3}$$

where $\mathrm{F}_{\mathrm{DS}^i}(\cdot)$ represents the i-th decoder stage.

After all decoder stages, we apply the UPFE (w/o LeakyReLU) to get high-frequency texture feature I^{HF}. Finally, the generated super-resolved image I^{SR} is obtained by:

$$I^{\mathrm{SR}} = I^{\mathrm{LR-HR}} + I^{\mathrm{HF}} \tag{4}$$

3.2 The Dual Pyramid Transformer Structure

Swin Transformer Module (STM). Inspired by the work of [Liu *et al.*, 2021], we utilize the basic block of Swin Transformer as the feature extraction module, namely Swin Transformer module (STM). Given an input of size $H \times W \times C$, Swin Transformer reshapes and splits it into non-overlapping windows of $M \times M$ size. Then it calculates self-attention separately in each window.

For a local window feature $X \in \mathbb{R}^{M^2 \times C}$, it is first linearly transformed to three parts, *i.e.*, the query $Q \in \mathbb{R}^{M^2 \times d}$, key $K \in \mathbb{R}^{M^2 \times d}$ and value $V \in \mathbb{R}^{M^2 \times d}$ respectively, where d is the query/key/value dimension. Then the self attention calculation for a local window can be formulated as:

$$\mathrm{Attention}(Q, K, V) = \mathrm{Softmax}(QK^T/\sqrt{d} + B)V, \tag{5}$$

where B is the learnable relative position bias. The attention function is calculated h times in parallel and concatenated to get multi-head self-attention (MSA).

The MLP layer, including two FC layers with a GELU nonlinearity, is applied after MSA layer for feature transformation. The LayerNorm (LN) is applied before each layer, and residual connection is employed after each layer.

Moreover, to introduce connections between neighboring local windows, the features are shifted by $(\lfloor \frac{M}{2} \rfloor, \lfloor \frac{M}{2} \rfloor)$. We employ regular and shifted window partitioning alternately to accomplish cross-window connections.

Dual Pyramid Transformer Block (DPTB). The DPTB is composed of the Encoder Transformer Block (ETB) and the Decoder Transformer Block (DTB). Each block contains a stack of STMs, following by a 3×3 convolutional layer with residual connection. At the beginning of each ETB, we use a 4×4 convolutional layer with stride 2 to downsample the size of feature maps by half and double the feature channels. As

for DTB, we use a 2×2 transposed convolution with stride 2 to upsample the size of feature maps and reduce the number of feature channels to one fourth at the end. An exception is that there is not the downsampling/upsampling operator at stage 1. The number of the STMs d_{STMs} depends on the depth of stage i:

$$d_{\mathrm{STMs}} = 2^{i-1} \tag{6}$$

The Dual Pyramid Structure. The proposed method hierarchically extracts shallow-layer coarse feature and deep-layer detailed feature by DPTBs. In the encoder stage, we progressively increase the number of STMs in DPTB while it is decreased progressively in the decoder stage. The whole framework looks like a dual pyramid, which improves the capability of the network to capture feature.

3.3 The Fusing Modulation Module

As illustrated in Figure 1(b), the multi-layer fusing modulation module (FMM) contains the extraction stage, the modulation stage, and the fusion stage. For the i-th FMM, we first apply a 3×3 convolutional layer as the extraction stage $\mathrm{F}_{\mathrm{ES}}(\cdot)$ to extract the feature X'_{cf^i} from the corresponding encoder stage as:

$$X'_{\mathrm{cf}^i} = \mathrm{F}_{\mathrm{ES}}(X_{\mathrm{cf}^i}) \tag{7}$$

Then, the modulation stage utilizes X'_{cf^i} to modulate the deep-layer detailed feature from $(i+1)$-th decoder stage to get the modulated features X_{mdf^i} as:

$$X_{\mathrm{mdf}^i} = (\sigma(X'_{\mathrm{cf}^i}) \otimes X_{\mathrm{df}^{i+1}}), \tag{8}$$

where $\sigma(\cdot)$ and $\otimes$ represent the sigmoid logistic function and the element-wise multiplication operator separately.

Finally, the fusion stage concatenates X'_{cf^i} and X_{mdf^i} as the input features of i-th decoder stage X_{mf^i}. X_{mf^i} can be formulated as:

$$X_{\mathrm{mf}^i} = \mathrm{Concat}[X'_{\mathrm{cf}^i}, X_{\mathrm{mdf}^i}] \tag{9}$$

The overall process of FMM is formulated as:

$$X_{\mathrm{mf}^i} = \mathrm{FMM}(X_{\mathrm{cf}^i}, X_{\mathrm{df}^{i+1}}) \tag{10}$$

3.4 The Bottom Pyramid Feature Extractor

We design the bottom pyramid feature extractor (BPFE) to refine the coarse feature from encoder and produce the detailed feature as the input of decoder stages. As shown in Figure 1(c), BPFE contains a coarse-to-fine layer (CFL) which is composed of two consecutive blocks with STMs and convolutional layers. X_{cf^L} is fed into the CFL to further get X'_{cf^L}:

$$X'_{\mathrm{cf}^L} = \mathrm{F}_{\mathrm{CFL}}(X_{\mathrm{cf}^L}) \tag{11}$$

At the end of the BPFE, we also apply the FMM, which modulates X'_{cf^L} with X_{cf^L} to get the detailed features X_{df} :

$$X_{\mathrm{df}} = \mathrm{FMM}(X_{\mathrm{cf}^L}, X'_{\mathrm{cf}^L}) \tag{12}$$

3.5 Loss Function

In this section, we implement two variants of the proposed algorithm (*i.e.*, IDPT and IDPT-GAN) by switching the loss function in the training phase. Similar to previous methods [Ma *et al.*, 2020], IDPT is designed to guarantee the consistency in pixel-level, while the target of IDPT-GAN is to achieve better perceptual quality.

Pixel-level loss: The L1-norm is used to constrain the mapping between generated SR images and HR images at the pixel level:

$$\mathcal{L}^{\mathrm{Pix}} = \frac{1}{N} \sum_{i=1}^{N} \| I^{\mathrm{SR}} - I^{\mathrm{HR}} \|_1, \tag{13}$$

where N is the number of training set, and the SR images can be formulated as $I^{\mathrm{SR}} = \mathrm{F}_{\mathrm{IDPT}}(I^{\mathrm{LR-HR}})$.

Adversarial loss: To make a fair comparison, we employ similar approach as [Ma *et al.*, 2020] to construct the adversarial loss. Specifically, we build the same discriminator D as [Ma *et al.*, 2020] and employ IDPT as the generator G. The generator aims to produce SR images that could fool the discriminator. Meanwhile, the discriminator tries to distinguish the ground-truth HR face images and the SR face images reconstructed by the generator. The loss functions of D and G are formulated as:

$$\mathcal{L}^{\mathrm{Dis}} = -\mathbb{E}[log(D(I^{\mathrm{HR}}))] - \mathbb{E}[log(1 - D(G(I^{\mathrm{LR-HR}})))] \tag{14}$$

$$\mathcal{L}^{\mathrm{Gen}} = -\mathbb{E}[log(D(G(I^{\mathrm{LR-HR}})))] \tag{15}$$

Perceptual loss: To improve the perceptual quality of generated images, we also apply the perceptual loss in our method. Similar to [Ma *et al.*, 2020], we employ the pretrained LightCNN [Wu *et al.*, 2018] to extract features. The loss function could be formulated as:

$$\mathcal{L}^{\mathrm{Pcp}} = \mathbb{E}[\| \phi(I^{\mathrm{SR}}) - \phi(I^{\mathrm{HR}}) \|_1], \tag{16}$$

where $\phi(\cdot)$ denotes the pretrained LightCNN.

For IDPT, we only employ the pixel-level loss $\mathcal{L}^{\mathrm{Pix}}$ as the loss function in the training phase. To train IDPT-GAN, the overall GAN-based loss function to optimize generator is formulated as:

$$\mathcal{L}^{G} = \mathcal{L}^{\mathrm{Pix}} + \lambda^{\mathrm{Adv}} \cdot \mathcal{L}^{\mathrm{Gen}} + \lambda^{\mathrm{Pcp}} \cdot \mathcal{L}^{\mathrm{Pcp}}, \tag{17}$$

where λ^{Adv} and λ^{Pcp} represent coefficients of adversarial loss and perceptual loss respectively.

4 Experiments

4.1 Datasets

The CelebA dataset [Liu *et al.*, 2015] is utilized to train IDPT in the experiments. To construct the training set, we first crop the face regions from each image by MTCNN [Zhang *et al.*, 2016] and resize them to the size of 128×128 without any pre-alignment as HR training images. Then, we downsample the HR images to the size of 16×16 by bicubic to obtain

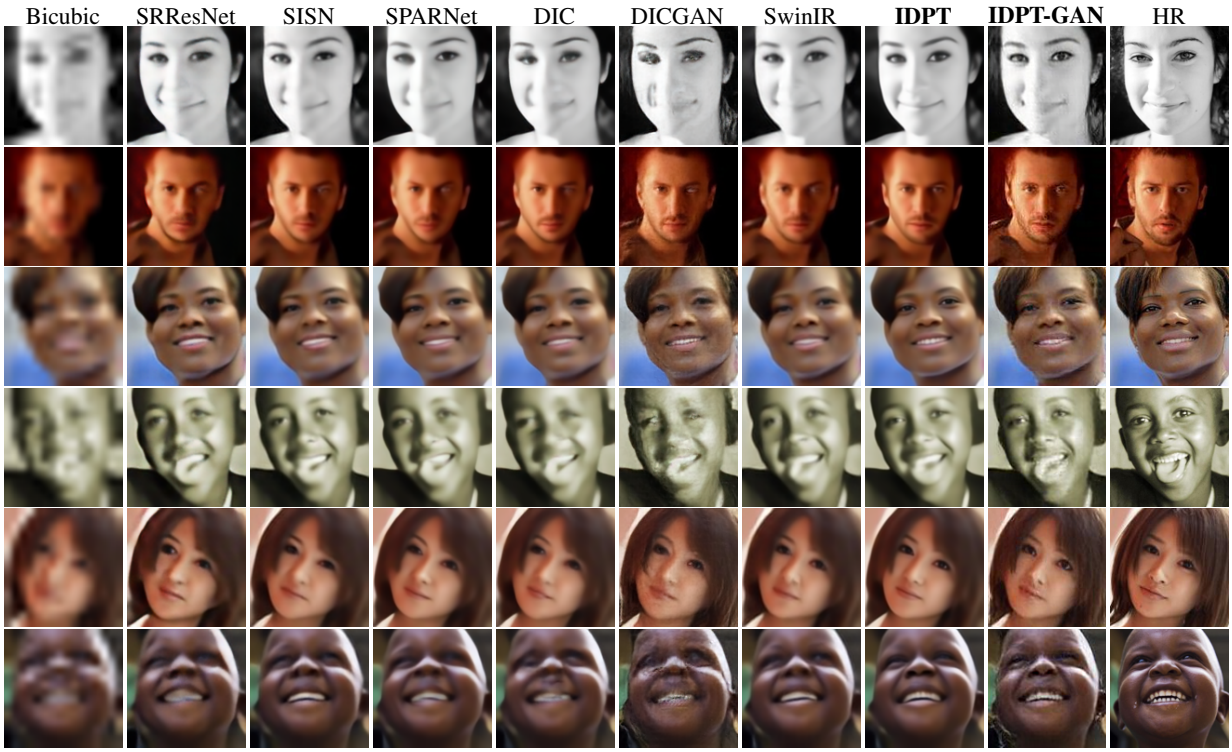

Figure 2: Visualization among different methods. The restored images from previous methods have undesirable artifacts on key facial parts. In contrast, the proposed IDPT and IDPT-GAN can better restore facial details, especially on eyes, lip and teeth. The proposed method performs well even under the scenarios with exaggerated expressions (e.g., smiling). Please zoom in for better comparison.

Method	Helen		CelebA	
	PSNR(dB)	SSIM	PSNR(dB)	SSIM
Bicubic	23.80	0.6745	23.56	0.6361
SRResNet	26.48	0.7962	26.38	0.7690
SISN	27.00	0.8074	26.76	0.7778
SPARNet	27.43	0.8201	27.17	0.7911
SwinIR	**27.50**	**0.8215**	27.18	0.7906
DIC	26.94	0.7994	**27.41**	**0.7983**
DICGAN	25.90	0.7516	26.34	0.7562
IDPT	**27.96**	**0.8355**	**27.59**	**0.8045**
IDPT-GAN	26.57	0.7837	26.19	0.7480

Table 1: Quantitative comparison with state-of-the-art FSR methods on CelebA and Helen dataset. Best and second best performance are **highlighted** in *red* and *blue* colors, respectively.

the corresponding LR training set. In the testing phase, we randomly select 1000 images from CelebA dataset and 50 images from Helen dataset [Le *et al.*, 2012] to conduct the experiments. The generated SR images are evaluated with PSNR and SSIM metrics calculated on the Y channel of transformed YCbCr space.

4.2 Implementation Details

Training Setting. For UPFE, it produces feature maps with channel number of 32 for the first encoder stage, while the parameter α of LeakyReLU is set to 0.01. In the dual pyramid structure, the maximum number of stages L is set to 4. Notice that L could be set as a smaller value to reduce the parameters of network. For STMs, the attention head number and window size are set to 2^{i-1} (i = 1, ..., L) and 8×8, respectively. We train IDPT by AdamW optimizer with β_1 = 0.9, β_2 = 0.99, $\epsilon = 10^{-8}$, and weight decay is set to 0.02. The learning rate is 2×10^{-4} and the batchsize is set to 16.

Loss Setting. For IDPT, we set λ^{Adv} = 0 and λ^{Pcp} = 0. For IDPT-GAN, we empirically set λ^{Adv} = 0.005 and λ^{Pcp} = 0.1.

4.3 Comparisons with State-of-the-Art Methods

Comparison of PSNR and SSIM scores with state-of-the-art FSR methods. We compare our proposed method with several state-of-the-art FSR methods such as SRResNet [Ledig *et al.*, 2017], SISN [Lu *et al.*, 2021], SPARNet [Chen *et al.*, 2020], SwinIR [Liang *et al.*, 2021], DIC [Ma *et al.*, 2020], and DICGAN [Ma *et al.*, 2020] quantitatively.[1] Table 1 presents the quantitative results on Helen and CelebA respectively. It is obvious that IDPT obtains the best PSNR and SSIM scores on both datasets and significantly outperforms other FSR methods by a large margin. Compared to two recent state-of-the-art FSR methods (*i.e.*, SwinIR and DIC) in PSNR value, IDPT outperforms SwinIR by **0.46dB** on Helen and **0.41dB** on CelebA, while surpasses DIC by **1.02dB** on Helen and **0.18dB** on CelebA.

Since the target of IDPT-GAN is to produce visually realistic texture, it could induce the decrease of quantitative values. Even so, IDPT-GAN also gets comparable performance in both PSNR and SSIM. This shows that our IDPT-GAN can

[1]The experiments are implemented on Mindspore and Pytorch.

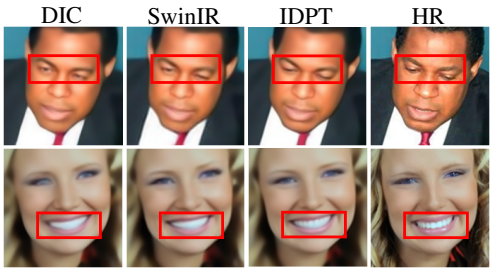

Figure 3: Visual details of the restored images from different FSR methods. IDPT restores correct eyes details and sharper details on lip and teeth.

maintain pixel-level consistency while improving the perceptual quality of generated results.

Visualized comparison with state-of-the-art FSR methods. Some generated results of different FSR methods are visualized in Figure 2. For face SR tasks, it is difficult to recover the details of facial components, especially when the human face has large variations on expressions. It can be seen that IDPT can produce clearer face images and vivid details on key facial components. Figure 3 shows more comparison results on image details. Specifically, we see IDPT recovers correct textures on eyes while other methods can not generate pleasant details. Besides, IDPT can recover sharper details on lip and teeth, while other methods produce blurry results.

Moreover, as shown in Figure 2, IDPT-GAN could generate more realistic and stable results with plausible details when compared with DICGAN. It also demonstrates the effectiveness of the proposed method.

4.4 Ablation Study

In this section, we conduct ablation studies on the Helen dataset to verify the effectiveness of the proposed method.

Effectiveness of the dual pyramid Transformer structure. Instead of straightly stacking cascaded feature reconstruction blocks, we propose the dual pyramid Transformer structure, which hierarchically reconstructs the high-frequency features by the coarse-to-fine mechanism. The number of STMs in each DPTB is set to 2^{i-1}, where i denotes the depth of the stage. We implement an ablation study to confirm its effectiveness. For comparison, we fix depths of STMs in all the DPTBs to 2, 4, and 8 respectively. The results are shown in Figure 4. As the depth of STMs increases, the growth of PSNR is not obvious and it could not obtain superior results when compared to the proposed pyramid structure. For example, when the depth is set to 8, our method outperforms it by **0.22dB**, while the number of parameters can be also reduced significantly.

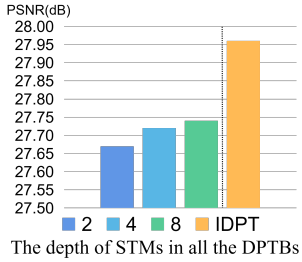

Figure 4: Results of straightly stacking STMs in all the DPTBs and the proposed dual pyramid structure (with yellow) in IDPT.

Name. of *methods*	IDPT	w/o BPFE	SwinIR
PSNR (dB)	**27.96**	27.66	27.50

Table 2: Ablation study of the Bottom Pyramid Feature Extractor.

ES	MS	PSNR (dB)
		27.75
	✓	27.84
✓		27.88
✓	✓	**27.96**

Table 3: Quantitative comparison of combining different components of the Fusing Modulation Module.

Effectiveness of the Fusing Modulation Module. The Fusing Modulation Module (FMM) contains two core components. The extraction stage (ES) is utilized to extract the coarse textures, while the modulation stage (MS) is employed to refine the detailed textures by the guidance of corresponding coarse features. Table 3 shows the comparison results when we combine different components in FMM. If we remove the whole module and simply use concatenation to replace it, the PSNR value decreases to the baseline of 27.75dB. If we separately introduce ES and MS to the neural network, it gains an improvement of **0.13dB** and **0.09dB** respectively. The whole FMM could remarkably improve the performance by **0.21dB**. All experiments validate the effectiveness of our proposed FMM.

Effectiveness of the Bottom Pyramid Feature Extractor. In our proposed method, BPFE constructs the connection of the dual pyramid structure and explore the deep-level detailed features from the coarse ones. In our ablation study, we remove BPFE and use a convolutional layer to substitute the connection. As shown in Table 2, BPFE introduces an improvement of **0.3dB** in PSNR value, which demonstrates the effectiveness of BPFE. In addition, after removing BPFE, our method still performs better than the second-best comparison approach (*i.e.*, SwinIR), which illustrates the superiority of our dual pyramid structure.

5 Conclusion

In this paper, we propose a Transformer-based approach called IDPT for the face super-resolution task. The designed pyramid encoder/decoder Transformer architecture could effectively enrich the relationship between shallow-layer coarse features and deep-layer detailed features. The multi-layer fusing modulation modules are also devised to refine the detailed features by the guidance of coarse ones at various spatial dimensions. Experimental results on CelebA and Helen datasets show the superiority of our proposed method.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grant No. 62002283, U21B2048, 62102307 and 62076195, the National Key Research and Development Project of China under Grant No. 2020AAA0105600, and the Fundamental Research Funds for the Central Universities. We also thank Mindspore for the support of this work.[2]

References

[Cao *et al.*, 2021] Jiezhang Cao, Yawei Li, Kai Zhang, and Luc Van Gool. Video super-resolution transformer. *arXiv preprint arXiv:2106.06847*, 2021.

[Chen *et al.*, 2018] Yu Chen, Ying Tai, Xiaoming Liu, Chunhua Shen, and Jian Yang. FSRNet: End-to-end learning face super-resolution with facial priors. In *CVPR*, pages 2492–2501, 2018.

[Chen *et al.*, 2020] Chaofeng Chen, Dihong Gong, Hao Wang, Zhifeng Li, and Kwan-Yee Wong. Learning spatial attention for face super-resolution. *IEEE Transactions on Image Processing*, 30:1219–1231, 2020.

[Dosovitskiy *et al.*, 2021] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, and et.al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

[Jiang *et al.*, 2021] Junjun Jiang, Chenyang Wang, Xianming Liu, and Jiayi Ma. Deep learning-based face super-resolution: A survey. *ACM Computing Surveys*, 55(1):1–36, 2021.

[Khan *et al.*, 2021] Salman Khan, Muzammal Naseer, Munawar Hayat, and et al. Transformers in vision: A survey. *arXiv preprint arXiv:2101.01169*, 2021.

[Kim *et al.*, 2019] Deokyun Kim, Minseon Kim, Gihyun Kwon, and et al. Progressive face super-resolution via attention to face landmark. In *BMVC*, 2019.

[Le *et al.*, 2012] Vuong Le, Jonathan Brandt, Zhe Lin, and et al. Interactive facial feature localization. In *ECCV*, pages 679–692. Springer, 2012.

[Ledig *et al.*, 2017] Christian Ledig, Lucas Theis, Ferenc Huszár, and et al. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, pages 4681–4690, 2017.

[Liang *et al.*, 2021] Jingyun Liang, Jiezhang Cao, Guolei Sun, and et al. SwinIR: Image restoration using swin transformer. In *CVPR*, pages 1833–1844, 2021.

[Liu *et al.*, 2015] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, pages 3730–3738, 2015.

[Liu *et al.*, 2017] Ling Liu, Fan Wang, Ke Zhou, Nai Ding, and Huan Luo. Perceptual integration rapidly activates dorsal visual pathway to guide local processing in early visual areas. *PLoS biology*, 15(11):e2003646, 2017.

[Liu *et al.*, 2021] Ze Liu, Yutong Lin, Yue Cao, and et al. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pages 10012–10022, 2021.

[Lu *et al.*, 2021] Tao Lu, Yuanzhi Wang, Yanduo Zhang, and et al. Face hallucination via split-attention in split-attention network. In *ACM MM*, pages 5501–5509, 2021.

[Ma *et al.*, 2020] Cheng Ma, Zhenyu Jiang, Yongming Rao, and et al. Deep face super-resolution with iterative collaboration between attentive recovery and landmark estimation. In *CVPR*, pages 5569–5578, 2020.

[Shi and Zhao, 2019] Jingang Shi and Guoying Zhao. Face hallucination via coarse-to-fine recursive kernel regression structure. *IEEE Transactions on Multimedia*, 21(9):2223–2236, 2019.

[Shi *et al.*, 2018] Jingang Shi, Xin Liu, Yuan Zong, Chun Qi, and Guoying Zhao. Hallucinating face image by regularization models in high-resolution feature space. *IEEE Transactions on Image Processing*, 27(6):2980–2995, 2018.

[Wu *et al.*, 2018] Xiang Wu, Ran He, Zhenan Sun, and Tieniu Tan. A light cnn for deep face representation with noisy labels. *IEEE Transactions on Information Forensics and Security*, 13(11):2884–2896, 2018.

[Yang *et al.*, 2021] Tao Yang, Peiran Ren, Xuansong Xie, and Lei Zhang. GAN prior embedded network for blind face restoration in the wild. In *CVPR*, pages 672–681, 2021.

[Yu and Porikli, 2016] Xin Yu and Fatih Porikli. Ultra-resolving face images by discriminative generative networks. In *ECCV*, pages 318–333. Springer, 2016.

[Yu *et al.*, 2018] Xin Yu, Basura Fernando, Bernard Ghanem, Fatih Porikli, and Richard Hartley. Face super-resolution guided by facial component heatmaps. In *ECCV*, pages 217–233, 2018.

[Yu *et al.*, 2022] Zitong Yu, Yuming Shen, Jingang Shi, Hengshuang Zhao, Philip Torr, and Guoying Zhao. Physformer: Facial video-based physiological measurement with temporal difference transformer. In *CVPR*, 2022.

[Yuan *et al.*, 2021] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zihang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *ICCV*, pages 558–567, 2021.

[Zhang *et al.*, 2016] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016.

[Zhang *et al.*, 2018] Kaipeng Zhang, Zhanpeng Zhang, Chia-Wen Cheng, Winston Hsu, Yu Qiao, Wei Liu, and Tong Zhang. Super-identity convolutional neural network for face hallucination. In *ECCV*, pages 183–198, 2018.

[2]https://www.mindspore.cn/