Multiprior Learning via Neural Architecture Search for Blind Face Restoration

Yanjiang Yu, Puyang Zhang, Kaihao Zhang, Wenhan Luo, Senior Member, IEEE, and Changsheng Li

Abstract—Blind face restoration (BFR) aims to recover high-quality (HQ) face images from low-quality (LQ) ones and usually resorts to facial priors for improving restoration performance. However, current methods still suffer from two major difficulties: 1) how to derive a powerful network architecture without extensive hand tuning and 2) how to capture complementary information from multiple facial priors in one network to improve restoration performance. To this end, we propose a face restoration searching network (FRSNet) to adaptively search the suitable feature extraction architecture within our specified search space, which can directly contribute to the restoration quality. On the basis of FRSNet, we further design our multiple facial prior searching network (MFPSNet) with a multiprior learning scheme. MFPSNet optimally extracts information from diverse facial priors and fuses the information into image features, ensuring that both external guidance and internal features are reserved. In this way, MFPSNet takes full advantage of semantic-level (parsing maps), geometric-level (facial heat maps), reference-level (facial dictionaries), and pixel-level (degraded images) information and, thus, generates faithful and realistic images. Quantitative and qualitative experiments show that the MFPSNet performs favorably on both synthetic and real-world datasets against the state-of-the-art (SOTA) BFR methods. The codes are publicly available at: https://github.com/YYJianG/MFPSNet.

Index Terms—Blind face restoration (BFR), facial prior-guided restoration, neural architecture search (NAS).

I. INTRODUCTION

Blind face restoration (BFR) recovers high-quality (HQ) images from low-quality (LQ) images with unknown degradation, such as blur [1], [2], noise [3], [4], JPEG compression artifacts [5], low-resolution [6], [7], or the combination of them (full degradation) [8], [9], [10], [11]. It has been widely applied to many real-world scenarios, such as old photograph renovation [12], LQ face recognition [13], face detection [14], and face reconstruction [15].

Compared with general image restoration, BFR can make use of abundant facial prior knowledge, such as parsing maps [9], [13], facial heat maps [6], and face reference priors [16], [17], [18], to recover details even if the images are severely degraded. Therefore, most BFR methods focus on designing various networks for learning an effective mapping from LQ images to HQ images and exploiting external guidance from facial priors [6], [8], [9], [18], [19]. To effectively extract image features with the increase of the image scale, they usually adopt a keep-upsampling architecture and deepen their neural network to achieve better performance. However, cutting-edge studies [20] have revealed that deeper architectures designed with extensive hand tuning for feature transformation do not necessarily guarantee optimal performance, though it takes a lot of workload. Moreover, facial priors provide different kinds of guidance for BFR, e.g., the parsing maps provide face structure information, and the facial heat maps contain shape and location information of facial components. These priors are pivotal to the image quality but lack guidance for generating details. The reference priors help recover detail-rich images by introducing abundant facial details from HQ-guided face images. However, the generated images are prone to exhibit low fidelity and deviate from the ground truth. Although there exists complementary information among different facial priors, how to extract that information and fuse them together to generate HQ face images with accurate global structures and local details is still an open question.

In this work, to explore the optimal architecture for the BFR task, we first propose a face restoration searching network (FRSNet), bringing more network-level variations in image feature extraction. In most BFR methods, HQ images are generated via encoder–decoder architectures, or recovered in a coarse-to-fine manner with the progressive improvement of the image feature scale. Therefore, the decoding process is inherently important for restoration performance. Instead of simply using a keep-upsampling architecture in the decoding process as in previous face restoration methods [6], [8], [9], [16], [19], we design an architecture search approach that aims to find the optimal network architecture for BFR. To be specific, we introduce neural architecture search (NAS) to search an architecture from a hierarchical upsampling search space (USS) tailored for BFR, which can contain multiple network-level candidate upsampling paths and cell-level candidate operations. Furthermore, we devise a degradation disentanglement strategy in the searching process to reveal the intrinsic correlations between full degradation and single degradation so as to improve the search performance.

On the basis of FRSNet, we turn our attention to multiple prior learning and propose a novel multiple facial prior searching network (MFPSNet), as shown in Fig. 1. Multiple prior learning can be considered as a multi-information extraction...
In this section, we review the most relevant studies on traditional (nonblind) face restoration, BFR, and previous NAS approaches.

A. Face Restoration

Natural image restoration is a comprehensive research field, which includes a variety of tasks, such as image deblurring [1], [13], [21], [22], [23], image denoising [3], [24], [25], [26], image artifact removal [5], [27], and image super-resolution [28], [29], [30], [31], [32], [33], [34], [35]. As an important and high-profile branch of image restoration, traditional (nonblind) face restoration methods [33], [36], [37], [38] have been proposed. Cao et al. [39] propose an attention-aware method to add details to different areas of face images progressively. Shen et al. [13] remove motion blur by learning facial semantic information. Dong et al. [5] adopt transfer learning to reduce artifacts in face images.

Despite the success of traditional face restoration methods, in most real-life scenarios, the degradation types of the real-world images are usually unpredictable. Sometimes, the degradation of images is a combination of several certain degradation types, which leads to limited generalization ability. In this way, our model avoids extensive hand tuning on the network structure design and achieves better restoration performance.

In summary, this article makes the following contributions.

1) We design a hierarchical USS and redesign a candidate operation set to automatically discover the optimal structures. To the best of our knowledge, this is the first attempt to design an architecture search algorithm for BFR. Among the network, a degradation disentanglement strategy is devised to reveal the intrinsic correlations of multiple degradation factors for improving restoration performance.

2) We further propose an MFPSNet to effectively extract multiple prior features and integrate them into image features in a better manner. Different from traditional NAS frameworks selecting a single architecture, the proposed method simultaneously performs searching over the structures of several subnetworks for BFR. In this way, the searched architecture can extract multiple prior features and, thus, help recover HQ facial images effectively. The proposed method is also flexible in incorporating more priors as the field evolves.

3) Extensive experiments demonstrate that the proposed methods outperform the state-of-the-art (SOTA) BFR methods on both synthetic and real-world testing datasets. Both the code and searched results are publicly available.
prior [17], HQ facial components [16], HQ-guided images [18], [19], [44], or pretrained HQ-image-generating generative adversarial networks (GANs) [10], [11]. Based on the powerful pretrained face image generator StyleGAN [45], the methods [10], [11] can generate HQ face images with high resolution. Although reference priors provide abundant texture and details for the restoration process, the recovered images show inconsistency with their input versions and are significantly different from the ground truth.

Facial priors provide various information, which has been verified to be helpful for BFR. However, there is no unified framework that can leverage multiple facial priors. In this article, one of our goals is to study how to effectively extract and fuse these priors to guide face restoration with accurate geometric and semantic details.

C. Neural Architecture Search

NAS [46] aims to automatically search for high-performance network architectures and has been applied to many tasks [47], [48], [49], [50]. Early NAS methods find better architectures at the cost of expensive computational load (e.g., thousands of GPU days). To overcome this, recent works propose their search method to reduce search space complexity [51], [52], [53] or improve search efficiency [47], [54], [55]. Cutting-edge approaches [52], [55] search for a repeatable cell structure instead of searching the entire network and reduce computation overheads to a great extent.

On this basis, works [20], [43], [56], [57] employing NAS for vision tasks achieved brilliant success. Inspired by the differentiable formulation proposed by a differentiable architecture search method (DARTS) [51], Liu et al. [20] develop a hierarchical search space for semantic segmentation to seek the optimal scale for image representation. Zhang et al. [58] redesign a bigger densely connected search space for semantic segmentation while keeping the searching time constant. For other vision tasks, Wang et al. [59] and Liang et al. [56] design their search space based on the basic image detection pyramidal representations, which are proved to be effective. Cheng et al. [60] design their search space for feature extraction and stereo matching.

Rapid progress on NAS for semantic segmentation or image classification has been witnessed as of late. Compared with that, only a few researchers are trying to tailor search space for image restoration tasks. Zhang et al. [61] propose memory-efficient hierarchical neural architecture search for image denoising (HiNAS) with a memory-efficient hierarchical search space for image denoising, accelerating the search speed. Gou et al. [62] redesign the proposed multi-scale neural architecture search for image restoration (CLEARER) search space consisting of different cell-level search spaces in different modules while keeping computational efficiency.

Although the HiNAS and CLEARER are also NAS-based image restoration methods, as alluded to earlier, BFR is technically different from general image restoration. Image restoration methods usually do not introduce any facial priors, whereas our proposed method innovatively incorporates three different priors in one network, which is vital to BFR. More specifically, we propose a tailored search space for BFR to extract and fuse prior features automatically.

III. OUR APPROACH

In this work, we leverage NAS to derive a powerful face restoration network by automatically searching the architecture. A well-designed USS is designed to help search for suitable architectures. In addition, we propose a training strategy based on the intrinsic correlations of multiple degradation factors enabling our model to disentangle the degradation at the architecture level.

Based on FRSNet, we further propose an MFPSNet to utilize multiple facial priors. A modularized search space is designed to extract features from different priors and choose the way to fuse those features. The modularized search space makes MFPSNet incorporate three or more priors easily.

A. Face Restoration Searching Network

The overview of our FRSNet is depicted in Fig. 2. We first use FRSNet to extract facial features from LQ images and then recover HQ images based on the extracted features. To optimally process image features, we design a USS (shown in Fig. 3) behind the residual-based backbone residual dense block (RDB) and utilize NAS to search the optimal upsampling path and the optimal feature transform operations from the search space. Based on an autoencoder-like architecture, the extracted facial features are mapped to HQ images. Moreover, with the help of our degradation disentanglement strategy, the searched architecture is able to remove complicated degradation and recover more details.

1) Upsampling Search Space: Following recent NAS works [20], [51], [60], [61], we define cell as the smallest repeatable network unit, which involves a set of candidate operations to be selected. The USS introduces a set of grid-arrangement cells and places optional upsample, keep, and downsample

![Fig. 2. Overview of the proposed FRSNet. The residual-based backbone RDB extracts facial features that are processed by a decoding architecture. We utilize the NAS framework to select the optimal decoding architecture from a USS. We illustrate the degradation disentanglement strategy on the left.](image-url)
operations between them to control the transformations of features and the information flow between cells. We redesign the cell and network-level search space to fit the BFR task and employ NAS to allow them to be jointly searched.

a) Cell-level search space: As the elementary unit of the search space, we denote cell at layer \( l \) as \( C_l \). \( C_l \) takes the output of the previous cell \( H_{l-1} \), the output of the cell before the previous cell \( H_{l-2} \) as input, and outputs a tensor \( H_l \). Each cell employs a sequence of \( B \) blocks inside, each of which outputs a tensor \( h_i^l \). The output of the current cell \( H_l \) is the concatenation of the outputs of all the blocks, defined as follows:

\[
H_l = \text{Cell}(H_{l-1}, H_{l-2}) = \text{Concat}\{h_i^l, i \leq B\}. \tag{1}
\]

In the search phase, each block takes the two inputs of the current cell and the outputs of previous blocks \( \{h_1^l, \ldots, h_{l-1}^l\} \) as input. We denote the input of block \( i \) in the current cell as \( I_i^l = \{H_{l-2}, H_{l-1}, h_1^l, \ldots, h_{l-1}^l\} \). Block applies a set of operations (e.g., 2-D convolution and skip connection) to all inputs. The output of the \( i \)th block is calculated as follows:

\[
h_i^l = \sum_{x \in \mathcal{O}} O_{j \rightarrow i}(x_j). \tag{2}
\]

We define \( O_{j \rightarrow i} \) with a continuous relaxation fashion described in [51]

\[
O_{j \rightarrow i}(x_j) = \sum_{o \in \mathcal{O}} \alpha_{j \rightarrow i}^{o} o(x_j) \tag{3}
\]

where \( \mathcal{O} \) denotes the set of candidate operations and \( \alpha_{j \rightarrow i}^{o} \) denotes the normalized weight for operation \( o \).

After the search phase is converged, every block selects the two most important inputs as the searched results. That is,

\[
o_{j \rightarrow i}^{r} = o_{j \rightarrow i}^{\alpha_{j \rightarrow i}^{\text{max}}}; \quad r = \text{argmax} \alpha_{j \rightarrow i}^{o}. \tag{4}
\]

b) Operation selection: The candidate operation selection is vital for search space design. Arbitrarily using a large set of operations leads to a superfluous search space, which is difficult to optimize, and, thus, results in low performance of searched architecture. By observing the widely used operations in restoration tasks, we redesign the set of operations in the pool. Instead of following the same configuration of DARTS [51], we empirically remove the separable convolution and pooling layer, which are infrequently used in the task of BFR. We make selections based on empirical evidence, focusing on 2-D convolution operations that are widely employed in various cutting-edge BFR models [8], [9], [10], [16], [37]. Thus, we include the following candidate operations in \( \mathcal{O} \):

1) 2-D convolution \((3 \times 3)\);
2) 2-D convolution \((5 \times 5)\);
3) 2-D convolution \((7 \times 7)\);
4) skip connection.

We discover that the larger set of operation pool leads to converging in local minima and poorer performance compared with ours. The reason for this particular and counter-intuitive discovery is that the algorithm is prone to select skip connections when the operation pool is large, leading to a low-capacity architecture. Please see the ablation study regarding this for more details.

c) Upsampling network-level search space: Fig. 3 illustrates the overview of our proposed USS. The USS takes low-scale features as input and outputs high-scale features through various candidate upsampling paths. Specifically, there are \( L_U \) layers (\( L_U \) is set to 6 in our paper), and each layer consists of four candidate upsampling rates of \( \{1, 2, 4, 8\} \). We associate a scalar \( \beta \) with each arrow in the search space, representing the weight of the preceding cell’s output. The forward propagation of the network-level search space is described as follows:

\[
H_i^l = \beta_{2s \rightarrow s} C_{cell}(H_{i-1}^{2s}, H_{i-2}^s) + \beta_{s \rightarrow 2s} C_{cell}(H_{i-1}^s, H_{i-2}^{2s}) + \beta_{i \rightarrow \beta_{i \rightarrow 2s}} C_{cell}(H_{i-1}^s, H_{i-2}^{2s}) \tag{5}
\]

where \( s \) denotes the upsampling rate and the scalars \( \alpha \) represent the architecture parameters of each cell. We normalize the scalars \( \beta \) by softmax

\[
\beta_{s \rightarrow 2s} + \beta_{s \rightarrow s} + \beta_{i \rightarrow \beta_{i \rightarrow 2s}} = 1 \quad \forall s, l. \tag{6}
\]

We follow the common practice of halving the channel number when the resolution is doubled and cancel the two-layer “stem” structure to preserve pixel-level information.

Fig. 3. Overview of the proposed USS. We illustrate the network-level search space on the left and the cell-level search space on the right. The number of layers \( L_U \) is set to 6.
We conduct an ablation study in the experimental section to the searched architecture for complex degradation removal. By (9), we can train the model to learn from simple degradation and singly degraded images as follows:

$$Z_{\text{full}} = f(Z_{\text{blur}}, Z_{\text{noise}}, Z_{\text{geo}}, Z_{\text{ref}}).$$

(8)

Recovering a face from a completely degraded image is significantly more challenging than from a single degraded image, presenting a challenge in searching for the optimal network structure. To enhance the effectiveness of network search, we incorporate the single degradation restoration task into the full degradation restoration task. Based on the concept of multitasking learning [63], knowledge transfer between related tasks can facilitate better learning in more challenging tasks. Inspired by this, we aim to search for a network architecture with the ability to disentangle complicated degradation.

Inspired by this, we aim to search for a network architecture with the ability to disentangle complicated degradation. It is worth noting that this strategy does not explicitly disentangle the degradation in feature space, but implicitly disentangle the degradation coupling. To achieve this, in the network search phase, we jointly train our model on fully degraded images and singly degraded images as follows:

$$Z_{\text{full}}, Z_{\text{single}} = f(I_{\text{full}}, I_{\text{single}}).$$

(9)

By (9), we can train the model to learn from simple degradation samples, which can then improve the performance of the searched architecture for complex degradation removal. We conduct an ablation study in the experimental section to demonstrate the effectiveness of searched architecture with and without the degradation disentanglement strategy.

### B. Multiple Facial Prior Search Framework

In BFR, facial priors are important for restoration performance [6], [9], [10], [11], [13], [16]. As aforementioned, one single facial prior may be insufficient to generate HQ images. Therefore, based on FRSNet, we further propose MFPSNet that adds two key modules: prior search module and feature fusion module, as shown in Fig. 4. We design three identical prior search modules as prior-specific encoders to extract features from: parsing maps, facial heat maps, and facial component dictionaries, which provide external semantic, geometric, and reference information, respectively. The fusion module aims at integrating multiple prior features into image features to generate information-rich image features. We introduce a flexible hierarchical search space into the prior search module and the feature fusion module to seek the optimal architectures.

1) **Prior Search Module:** We expect the prior search module to handle various unknown prior information beyond those mentioned in this article. The uncertainty and diversity of prior facial information require the search space of the prior search module to include sufficient potential topological structures, so as to provide the best receptive fields and information extraction methods for different prior information. In each prior search module, we set the network-level search space to require consistent resolution between inputs and outputs.

As the prior search module aims to extract prior features, that the search space includes sufficient topological structures. To aggregate various prior information, we design a fusion module to jointly incorporate multiple prior features into image features. As illustrated in Fig. 4, NAS is utilized to optimize the architecture of the fusion module. We follow the exact configurations of the prior search module, except that we set the number of network-level layers from 2 to 6 to ensure that the search space includes sufficient topological structures. As the prior search module aims to extract prior features, which are functionally similar to the USS, we set the candidate operation in cell-level search space as O, which has been experimentally validated to be effective.

2) **Feature Fusion Module:** To aggregate various prior information, we design a fusion module to jointly incorporate multiple prior features into image features. As illustrated in Fig. 4, NAS is utilized to optimize the architecture of the fusion module. We follow the exact configurations of the prior search module, except that we set the number of network-level layers from 2 to 6 to ensure that the search space includes sufficient topological structures.
layers $L$ to 8, with the purpose of achieving a balance between the computation cost and restoration quality.

C. Loss Function and Optimization

To optimize our network, our loss function is composed of the widely used $L_1$ loss and a perceptual loss [64] as follows:

$$\mathcal{L} = ||\hat{I}_{HQ} - I_{HQ}||_1 + \lambda_{per}||\phi(\hat{I}_{HQ}) - \phi(I_{HQ})||_1$$  \hspace{1cm} (10)

where $\phi$ is the pretrained a convolutional Neural Network based architecture (VGG-19) [65] and $\lambda_{per}$ is a balancing parameter.

Benefiting from the continuous relaxation, we alternately optimize network weights $w$ and architecture parameters $\alpha$ and $\beta$ in model $\Psi$ via stochastic gradient descent. We use the first-order approximation [20] and divide the training set $\mathcal{D}$ into two parts $\mathcal{D}_A$ and $\mathcal{D}_B$ to optimize network weights and architecture parameters, respectively. After the optimization convergence, we decode the cell architecture by retaining the first two operations in each block [51] and then decode the architecture parameters, respectively. After the optimization, we decode the architecture by retaining the first two operations in each block [51] and then decode the architecture by finding the path with the maximum probability [20]. We retrain the decoded architecture from scratch to achieve higher performance. The whole algorithm for the proposed multiple facial prior searching is summarized in Algorithm 1.

Algorithm 1: MFPSNet Searching and Training Process

Input: the training set $\mathcal{D}$, the model $\Psi$ to be searched

Output: the searched architecture $\Psi'$ and the trained parameters of $\Psi'$

1: Initialize $\Psi$;
2: Divide the training set $\mathcal{D}$ into $\mathcal{D}_A$ and $\mathcal{D}_B$;
3: for each iteration do
4: Update network weights $w$ by $\nabla_w \mathcal{L}_{\mathcal{D}_A}(w, \alpha, \beta)$;
5: Update architecture $\alpha, \beta$ by $\nabla_{\alpha,\beta} \mathcal{L}_{\mathcal{D}_B}(w, \alpha, \beta)$;
6: end for
7: Decode the searched architecture $\Psi'$ from $\Psi$;
8: Initialize $\Psi'$;
9: for each iteration do
10: Update weights $w$ of $\Psi'$ by $\nabla_w \mathcal{L}_{\mathcal{D}}(w)$;
11: end for

IV. EXPERIMENTS

In this section, we first introduce the training and testing datasets and how we synthesize datasets. The implementation details of the proposed method are also introduced. Then, we introduce metrics for quantitative evaluation and compare our methods with several SOTA BFR methods on synthetic and real-world images. In addition, we compare our methods with SOTA NAS-based image restoration methods. Moreover, an ablation study is conducted to validate the modules’ effectiveness and the proposed training strategy. Finally, we discuss the limitations of our proposed MFPSNet and point out potential future studies.

A. Datasets

All methods in this section are trained on one dataset and evaluated on two synthetic and one real-world datasets.

1) Training Datasets: We search and retrain our FRSNet and MFPSNet on the Flickr-Faces-HQ (FFHQ) dataset [45] of 70,000 HQ face images. To build HQ–LQ image samples for training, we resize all images to $128 \times 128 \times 3$ as HQ images and synthesize full degraded LQ images from HQ images using the following degradation model [8], [9], [10], [11]:

$$I_{full} = ((I_{HQ} \otimes k_\sigma) \downarrow_r + n_\delta)_{JPEG_q}$$  \hspace{1cm} (11)

where $\otimes$ represents the convolution operation and $k_\sigma$ denotes the blur kernel with a parameter $\sigma$. $\downarrow_r$ is the downsampling operation with a scale factor $r$. $n_\delta$ indicates the Gaussian noise with a standard deviation $\delta$, and $JPEG_q$ is the JPEG compression operation with a quality factor $q$. We randomly select $\sigma, r, \delta$, and $q$ for each HQ image to generate one LQ image with full degradation and multiple LQ images with single degradation using the mentioned degradation operations for our degradation disentanglement strategy.

2) Synthetic Testing Datasets: For evaluation, we construct two synthetic datasets, widely used CelebA-HQ [66] and CASIA-WebFace [67]. For CelebA-HQ, we randomly select 1300 HQ images, which has no overlap with the FFHQ. We synthesize HQ–LQ image samples in the same way as the training data. For the CASIA-WebFace dataset, we randomly select 1300 HQ images, and the way to generate paired samples is the same as training data.

3) Real-World Testing Datasets: To evaluate the generalization abilities of our methods, we collect 981 real-world LQ face photographs with diverse ages and races from the Internet to conduct a qualitative and quantitative evaluation. Those photographs exhibit complicated degradation and include people with diverse ages and races.

B. Implementation and Searched Architectures

We search the architecture of FRSNet and MFPSNet with 50 epochs, respectively. The batch size is set to 4 due to the GPU memory constraint. The architecture parameters $\alpha$ and $\beta$ are initialized with a standard Gaussian sampler and updated after the tenth epoch to avoid local minimum. We adopt the Adam optimizer with a learning rate 0.0001 and the weight decay 0.001 to optimize $\alpha$ and $\beta$. We adopt the SGD optimizer with a cosine learning rate that decays from 0.001 to 0.0001 to optimize network weights $w$. $\lambda_{per}$ is set to 0.001. The searched architecture from US$\mathcal{S}$ is shown in Fig. 5, and other searched results are in Figs. 6–9. Notably, we only search our network architecture once on the FFHQ dataset and retrain the searched architecture weights on each task individually. Also, the training performance heavily relies on the quality of facial priors, which, however, is difficult to extract from severely degraded images. Therefore, we incorporate facial priors after training ten epochs to ensure the quality of the generated facial priors.

Once the search phase converges, we proceed to decode the searched architecture and retrain it from scratch. Both the FRSNet and MFPSNet are retrained for 50 epochs with a batch size of 8. We use the Adam optimizer with a learning rate of 0.0001 and a weight decay of 0.001.
C. Comparison With SOTA Methods

To evaluate our proposed models, we compare FRNNet and MFPSNet with other SOTA methods on the synthetic and real-world testing datasets. Following [9], [10], and [11], we employ widely used pixelwise metrics peak signal-to-noise ratio (PSNR) [70], structural similarity index (SSIM) [70], and multi-scale SSIM (MS-SSIM) [71] to evaluate the performance on CelebA-HQ with ground truth. In order to verify our model on real-world images without ground truth, we utilize learned perceptual image patch similarity (LPIPS) [72] scores to assess the perceptual realism of the generated faces. In addition,
TABLE I
QUANTITATIVE COMPARISONS WITH THE SOTA METHODS ON THE CelebA-HQ FOR BFR. BOLD AND UNDERLINE INDICATE THE BEST AND THE SECOND-BEST PERFORMANCE

| Methods         | PSNR↑ | SSIM↑ | MS-SSIM↑ |
|-----------------|-------|-------|----------|
| HiFaceGAN       | 17.16 | 0.4032 | 0.7584   |
| PULSE           | 19.64 | 0.4433 | 0.7775   |
| GFP-GAN         | 18.84 | 0.4187 | 0.7907   |
| PSFR-GAN        | 18.36 | 0.4494 | 0.7978   |
| FSRNet          | 19.12 | 0.4330 | 0.7694   |
| DFDNet          | 19.77 | 0.5020 | 0.7984   |
| FRSNet (ours)   | 19.78 | 0.5092 | 0.8246   |
| MFPSNet (ours)  | 20.27 | 0.5137 | 0.8312   |

TABLE II
QUANTITATIVE COMPARISON WITH THE SOTA METHODS ON THE Celeba-HQ FOR THE FOUR FACE RESTORATION SUBTASKS. BOLD AND UNDERLINE INDICATE THE BEST AND THE SECOND-BEST PERFORMANCE

| Task                     | Methods         | PSNR↑ | SSIM↑ | MS-SSIM↑ |
|--------------------------|-----------------|-------|-------|----------|
| Face Deblurring          | Deepdeblur [68] | 21.64 | 0.5932 | 0.8968   |
|                          | DeblurGANv2 [1] | 22.07 | 0.5804 | 0.9212   |
|                          | PSFR-GAN [9]    | 21.77 | 0.5877 | 0.9030   |
|                          | HiFaceGAN [8]   | 22.91 | 0.6273 | 0.9323   |
|                          | FRSNet (ours)   | 22.73 | 0.6173 | 0.9117   |
|                          | MFPSNet (ours)  | 23.17 | 0.6429 | 0.9368   |
| Face Denoising           | RIDNet [26]     | 24.17 | 0.6803 | 0.8892   |
|                          | VDNET [3]       | 26.35 | 0.7593 | 0.9620   |
|                          | PSFR-GAN [9]    | 26.12 | 0.7619 | 0.9621   |
|                          | HiFaceGAN [8]   | 24.03 | 0.6495 | 0.9402   |
|                          | FRSNet (ours)   | 26.46 | 0.7663 | 0.9658   |
|                          | MFPSNet (ours)  | 26.95 | 0.8096 | 0.9768   |
| Face Artifacts Removal   | ARCNN [5]       | 28.56 | 0.8701 | 0.9866   |
|                          | PSFR-GAN [9]    | 27.42 | 0.8330 | 0.9783   |
|                          | HiFaceGAN [8]   | 26.23 | 0.8069 | 0.9760   |
|                          | FRSNet (ours)   | 27.64 | 0.8202 | 0.9792   |
|                          | MFPSNet (ours)  | 27.81 | 0.8216 | 0.9826   |
| Face Super Resolution    | HAN [69]        | 18.03 | 0.4357 | 0.7845   |
|                          | FSRNet [6]      | 20.32 | 0.5442 | 0.8449   |
|                          | PSFR-GAN [9]    | 19.28 | 0.5033 | 0.8319   |
|                          | HiFaceGAN [8]   | 19.66 | 0.5217 | 0.8478   |
|                          | FRSNet (ours)   | 20.47 | 0.5544 | 0.8523   |
|                          | MFPSNet (ours)  | 21.82 | 0.5912 | 0.8835   |

TABLE III
QUANTITATIVE COMPARISONS WITH THE SOTA METHODS ON THE CASIA-WEBFACE FOR BFR. BOLD AND UNDERLINE INDICATE THE BEST AND THE SECOND-BEST PERFORMANCE

| Methods         | PSNR↑ | SSIM↑ | MS-SSIM↑ |
|-----------------|-------|-------|----------|
| HiFaceGAN       | 16.77 | 0.3548 | 0.6865   |
| PULSE           | 19.73 | 0.4502 | 0.7886   |
| GFP-GAN         | 17.90 | 0.4092 | 0.7197   |
| PSFR-GAN        | 18.68 | 0.4410 | 0.7541   |
| FSRNet          | 19.47 | 0.4922 | 0.7478   |
| DFDNet          | 19.99 | 0.5066 | 0.7657   |
| FRSNet (ours)   | 20.64 | 0.5478 | 0.8013   |
| MFPSNet (ours)  | 21.91 | 0.5920 | 0.8531   |

TABLE IV
QUANTITATIVE COMPARISON WITH THE SOTA METHODS ON THE CASIA-WEBFACE FOR THE FOUR FACE RESTORATION SUBTASKS. BOLD AND UNDERLINE INDICATE THE BEST AND THE SECOND-BEST PERFORMANCE

| Task                     | Methods         | PSNR↑ | SSIM↑ | MS-SSIM↑ |
|--------------------------|-----------------|-------|-------|----------|
| Face Deblurring          | Deepdeblur [68] | 22.37 | 0.6379 | 0.8686   |
|                          | DeblurGANv2 [1] | 22.90 | 0.5947 | 0.9171   |
|                          | PSFR-GAN [9]    | 22.48 | 0.6213 | 0.9044   |
|                          | HiFaceGAN [8]   | 17.11 | 0.5007 | 0.8027   |
|                          | FRSNet (ours)   | 23.21 | 0.6562 | 0.9031   |
|                          | MFPSNet (ours)  | 23.47 | 0.6676 | 0.9033   |
| Face Denoising           | RIDNet [26]     | 25.52 | 0.6858 | 0.8703   |
|                          | VDNET [3]       | 27.26 | 0.7977 | 0.9526   |
|                          | PSFR-GAN [9]    | 27.49 | 0.7556 | 0.9594   |
|                          | HiFaceGAN [8]   | 19.66 | 0.6530 | 0.8628   |
|                          | FRSNet (ours)   | 27.90 | 0.8089 | 0.9614   |
|                          | MFPSNet (ours)  | 28.21 | 0.8351 | 0.9700   |
| Face Artifacts Removal   | ARCNN [5]       | 32.28 | 0.9130 | 0.9865   |
|                          | PSFR-GAN [9]    | 30.96 | 0.8860 | 0.9824   |
|                          | HiFaceGAN [8]   | 18.72 | 0.7369 | 0.8587   |
|                          | FRSNet (ours)   | 30.65 | 0.8839 | 0.9822   |
|                          | MFPSNet (ours)  | 31.05 | 0.8865 | 0.9828   |
| Face Super Resolution    | HAN [69]        | 18.83 | 0.4106 | 0.7631   |
|                          | FSRNet [6]      | 21.44 | 0.6380 | 0.8594   |
|                          | PSFR-GAN [9]    | 19.81 | 0.5228 | 0.8192   |
|                          | HiFaceGAN [8]   | 16.21 | 0.4404 | 0.7373   |
|                          | FRSNet (ours)   | 22.01 | 0.6465 | 0.8715   |
|                          | MFPSNet (ours)  | 23.22 | 0.6760 | 0.8954   |

we employ Fréchet inception distance (FID) [73] scores to quantify the statistical distance between the restoration outcomes and a reference HQ face dataset.

We first evaluate our searched architecture on four sub-tasks, including face denoising, face deblurring, face JPEG artifacts removal, and face super-resolution (hallucination). Considering that most restoration methods are proposed for general image restoration instead of face image restoration, for a fair comparison, we compare MFPSNet with both methods designed for BFR tasks, including HiFaceGAN [8] and progressive semantic-aware style transformation for blind face restoration (PSFR-GAN) [9], and task-specific methods designed for general image restoration. Then, to evaluate our methods on the BFR task, we compare ours with BFR methods without extra facial priors: HiFaceGAN and self-supervised photo upsampling via latent space exploration of generative models (PULSE) [37], and BFR methods utilizing facial priors: towards real-world blind face restoration with generative facial prior (GFP-GAN) (GAN prior) [10], PSFR-GAN (parsing map), and blind face restoration via deep multi-scale component dictionaries (DFDNet) (reference prior) [16]. We also compare with the FSRNet (facial heat maps) [6], which uses facial priors for face super-resolution. For a fair comparison, we retrain them on the same training set instead of using their pretrained models, since their datasets are generated from different seeds.

1) Results on Synthetic Dataset: The quantitative results of the BFR task on CelebA-HQ are reported in Table I, and the results of the four restoration sub-tasks on CelebA-HQ are reported in Table II. The results of the BFR task on
Fig. 10. Qualitative comparison on the synthetic test dataset based on CelebA-HQ for BFR. The proposed MFPSNet shows remarkable results for LQ images with severe degradation and different complexions and poses. The generated images show accurate face contours and faithful facial components.

Fig. 11. Qualitative comparison on the synthetic test dataset based on CASIA-WebFace for BFR. The generated results of the proposed MFPSNet are clearer and show higher fidelity.

CASIA-WebFace are reported in Table III and the results of the four restoration sub-tasks on CASIA-WebFace are reported in Table IV. On all the tasks, our FRSNet and MFPSNet achieve the best performance, except the face artifacts removal task. This indicates that our results are closer to the ground truth. In general, our model is designed to handle tasks with heavy degradation. With the guidance of external facial priors, our model obtains information beyond the input images to generate HQ images. For face artifact removal in which the degradation is relatively milder, our method outperforms other BFR methods and is only second to compression artifacts reduction by a deep convolutional network (ARCNN), which is designed specifically for artifacts removal.

The exemplar qualitative results of different methods are shown in Figs. 10 and 11. Suffering from the unstable training of GAN, the StyleGAN-based methods, GFP-GAN, PULSE,
and HiFaceGAN, generate clear but low-fidelity images, and the results are pretty different from the ground truth. Compared with them, our MFPSNet produces photorealistic face images under the guidance of facial priors. Furthermore, instead of introducing a single facial prior as FSRNet, DFDNet, and PSFR-GAN do, our MFPSNet obtains abundant information from multiple facial priors and is able to generate accurate facial contour and facial components, e.g., the position of the jaw and the facial expression, even in the case of severely degraded input images.

2) Results on Real-World Dataset: We qualitatively (Fig. 12) and quantitatively (Table V) compare different BFR methods on the real-world testing dataset. The results of GFPGAN exhibit many artifacts. HiFaceGAN attempts to improve image quality by brightening the images leading to optical inconsistency. PULSE does generate clear and sharper face images that are obviously not the same person as inputs. The FSRNet and PSFR-GAN, which utilize facial heat maps and parsing maps, seem to encounter problems recovering correct facial components (first and second rows). The DFDNet mainly focuses on facial components and, thus, neglects the rest regions. In contrast, the results of our method are most faithful to input images. Our model removes most of the degradation and restores natural texture, validating the effectiveness of our model in natural scenes.

D. Ablation Study

To study the effectiveness of different components of our models, we conduct an ablation study by comparing several variants of FRSNet and MFPSNet.

**TABLE VI**

**Quantitative Comparisons With NAS-Based Image Restoration Method CLEARER**

| Methods          | PSNR↑ | SSIM | MS-SSIM↑ |
|------------------|-------|------|---------|
| CLEARER          | 18.61 | 0.4677 | 0.7799 |
| MFPSNet (ours)   | **21.91** | **0.5920** | **0.8531** |

1) FRSNet Variants: The following are different variants of FRSNet.
2) MFPSNet Variants: The definitions of variants of MFPSNet are listed below.

1) FRSNet-w/o-uss represents FRSNet without the USS.
2) FRSNet-\text{O}_{\text{large}} adopts a larger candidate operation set commonly used in [20] and [51], including eight operations, such as various convolutions, pooling layers, and connection mechanisms.
3) FRSNet-w/o-Dis denotes training FRSNet without the degradation disentanglement strategy.
4) MFPSNet-prior-PHD is the MFPSNet without the fusion module, while all priors are introduced.
5) MFPSNet-random-searched is the MFPSNet retrained from a randomly selected architecture.

3) Analysis: The results in Table VII show that our FPSNet and MFPSNet achieve better performance than their variants. We illustrate sample results of the proposed methods and their variants in Fig. 13. Specifically, without the USS, the performance of FRSNet-w/o-uss is inferior. This is in line with the exemplar results shown in Fig. 13, where the facial details of the results from FRSNet-w/o-uss are lost to a great extent. We suspect that the simple upsampling operation cannot fully utilize the facial detail information contained in image features. The performance of FRSNet-\(O_{large}\) demonstrates that the larger set of operations hurts the performance of restoration. The reason is that the algorithm tends to select skip connections when the operation pool is large, resulting in a low-capacity architecture. Similar observations have been reported in [74]. Compared with the time-consuming manual process of designing neural network structures, NAS significantly reduces the effort required and only requires basic knowledge of the target task's model. By discarding the degradation disentangle strategy, FRSNet-w/o-Dis generates more blurry images and has more artifacts when compared with FRSNet. This implies the validity of degradation disentanglement on architecture search.

In terms of the MFPSNet variants, the restoration performance improves, as facial priors are incrementally introduced, validating the effectiveness of multiple facial prior guidance. As shown in Fig. 13, the MFPSNet-prior-P generates images with a clearer structure compared with FRSNet. MFPSNet-prior-PH brings finer facial components, and MFPSNet-prior-PHD recovers HQ images with diverse facial details making the facial component more realistic. The difference between MFPSNet and MFPSNet-prior-PHD indicates the contribution of the fusion module in aggregating multiple prior information. The result of MFPSNet-prior-P-w/o-nas shows the effectiveness of the prior search module compared with a hand-designed network. MFPSNet-random-searched is the MFPSNet retrained from a randomly selected architecture. The quantitative result of MFPSNet-random-searched is shown in Table VII to evaluate the selected architecture. The MFPSNet-random-searched cannot effectively extract images and prior features, leading to lower performance than the searched architecture.

### E. Comparison With NAS Method

CLEARER [62] is a NAS-based method for nature image restoration. Different from face restoration methods compared above, CLEARER is designed to restore nature images, and it is indeed unfair to compare CLEARER with other face

| Method                      | PSNR↑ | SSIM↑ |
|-----------------------------|-------|-------|
| FRSNet-w/o-uss              | 20.41 | 0.5254|
| FRSNet-\(O_{large}\)       | 20.48 | 0.5356|
| FRSNet-w/o-Dis              | 20.57 | 0.5457|
| FRSNet                      | 20.64 | 0.5478|
| MFPSNet-prior-P-w/o-nas     | 20.72 | 0.5523|
| MFPSNet-prior-P             | 20.92 | 0.5664|
| MFPSNet-prior-PH            | 21.11 | 0.5847|
| MFPSNet-prior-PHD           | 21.72 | 0.5820|
| MFPSNet-random-searched     | 21.12 | 0.5791|
| MFPSNet                     | 21.91 | 0.5920|
restoration methods. To search the best architecture as possible as we can, we train CLEARER under a set of hyperparameters and report the best performance. Even so, as shown in Table VI, our MFPSNet outperforms CLEARER significantly. Some failure examples of CLEARER are provided in Fig. 14. It is observed that the generated images suffer from artifacts and noise and fail to recover finer details as our method. The primary reason is that, CLEARER cannot utilize any facial priors that are vital to BFR.

F. Discussion and Limitations

Two failure cases are shown in Fig. 15. The proposed MFPSNet fails to generate clear contours of sunglasses as well as a clear background when the input image is severely degraded. It is probably because the facial priors we introduce lack the information of ornaments and background. One solution is to introduce priors containing diverse information beyond faces, which can be future work.

G. Computational Complexity and Convergence Analysis

In this section, we analyze the computational complexity and memory cost of the proposed method and its variants mentioned above. Tables VIII and IX show the floating point operations per second (FLOPs) and parameters of the proposed method in the search and retraining stages, respectively. For the proposed MFPSNet, the entire architecture search stage takes about one day on two GTX3090 GPUs. Then, it takes one day on one GPU to retrain the searched architecture. A convergence analysis result is shown in Fig. 16. Please notice that we incorporate facial priors after training ten epochs to ensure the quality of the generated facial priors. We also further carry out the training for longer epochs (60, 80, and 100), but no benefit is observed.

V. Conclusion

In this article, we resort to multiple facial priors to recover HQ images and optimize our network architecture via NAS. Specifically, we propose two novel networks. The first is FRSNet, which optimally upsamples image features through our USS, leading to improved performance. Based on FRSNet, we further propose MFPSNet, which extracts multiple prior features and fuses them, driving the model to recover faithful and realistic face images. Experimental results on both synthetic and real-world datasets demonstrate the advance of the proposed methods in all restoration subtasks and the BFR task.

REFERENCES

[1] O. Kupyn, T. Martyniuk, J. Wu, and Z. Wang, “DeblurGAN-v2: Deblurring (orders-of-magnitude) faster and better,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 8877–8886.
[2] K. Zhang et al., “Deep image deblurring: A survey,” Int. J. Comput. Vis., vol. 130, no. 9, pp. 2103–2130, Sep. 2022.
[3] Z. Yue, H. Yong, Q. Zhao, L. Zhang, and D. Meng, “Variational denoising network: Toward blind noise modeling and removal,” 2019, arXiv:1908.11314.
[4] L. Liu, Q. Feng, C. L. P. Chen, and Y. Wang, “Noise robust face hallucination based on smooth correntropy representation,” IEEE Trans. Neural Netw. Learn. Syst., vol. 33, no. 10, pp. 5953–5965, Oct. 2022.
[5] C. Dong, Y. Deng, C. C. Loy, and X. Tang, “Compression artifacts reduction by a deep convolutional network,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 576–584.
[6] Y. Chen, Y. Tai, X. Liu, C. Shen, and J. Yang, “FSRNNet: End-to-end learning face super-resolution with facial priors,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 2492–2501.
[7] K. Jiang, Z. Wang, P. Yi, T. Lu, J. Jiang, and Z. Xiong, “Dual-path deep fusion network for face image hallucination,” IEEE Trans. Neural Netw. Learn. Syst., vol. 33, no. 1, pp. 378–391, Jan. 2022.
[8] L. Yang et al., “HiFaceGAN: Face renovation via collaborative suppression and replenishment,” in Proc. 28th ACM Int. Conf. Multimedia, Oct. 2020, pp. 1551–1560.
[9] C. Chen, X. Li, L. Yang, X. Lin, L. Zhang, and K. K. Wong, “Progressive semantic-aware style transformation for blind face restoration,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 11891–11900.
A. Bulat and G. Tzimiropoulos, “Super-FAN: Integrated facial landmark localization and super-resolution of real-world low resolution faces in arbitrary poses with GANs,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 109–117.

K. Zhang et al., “Benchmarking ultra-high-definition image super-resolution,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2021, pp. 14749–14758.

H. Huang, R. He, Z. Sun, and T. Tan, “Wavelet-SRNet: A wavelet-based CNN for multi-scale face super resolution,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 1698–1706.

S. Menon, A. Damian, S. Hu, N. Ravi, and C. Rudin, “PULSE: Self-supervised photo upsampling via latent space exploration of generative models,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 2433–2442.

K. Zhang et al., “EDFace-Celeb-M: Benchmarking face hallucination via a million-scale dataset,” 2021, arXiv:2110.05031.

Q. Cao, L. Lin, Y. Shi, X. Liang, and G. Li, “Attention-aware face hallucination via deep reinforcement learning,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 1656–1664.

A. N. Zamani, M. K. Awang, N. Omar, and S. A. Nazeer, “Image quality assessments and restoration for face detection and recognition system images,” in Proc. 2nd Asia Int. Conf. Modelling Simulation (AMS), May 2008, pp. 505–510.

E. Richardson, M. Sela, R. Or-El, and R. Kimmel, “Learning detailed face reconstruction from a single image,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 5553–5562.

Z. Wan et al., “Bringing old photos back to life,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 2744–2754.

C. Dong, C. C. Loy, K. He, and X. Tang, “Image super-resolution using deep convolutional networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2014, pp. 808–815.

K. Zhang, W. Zuo, and L. Zhang, “Beyond a single convolutional neural network,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2015, pp. 1132–1140.

X. Fu, Z.-J. Zha, F. Wu, X. Ding, and J. Paisley, “JPEG artifacts reduction via deep convolutional sparse coding,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 2501–2508.

C. Dong, C. C. Loy, K. He, and X. Tang, “Image super-resolution using deep convolutional networks,” in Proc. IEEE Trans. Pattern Anal. Mach. Intell., vol. 38, no. 2, pp. 295–307, Feb. 2016.

B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, “Enhanced deep residual networks for single image super-resolution,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), Jul. 2017, pp. 1132–1140.

Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, “Image super-resolution using very deep residual channel attention networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), Jul. 2017, pp. 1132–1140.

M. S. M. Sajjadi, B. Scholkopf, and M. Hirsch, “EnhanceNet: Single image super-resolution through automated texture synthesis,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 4491–4500.

X. Wang et al., “ESRGAN: Enhanced super-resolution generative adversarial networks,” in Proc. Eur. Conf. Comput. Vis. (ECCV) Workshops, 2018, pp. 63–70.
X. Zhang et al., “DCNAS: Densely connected neural architecture search for semantic image segmentation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 13951–13962.

N. Wang et al., “NAS-Fcos: Fast neural architecture search for object detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 11940–11948.

X. Cheng et al., “Hierarchical neural architecture search for deep stereo matching.” 2020, arXiv:2010.13501.

H. Zhang, Y. Li, H. Chen, and C. Shen, “Memory-efficient hierarchical neural architecture search for image denoising,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 3654–3663.

Y. Gou, B. Li, Z. Liu, S. Yang, and X. Peng, “CLEARER: Multi-scale neural architecture search for image restoration,” in Proc. Adv. Neural Inf. Process. Syst., vol. 33, 2020, pp. 17129–17140.

S. Vandenhende, S. Georgoulis, W. Van Gansbeke, M. Proesmans, D. Dai, and L. Van Gool, “Multi-task learning for dense prediction tasks: A survey,” 2020, arXiv:2004.13379.

J. Johnson, A. Alahi, and L. Fei-Fei, “Perceptual losses for real-time style transfer and super-resolution,” in Proc. Eur. Conf. Comput. Vis. Cham, Switzerland: Springer, 2016, pp. 694–711.

K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” 2014, arXiv:1409.1556.

T. Karras, T. Aila, S. Laine, and J. Lehtinen, “Progressive growing of GANs for improved quality, stability, and variation,” 2017, arXiv:1710.10196.

D. Yi, Z. Lei, S. Liao, and S. Z. Li, “Learning face representation from scratch,” 2014, arXiv:1411.7923.

S. Nah, T. H. Kim, and K. M. Lee, “Deep multi-scale convolutional neural network for dynamic scene deblurring,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 257–265.

B. Niu et al., “Single image super-resolution via a holistic attention network,” in Proc. Eur. Conf. Comput. Vis. Cham, Switzerland: Springer, 2020, pp. 191–207.

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

Z. Wang, E. P. Simoncelli, and A. C. Bovik, “Multiscale structural similarity for image quality assessment,” in Proc. 37th Asilomar Conf. Signals, Syst. Comput., Nov. 2003, pp. 1398–1402.

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

M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, “GANs trained by a two time-scale update rule converge to a local Nash equilibrium,” in Proc. Adv. Neural Inf. Process. Syst., vol. 30, 2017, pp. 1–12.

A. Zela, T. Elsken, T. Saikia, Y. Marrakchi, T. Brox, and F. Hutter, “Understanding and robustifying differentiable architecture search,” 2019, arXiv:1909.09656.

Yanjiang Yu received the B.E. degree in computer science and technology from the Beijing University of Technology (BJUT), Beijing, China, in 2021. He is currently pursuing the master’s degree in computer science and technology with the Beijing Institute of Technology (BIT), Beijing. His research interests include computer vision and graph neural network.

Puyang Zhang received the B.E. degree in computer science and technology from the Beijing Institute of Technology (BIT), Beijing, China, in 2022, where he is currently pursuing the master’s degree in computer science and technology. His research interests include computer vision and graph neural network.

Kaihao Zhang is currently pursuing the Ph.D. degree with the College of Engineering and Computer Science, Australian National University, Canberra, ACT, Australia. He has more than 30 refereed publications in international conferences and journals, including Conference on Computer Vision and Pattern Recognition (CVPR), International Conference on Computer Vision (ICCV), European Conference on Computer Vision (ECCV), Conference on Neural Information Processing Systems (NeurIPS), Association for the Advancement of Artificial Intelligence (AAAI), and ACM Multimedia Conference (ACMMM), IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE (TPAMI), International Journal of Computer Vision (IJCV), IEEE TRANSACTIONS ON IMAGE PROCESSING (TIP), and IEEE TRANSACTIONS ON MULTIMEDIA (TMM). His research interests focus on computer vision and deep learning.

Wenhan Luo (Senior Member, IEEE) received the B.E. degree from the Huazhong University of Science and Technology, Wuhan, China, in 2009, the M.E. degree from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2012, and the Ph.D. degree from the Imperial College London, London, U.K., in 2016. He worked as a Research Scientist at Tencent, Shenzhen, China, and Amazon, Palo Alto, CA, USA. He is currently an Associate Professor at Sun Yat-sen University, Guangzhou, China. He has published over 70 papers in top conferences and leading journals, including International Conference on Machine Learning (ICML), Conference on Neural Information Processing Systems (NeurIPS), Conference on Computer Vision and Pattern Recognition (CVPR), International Conference on Computer Vision (ICCV), European Conference on Computer Vision (ECCV), Association for Computational Linguistics (ACL), Association for the Advancement of Artificial Intelligence (AAAI), International Conference on Learning Representations (ICLR), IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE (TPAMI), International Journal of Computer Vision (IJCV), and IEEE TRANSACTIONS ON IMAGE PROCESSING (TIP). His research interests include several topics in computer vision and machine learning, such as image/video synthesis and image/video quality restoration.

Dr. Luo has been a reviewer, a senior PC member, and a guest editor for several prestigious journals and conferences.

Changsheng Li received the B.E. degree from the University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 2008, and the Ph.D. degree in pattern recognition and intelligent system from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2013. During his Ph.D., he once studied as a Research Assistant with The Hong Kong Polytechnic University, Hong Kong, from 2009 to 2010. He worked with IBM Research, Beijing; Alibaba Group, Beijing; and UESTC. He is currently a Professor with the Beijing Institute of Technology, Beijing. He has more than 60 refereed publications in international journals and conferences, including IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, International Journal of Computer Vision, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, IEEE TRANSACTIONS ON COMPUTERS, IEEE TRANSACTIONS ON MULTIMEDIA, Pattern Recognition (PR), Conference on Computer Vision and Pattern Recognition (CVPR), Knowledge Discovery and Data Mining (KDD), Association for the Advancement of Artificial Intelligence (AAAI), International Joint Conference on Artificial Intelligence (IJCAI), Conference on Information and Knowledge Management (CIKM), ACM Multimedia Conference (MM), and ACM International Conference on Multimedia Retrieval (ICMR). His research interests include machine learning, data mining, and computer vision.

Dr. Li won the National Science Fund for Excellent Young Scholars in 2021.