RestoreFormer: High-Quality Blind Face Restoration from Undegraded Key-Value Pairs

Zhouxia Wang[1] Jiawei Zhang[2] Runjian Chen[1] Wenping Wang[1] Ping Luo[1*]
[1] The University of Hong Kong, [2] SenseTime Research

Figure 1. Comparisons with state-of-the-art face restoration methods on some degraded real-world images. The restored results of our RestoreFormer contain more texture details and complete structures which make our results more natural and authentic.

Abstract

Blind face restoration is to recover a high-quality face image from unknown degradations. As face image contains abundant contextual information, we propose a method, RestoreFormer, which explores fully-spatial attentions to model contextual information and surpasses existing works that use local operators. RestoreFormer has several benefits compared to prior arts. First, unlike the conventional multi-head self-attention in previous Vision Transformers (ViTs), RestoreFormer incorporates a multi-head cross-attention layer to learn fully-spatial interactions between corrupted queries and high-quality key-value pairs. Second, the key-value pairs in ResotreFormer are sampled from a reconstruction-oriented high-quality dictionary, whose elements are rich in high-quality facial features specifically aimed for face reconstruction, leading to superior restoration results. Third, RestoreFormer outperforms advanced state-of-the-art methods on one synthetic dataset and three real-world datasets, as well as produces images with better visual quality. Code is available at https://github.com/wzhouxiff/RestoreFormer.git.

1. Introduction

Blind face restoration aims at restoring a high-quality face from a degraded one that has suffered from complex and diverse degradations, such as down-sampling, blur, noise, compression artifact, *etc*. Since the degradations are unknown in the real world, restoration is a challenging task.

Although there are some works [3, 18, 39] tending to restore high-quality face only based on the information in the degraded one, most of the existing works have demonstrated that priors play a critical role in blind face restoration. These priors include geometric priors [5,7,21,32,41,42,46], references [10, 24, 26, 27], and generative priors [14, 29, 35, 37]. Geometric priors can be landmarks [7, 21], facial pars-

*This work is supported by the General Research Fund of HK No.27208720 and 17212120.

ing maps [5, 32], or facial component heatmaps [41]. They are considered to be helpful to reconstruct the facial structure. However, since most of them are estimated from the corrupted faces, their performance is restricted by the quality of the corrupted inputs. Reference priors are from high-quality exemplars [10,26,27] or facial component dictionaries [24]. Whereas, the high-resolution exemplars with the same identity of the degraded image are not always accessible and the existing dictionaries-based methods only consider facial components, *e.g*. eyes, mouth, and nose. Generative priors encapsulated in a well-trained high-quality face generator are also adopted in blind face restoration. By exploring an appropriate latent vector from the latent space of a generator [14, 29] or straightly projecting the degraded face into the latent space [35, 37], their generators are possible to generate a high-quality face with realness.

In these prior-based works, there are two sources of information: the degraded face with identity information and the priors with high-quality facial details. For restoring faces with realness and fidelity, it is important to fuse these two kinds of information. Most of the existing arts simply combine them by concatenation [10,26,27]. Also, there exist works [5,24,37] proposing to fuse these two kinds of information by Spatial Feature Transformer (SFT) [38].However, SFT fuses the information pixel-wisely which neglects the abundant facial context and ends up with sub-optimal restored results. Therefore, we propose a RestoreFormer, which aims for exploring fully-spatial attentions to globally model contextual information and finally transforms the feature from the degraded face into another one close to the ground-truth face feature according to its corresponding high-quality facial priors. Different from existing ViTs works [4, 6, 11, 47] that tend to implement fully-spatial attentions with multi-head self-attention, our RestoreFormer proposes a multi-head cross-attention layer. Specifically, it takes the features of a corrupted face as queries while their key-value pairs are from high-quality facial priors. By globally and spatially incorporating the corrupted facial features with their corresponding high-quality priors, the proposed method can simultaneously restore a face with realness and fidelity.

Besides, the high-quality dictionary (denoted as HQ Dictionary) proposed in this paper is a reconstruction-oriented one. It is learned from plenty of undegraded faces by a high-quality face generation network motivated by the idea of vector quantization [30]. Therefore, it is rich in high-quality facial details that are learned for face restoration. Compared to the previous Component Dictionaries proposed by Li *et al*. [24], whose elements are features of face components generated from amounts of high-quality faces with an off-line approach, our HQ Dictionary has two advantages: (1) HQ Dictionary owns rich and diverse details specifically aimed for high-quality face reconstruction, while the priors generated with an off-line recognition-oriented model, such as VGG [33], may not have such abilities. (2) HQ Dictionary involves all the areas of a face while the Component Dictionaries [24] only provide priors for eyes, nose, and mouth which restrict the ability for face restoration.

In conclusion, our main contributions are as follows:

- We propose a RestoreFormer to learn fully-spatial interactions between corrupted queries and high-quality key-value pairs which can attain a high-quality face with realness and fidelity from a degraded face.
- We learn a new HQ Dictionary as priors in RestoreFormer. Its reconstruction-oriented property plays a critical role in face restoration.
- Extensive experiments show that our RestoreFormer outperforms advanced state-of-the-art methods on both synthetic and real-world datasets, as well as restores faces with better visual quality.

2. Related Works

Blind Face Restoration Blind face restoration aims at restoring high-quality faces from complex and unknown degradations. Previous works have shown that additional priors play a critical role in this task and they can be coarsely categorized into three types: geometric priors [5, 7, 21, 25, 32, 41, 42, 46], references [10, 24, 26, 27], and generative priors [14, 29, 35, 37].

The methods based on geometric priors tend to progressively restore faces with landmark heatmaps [7, 21] or facial component heatmaps [5, 32]. Since these geometric priors are mainly generated from low-quality faces, the corrupted face limits the performance of restoration. On the other hand, reference-based works need the references to be in the same identity with the degraded face, which is not always accessible [10, 26, 27]. Although Li *et al*. [24] alleviate this constraint by collecting component dictionaries consisting of high-quality facial component features as general references, the facial details in these component dictionaries are limited since they are extracted with an off-line recognition-oriented model and only focus on some facial components. Besides, some works tend to exploit the generative priors encapsulated in a high-quality face generation model for blind face restoration. They implement it by exploring a latent vector with an expensive target-specific optimization [29] or projecting the degraded face into the latent space directly [35, 37]. As they [29, 35] fail to consider the identity information during training, their restored lack fidelity. Although Wang *et al*. [37] combine their generative priors with the degraded face with a spatial feature transformer layer, the locally combining method ignores the rich facial context in the face image

Vision Transformer Transformer is a kind of deep neural network originally used in natural language processing field [2, 9, 34]. Due to its competitive representation ability, it begins to be applied to computer vision tasks,

such as recognition [11], detection [4, 47], and segmentation [36]. The low-level vision tasks also get benefits from it in [6,12,31,40,44,45]. Chen *et al*. [6] exploits the advantage of the transformer on large scale pre-training to construct a complex model covered several image processing tasks, such as denoise, deraining, and super-resolution. Esser *et al*. [12] apply the transformer to generate a high-resolution image by predicting a sequence of codebook-indices of their encoders, which makes full use of the strong representative capacity of the transformer within an acceptable computational resource. In [45], Zhu *et al*. adopt the transformer to obtain the global structure of the face which is helpful for photo-sketch synthesis.

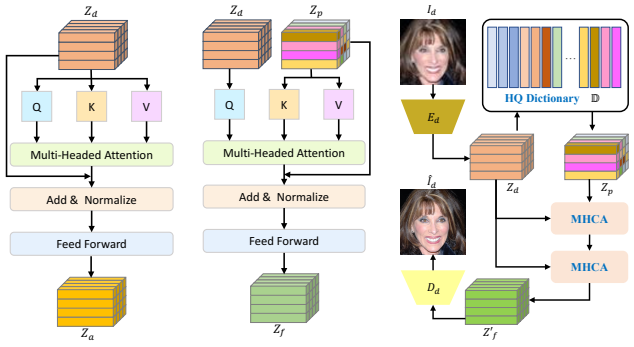
3. Methodology

This section introduces the proposed RestoreFormer for restoring high-quality faces from unknown degradations with an HQ Dictionary consisting of reconstruction-oriented high-quality priors. The whole pipeline is shown in Figure 2 (c). An encoder $\mathbf{E}_d$ is first deployed to extract representation $\boldsymbol{Z}_d$ of the degraded face $\boldsymbol{I}_d$ and its nearest high-quality priors $\boldsymbol{Z}_p$ are fetched from the HQ Dictionary $\mathbb{D}$. Then two consecutive transformers implemented with multi-head cross-attention (denoted as MHCA) are utilized to fuse the features of degraded images and priors. Finally a decoder $\mathbf{D}_d$ is applied on the fused representation $\boldsymbol{Z}'_f$ to restore a high-quality face $\hat{\boldsymbol{I}}_d$. Details of each step will be presented in Sec. 3.1.

To obtain the HQ Dictionary $\mathbb{D}$, we incorporate the idea of vector quantization [30] and propose a high-quality face generation network to learn $\mathbb{D}$ from plenty of undegraded faces. Compared to previous works [24] whose component dictionaries are extracted with an off-line recognition model VGG [33], the priors in $\mathbb{D}$ are reconstruction-oriented and can provide rich facial details for the restoration of degraded faces. The specific procedure of getting the reconstruction-oriented HQ Dictionary will be introduced in Sec. 3.2.

3.1. RestoreFormer

Even though facial image contains abundant global contextual information, *e.g*. eyes and teeth, the existing arts [5, 24, 37] only apply local operators for blind face restoration. Recently, ViT (Vision Transformer) [34] is proposed to consider the contextual information in images. However, most of the ViT-based methods [4, 6, 11, 47] only consider one source of information, *i.e*. the degraded face in our task, by multi-head self-attention (namely MHSA) and it cannot be directly applied into face restoration which needs to combine the information from degraded image and priors. Thus, we propose transformers with the multi-head cross-attention mechanism (MHCA) to fully-spatially fuse two sources of information to restore face with realness and fidelity.In this subsection, we first explain the MHCA by comparing it with MHSA and then give a detailed description of RestoreFormer built upon MHCA.



(a) **MHSA** (b) **MHCA** (c) **RestoreFormer**

Figure 2. **Framework of RestoreFormer**. (a) MHSA is a transformer with multi-head self-attention used in most of previous ViTs [4, 6, 11, 47]. Its queries, keys, and values are from the degraded information $\boldsymbol{Z}_d$. (b) MHCA is a transformer with a multi-head cross-attention used in the proposed RestoreFormer. It is designed to spatially fuse both degraded information $\boldsymbol{Z}_d$ and its corresponding high-quality priors $\boldsymbol{Z}_p$ by taking $\boldsymbol{Z}_d$ as queries while $\boldsymbol{Z}_p$ as key-value pairs. (c) is the whole pipeline of RestoreFormer. An encoder $\mathbf{E}_d$ is first deployed to extract representation $\boldsymbol{Z}_d$ of the degraded face $\boldsymbol{I}_d$ and its nearest high-quality priors $\boldsymbol{Z}_p$ are fetched from the HQ Dictionary $\mathbb{D}$. Then two MHCAs are utilized to fuse the degraded features $\boldsymbol{Z}_d$ and priors $\boldsymbol{Z}_p$. Finally, a decoder $\mathbf{D}_d$ is applied on the fused representation $\boldsymbol{Z}'_f$ to restore a high-quality face $\hat{\boldsymbol{I}}_d$. **The detailed structures of RestoreFormer are in the supplemental materials.**

MHSA. As Figure 2 (a) shown, MHSA used in most of the previous ViTs [4,6,11,47] tends to globally attend contents from $\boldsymbol{Z}_d \in \mathbb{R}^{H' \times W' \times C}$ (H', W' are spatial size of the feature map while C is the number of channels) which is extracted from the degraded input in our task. And the queries $\boldsymbol{Q}$, keys $\boldsymbol{K}$, and values $\boldsymbol{V}$ can be represented as:

$$\boldsymbol{Q} = \boldsymbol{Z}_d\boldsymbol{W}_q + \boldsymbol{b}_q\ ,\ \boldsymbol{K} = \boldsymbol{Z}_d\boldsymbol{W}_k + \boldsymbol{b}_k\ ,\ \boldsymbol{V} = \boldsymbol{Z}_d\boldsymbol{W}_v + \boldsymbol{b}_v, \tag{1}$$

where $\boldsymbol{W}_{q/k/v} \in \mathbb{R}^{C\times C}$ and $\boldsymbol{b}_{q/k/v} \in \mathbb{R}^{C}$ are learnable parameters.

For getting more powerful representations, multi-head attention [34] is adopted on $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$. First, $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$ are separated into N_h blocks along the channel dimension to obtain $\{\boldsymbol{Q}_1, \boldsymbol{Q}_2, \ldots, \boldsymbol{Q}_{N_h}\}$, $\{\boldsymbol{K}_1, \boldsymbol{K}_2, \ldots, \boldsymbol{K}_{N_h}\}$, and $\{\boldsymbol{V}_1, \boldsymbol{V}_2, \ldots, \boldsymbol{V}_{N_h}\}$. For each block, it has $C_h = \frac{C}{N_h}$ channels. Their attention maps can be represented as:

$$\boldsymbol{Z}_i = \text{softmax}(\frac{\boldsymbol{Q}_i\boldsymbol{K}_i^\intercal}{\sqrt{C_h}})\boldsymbol{V}_i, i = 1, 2, \ldots, N_h \tag{2}$$

and the final output of multi-head attention is the concatenation of $\boldsymbol{Z}_i$:

$$\boldsymbol{Z}_{mh} = \underset{i=1,\ldots,N_h}{\text{concat}}\ \boldsymbol{Z}_i. \tag{3}$$

Similar to [34], $\boldsymbol{Z}_{mh}$ is regarded as residual. $\boldsymbol{Z}_{mh}$ and $\boldsymbol{Z}_d$ are added before sending the summation into a normalization layer and a feed forward network sequentially as:

$$\boldsymbol{Z}_a = \text{FFN}(\text{LN}(\boldsymbol{Z}_{mh} + \boldsymbol{Z}_d)), \tag{4}$$

where LN is the layer normalization, FFN is the feed-forward network composed by two convolution layers, and $\boldsymbol{Z}_a$ is the finally globally attended feature map.

MHCA. Different from MHSA, our MHCA aims for spatially fusing the information from the degraded face and its corresponding priors that can respectively provide identity information and high-quality facial details for face restoration. Therefore, as Figure 2 (b) shown, our MHCA takes features $\boldsymbol{Z}_d$ from the degraded face, as queries $\boldsymbol{Q}$, while the keys $\boldsymbol{K}$ and values $\boldsymbol{V}$ are from its high-quality facial priors $\boldsymbol{Z}_p \in \mathbb{R}^{H' \times W' \times C}$:

$$\boldsymbol{Q} = \boldsymbol{Z}_d \boldsymbol{W}_q + \boldsymbol{b}_q \ , \ \boldsymbol{K} = \boldsymbol{Z}_p \boldsymbol{W}_k + \boldsymbol{b}_k \ , \ \boldsymbol{V} = \boldsymbol{Z}_p \boldsymbol{W}_v + \boldsymbol{b}_v, \tag{5}$$

Following multi-head attention in MHSA according to Eq. 2 and Eq. 3, $\boldsymbol{Z}_{mh}$ in MHCA can be estimated similarly. To generate features with more face details, $\boldsymbol{Z}_{mh}$ is added by $\boldsymbol{Z}_p$ before LN and FNN to get the final fused features $\boldsymbol{Z}_f$:

$$\boldsymbol{Z}_f = \mathrm{MHCA}(\boldsymbol{Z}_d, \boldsymbol{Z}_p) = \mathrm{FFN}(\mathrm{LN}(\boldsymbol{Z}_{mh} + \boldsymbol{Z}_p)). \tag{6}$$

RestoreFormer. The whole pipeline of the proposed RestoreFormer based on MHCA is shown in Figure 2 (c). First, a degraded image $\boldsymbol{I}_d$ is sent into an image encoder $\mathbf{E}_d$, which is composed of 12 residual blocks and 5 average poolings, to extract representations $\boldsymbol{Z}_d$. Then we fetch priors from a reconstruction-oriented HQ Dictionary $\mathbb{D} = \{\boldsymbol{d}_m\}_{m=1}^{M} (d_m \in \mathbb{R}^C)$. $\mathbb{D}$ consists of M high-quality facial priors and the learning of the HQ Dictionary will be explained in Sec. 3.2. By finding the most similar priors of feature vectors in $\boldsymbol{Z}_d$ from $\mathbb{D}$, we get the priors $\boldsymbol{Z}_p$:

$$\boldsymbol{Z}_p^{(i,j)} = \underset{\boldsymbol{d}_m \in \mathbb{D}}{\arg\min} \| \boldsymbol{Z}_d^{(i,j)} - \boldsymbol{d}_m \|_2^2, \tag{7}$$

where $\boldsymbol{Z}_p^{(i,j)}$ and $\boldsymbol{Z}_d^{(i,j)}$ indicate the feature vector on the location (i, j) of $\boldsymbol{Z}_q$ and $\boldsymbol{Z}_d$, respectively. $||\cdot||_2$ is the L2-norm.

Given $\boldsymbol{Z}_p$ and $\boldsymbol{Z}_d$, two consecutive MHCAs are applied and we can get a refined representation $\boldsymbol{Z}'_f$ as:

$$\boldsymbol{Z}'_f = \mathrm{MHCA}(\boldsymbol{Z}_d, \mathrm{MHCA}(\boldsymbol{Z}_d, \boldsymbol{Z}_p)). \tag{8}$$

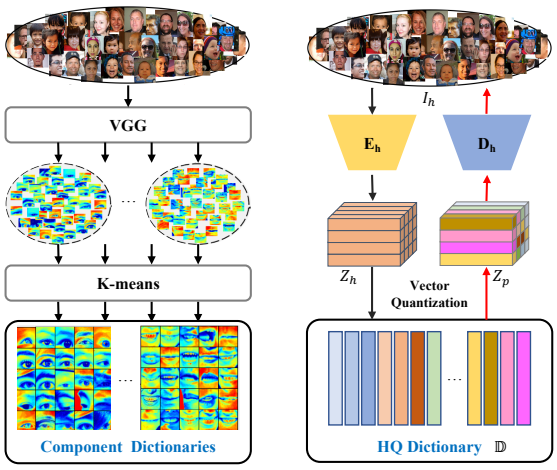
Finally, $\boldsymbol{Z}'_f$ is fed into a decoder $\mathbf{D}_d$ with 12 residual blocks and 5 nearest neighbour upsampling to recover the high-quality image $\hat{\boldsymbol{I}}_d \in \mathbb{R}^{H \times W \times 3}$.

Learning. To train RestoreFormer, our losses involve several aspects, including pixel-level, component-level, and image-level. Following are the detailed discussions.

Pixel-level losses. In pixel-level, we adopt two widely-used losses for face restoration: $L1$ loss and perceptual loss [19, 23]. They are expressed as:

$$\mathcal{L}_{l1} = |\boldsymbol{I_h} - \hat{\boldsymbol{I}}_{\boldsymbol{d}}|_1 \ ; \ \mathcal{L}_{per} = \|\phi(\boldsymbol{I}_h) - \phi(\hat{\boldsymbol{I}}_d)\|_2^2 \tag{9}$$

where $\boldsymbol{I_h}$ is the ground truth high-quality image; ϕ is the pretrained VGG-19 [33] and the feature maps are extracted from $\{conv1, \ldots, conv5\}$.



(a) Component Dictionaries (b) HQ Dictionary

Figure 3. **Comparison of Prior Dictionary**. (a) Component Dictionaries, proposed in DFDNet [24], are off-line generated by a VGG network [33] and clustered with K-means. They only consider eyes, nose, and mouth. (b) HQ Dictionary, proposed in this paper, is learned by a high-quality face generation network incorporating the idea of vector quantization [30]. The high-quality priors in the HQ Dictionary are reconstruction-oriented and provide more facial details for the restoration of degraded faces. Besides, the priors in the HQ Dictionary involve all the facial regions.

Besides, for accurately matching high-quality priors from the HQ Dictionary, we force the extracted features $\boldsymbol{Z}_d$ to approach their selected priors $\boldsymbol{Z}_p$. That is:

$$\mathcal{L}_p = \|\boldsymbol{Z}_p - \boldsymbol{Z}_d\|_2^2. \tag{10}$$

Component-level losses. Since eyes and mouth play an important role in the overview of a face, we also adopt a discrimination loss and feature style loss on the facial areas of eyes and mouth for further enhancing their restored quality. Following [37], we only focus on regions $r \in \{$left eye, right eye, mouth$\}$ and the loss functions are formulated as:

$$\begin{aligned} \mathcal{L}_{disc} &= \sum_r [\log D_r(R_r(\boldsymbol{I_h})) + \log(1 - D_r(R_r(\hat{\boldsymbol{I}}_d)))], \\ \mathcal{L}_{style} &= \sum_r \| \mathrm{Gram}(\varphi(R_r(\boldsymbol{I_h})) - \mathrm{Gram}(\varphi(R_r(\hat{\boldsymbol{I}}_d))\|_2^2, \end{aligned} \tag{11}$$

where $R_r(\cdot)$ is ROI align [15] and φ denotes the multi-resolution features of discriminator D_r trained on region r. Gram denotes the Gram matrix [13] that calculates the feature correlations to measure the style difference.

Image-level losses. The proposed method aims for attaining faces with high realness and fidelity. Therefore, in image-level, we adopt an adversarial loss for improving the realness of the restored face and an identity loss [37] for keeping its fidelity as:

$$\begin{aligned} \mathcal{L}_{adv} &= [\log D(\boldsymbol{I}_h) + \log(1 - D(\hat{\boldsymbol{I}}_d))], \\ \mathcal{L}_{id} &= \|\eta(\boldsymbol{I_h}) - \eta(\hat{\boldsymbol{I}}_{\boldsymbol{d}})\|_2^2, \end{aligned} \tag{12}$$

where D is the discriminator trained on the face image and η denotes the identity feature extracted from a well-trained face recognition ArcFace [8] model.

In the end, with all the loss functions proposed above, the final loss to train RestoreFormer is:

$$\begin{aligned}\mathcal{L}_{RF} = \mathcal{L}_{l1} + \lambda_{per}\mathcal{L}_{per} + \lambda_{p}\mathcal{L}_{p} + \lambda_{disc}\mathcal{L}_{disc} \\ + \lambda_{style}\mathcal{L}_{style} + \lambda_{adv}\mathcal{L}_{adv} + \lambda_{id}\mathcal{L}_{id},\end{aligned} \tag{13}$$

where $\lambda_{...}$ are the weighting factors for different losses.

3.2. HQ Dictionary

In this subsection, we introduce the generation of the HQ Dictionary $\mathbb{D} = \{\boldsymbol{d}_m\}_{m=0}^{M}, d_m \in \mathbb{R}^C$ used in RestoreFormer.

As shown in Figure 3, different from [24] whose component dictionaries are generated from an off-line recognition-orientated feature extractor VGG [33], we aim for getting a reconstruction-oriented high-quality dictionary that can provide richer facial details for face restoration. Therefore, we deploy a high-quality face generation network motivated from vector quantization [30] to learn a high-quality dictionary $\mathbb{D}$ from plenty of undegraded faces.

The framework of this face generation network is shown in Figure 3 (b). First, an encoder $\mathbf{E}_h$ is used to extract the representation $\boldsymbol{Z}_h \in \mathbb{R}^{H' \times W' \times C}$ from a high-quality undegraded image $\boldsymbol{I}_h \in \mathbb{R}^{H \times W \times 3}$. Then rather than decoding $\boldsymbol{Z}_h$ with a decoder $\mathbf{D}_h$ directly, we quantize feature vectors of $\boldsymbol{Z}_h$ by their nearest elements in $\mathbb{D}$ and finally get $\boldsymbol{Z}_p \in \mathbb{R}^{H' \times W' \times C}$:

$$\boldsymbol{Z}_p^{(i,j)} = \underset{\boldsymbol{d}_m \in \mathbb{D}}{\arg\min} \| \boldsymbol{Z}_h^{(i,j)} - \boldsymbol{d}_m \|_2^2, \tag{14}$$

where $\boldsymbol{Z}_p^{(i,j)}$ and $\boldsymbol{Z}_h^{(i,j)}$ are the feature vectors on the position (i, j) of $\boldsymbol{Z}_p$ and $\boldsymbol{Z}_h$, respectively. Taking $\boldsymbol{Z}_p$ as input, the decoder $\mathbf{D}_h$ can reconstruct a high-quality face $\hat{\boldsymbol{I}}_h \in \mathbb{R}^{H \times W \times 3}$. Noted that the structures of $\mathbf{E}_h$ and $\mathbf{D}_h$ are the same with that of $\mathbf{E}_d$ and $\mathbf{D}_d$ in Sec. 3.1.

Learning. The elements $\boldsymbol{d}_m$ in $\mathbb{D}$ are randomly initialized by a uniform distribution. For updating them to capture high-quality facial information, we adopt a dictionary learning algorithm, Vector Quantization (VQ) [30], to move $\boldsymbol{Z}_p$ towards $\boldsymbol{Z}_h$ as:

$$\mathcal{L}'_d = \| \mathrm{sg}[\boldsymbol{Z}_h] - \boldsymbol{Z}_p \|_2^2 \tag{15}$$

where $sg[\cdot]$ denotes the stop-gradient operation. Noted that since $\boldsymbol{Z}_p$ consists of the elements in $\mathbb{D}$ according to Eq. 14, $\mathbb{D}$ is updated through $\boldsymbol{Z}_p$. To keep the encoder $\mathbf{E}_h$ and dictionary $\mathbb{D}$ in the same learning space, a commitment loss [30] is also adopted:

$$\mathcal{L}'_c = \| \boldsymbol{Z}_h - \mathrm{sg}[\boldsymbol{Z}_p] \|_2^2. \tag{16}$$

As the above two losses make $\boldsymbol{Z}_p$ close to $\boldsymbol{Z}_h$ which is extracted from high-quality undegraded image $\boldsymbol{I}_h$, $\boldsymbol{Z}_p$ contains facial detail information which can benefit face restoration. And we consider $\mathbb{D} = \{\boldsymbol{d}_m\}_{m=0}^{M}, d_m \in \mathbb{R}^C$ as facial prior in RestoreFormer.

Besides the two losses for dictionary, an $L1$ loss, a perceptual loss, and an adversarial loss are also applied to the final reconstructed result $\hat{\boldsymbol{I}}_h$ to make sure $\boldsymbol{Z}_p$ has sufficient information to restore high-quality image $\boldsymbol{I}_h$:

$$\begin{aligned}\mathcal{L}'_{l1} = \|\boldsymbol{I}_h - \hat{\boldsymbol{I}}_h\|_1 \ ; \ \mathcal{L}'_{per} = \|\phi(\boldsymbol{I}_h) - \phi(\hat{\boldsymbol{I}}_h)\|_2^2 \\ \mathcal{L}'_{adv} = [\log D(\boldsymbol{I}_h) + \log(1 - D(\hat{\boldsymbol{I}}_h))].\end{aligned} \tag{17}$$

Noted that, since Eq. 14 is non-differentiable, the gradient of $\boldsymbol{Z}_h$ is simply copied from $\boldsymbol{Z}_p$ [30].

The final loss is:

$$\mathcal{L}_{Dict} = \mathcal{L}'_{l1} + \lambda_{per}\mathcal{L}'_{per} + \lambda_{adv}\mathcal{L}'_{adv} + \lambda_d\mathcal{L}'_d + \lambda_c\mathcal{L}'_c, \tag{18}$$

where $\lambda'_{...}$ are the weighting factors.

4. Experiments and Analysis

4.1. Datasets

Training Datasets. The HQ Dictionary is trained on the FFHQ [20] dataset. It contains 70000 high-quality images and all are resized to 512×512. Since the proposed RestoreFormer needs degarded image and high-quality image pairs for training, we synthesize degraded images on FFHQ dataset by the degrading model proposed in [26, 27, 37]:

$$\mathbf{I_d} = \{[(\mathbf{I_h} \otimes \mathbf{k}_\sigma) \downarrow_r + \mathbf{n}_\delta]_{JPEG_q}\} \uparrow_r . \tag{19}$$

Specifically, a high-quality image $\mathbf{I_h}$ is firstly blurred by Gaussian blur kernel $\mathbf{k}_\sigma$ whose sigma is σ. Then, it will be bilinearly downsampled with a scale factor r and added with white Gaussian noise $\mathbf{n}_\delta$ with sigma δ. Finally, a JPEG compression with quality factor q will be adopted to generate the final degraded image.And it will be resized to the same size of $\mathbf{I_h}$ by bilinear upsampling as the degraded input $\mathbf{I_d}$ of our network similar to existing arts [26, 27, 37]. In this paper, σ, r, δ, and q are randomly sampled from $\{0.2 : 10\}$, $\{1 : 8\}$, $\{0 : 20\}$, and $\{60 : 100\}$, respectively.

Testing Datasets. We evaluate our method on a synthetic dataset: CelebA-Test and three real-world datasets: LFW-Test, CelebChild-Test, and WebPhoto-Test. CelebA-Test consists of 3000 images and it is synthesized by applying the degrading described above on the testing set of CelebA-HQ images [28]. For LFW-Test, it consists of the first image of each identity in the validation partition of the original LFW [17] and there are 1711 images. Another two real-world datasets are collected by Wang *et al*. [37] from the Internet. Specifically, CelebChild-Test contains 180 child faces of celebrities and WebPhoto-Test consists of 407 real life faces.

Methods	FID↓	PSNR↑	SSIM↑	LPIPS↓	IDD↓
Input	132.69	24.96	0.6624	0.4989	0.9308
DFDNet [24]	52.92	24.10	0.6092	0.4478	0.7581
PSFRGAN [5]	43.88	24.45	0.6308	0.4186	0.7163
Wan *et al*. [35]	70.21	23.00	0.6189	0.4778	0.8018
PULSE [29]	67.75	21.61	0.6287	0.4657	1.2019
GFP-GAN [37]	42.39	24.46	0.6684	0.3551	0.6034
RestoreFormer	41.45	24.42	0.6404	0.3650	0.5650
GT	43.43	∞	1	0	0

Table 1. Quantitative comparisons on **CelebA-Test**. Our RestoreFormer has better performance based on FID and IDD which indicates the realness and identity preserving property of our method. It also gets a comparable results on PSNR, SSIM, and LPIPS.

Methods	**LFW-Test**	**CelebChild-Test**	**WebPhoto-Test**
Input	137.56	144.42	170.11
DFDNet [24]	62.57	111.55	100.68
PSFRGAN [5]	53.92	106.61	84.98
Wan *et al*. [35]	73.19	115.70	100.40
PULSE [29]	64.86	102.74	86.45
GFP-GAN [37]	49.96	111.78	87.35
RestoreFormer	47.75	101.22	77.33

Table 2. Quantitative comparisons on three **real-world dataset** in terms of FID. RestoreFormer performs the best.

	LFW-Test	**WebPhoto-Test**
Methods	**RestoreFormer**	
DFDNet [24]	15.41%/**84.59%**	28.78%/**71.22%**
PSFRGAN [5]	9.96%/**90.04%**	20.90%/**79.10%**
GFP-GAN [37]	9.89%/**90.11%**	10.40%/**89.60%**

Table 3. User study results on **LFW-Test** and **WebPhoto-Test**. For "a/b", a is the percentage where the compared method is considered better than our RestoreFormer, and b is the percentage where our RestoreFormer is considered better than the compared method.

4.2. Experimental Settings and Metrics

Settings. The size of the input image is $512 \times 512 \times 3$ and the size of $\boldsymbol{Z}_d$ is $16 \times 16 \times 256$. The HQ Dictionary contains $M = 1024$ elements and the length of each element is 256. The batch size is 16 and the weighting factors of the loss function are $\lambda_{per} = 1.0$, $\lambda_p = 0.25$, $\lambda_{disc} = 1.0$, $\lambda_{style} = 2000$, $\lambda_{adv} = 0.8$, $\lambda_{id} = 1.5$, $\lambda_d = 1.0$, and $\lambda_c = 0.25$.

During training, HQ Dictionary is trained by Adam optimizer [22] and the learning rate is set to $7e^{-5}$ at the beginning. Then, the learning rate is decayed by 10 after $6e^5$ iterations. The dictionary is trained until $8e^5$ iterations. We also optimize the RestoreFormer with Adam. Since $\mathbf{E}_d$ and $\mathbf{D}_d$ in RestoreFormer are initialized by $\mathbf{E}_h$ and $\mathbf{D}_h$ for dictionary learning, the learning rate of RestoreFormer is set to $7e^{-6}$ and trained by $6e^4$ iterations.

Metrics. Our evaluation is based on both the realness and fidelity of the restored faces. To measure the realness, except a widely-used non-reference metric FID [16], we also deploy a user study for further evaluating the visual performance of the restored results from the perspective of humans. As for the facial fidelity, we adopt two pixel-wise metrics: PSNR and SSIM and a perceptual metric: LPIPS [43]. Since identity recognition is a more straight and convincing approach for evaluating the fidelity of faces, we introduce an identity distance (denoted as IDD) that is implemented by measuring the distance of the features extracted from ArcFace [8] with angle.

4.3. Comparison with State-of-the-art Methods

To validate the effectiveness of our proposed method on blind face restoration, we compare its performance with several state-of-the-art face restoration methods, including DFDNet [24], PSFRGAN [5], Wan *et al*. [35], PULSE [29], and GFP-GAN [37]. These methods cover different types of priors, *e.g*. reference (DFDNet), geometric priors (PSFRGAN), and generative priors (Wan *et al*., PULSE, and GFP-GAN).

Synthetic Dataset. We first compare our RestoreFormer with other methods on CelebA-Test. The quantitative results of each method are shown in Table 1. Our RestoreFormer has a better performance based on FID and IDD which indicates its restored faces are closer to the real face and have a more similar identity with their ground truth at the same time. It also has comparable results on the pixel-wise and perceptual metrics: PSNR, SSIM, and LPIPS, although they have been proven not that consistent with the subjective evaluation of human beings [1, 23]. As to the visual results, PULSE [29] can generate visually pleasant results in Figure 4. However, it cannot preserve the human identity compared with RestoreFormer. Even though the left eyebrow and eyeglasses can be detected by DFDNet [24] and GFP-GAN [37] in the first and second row (blue box) of Figure 4, they are only partially reconstructed. This may be because only local information is considered when fusing degraded information and priors. With the help of MHCA, RestoreFormer can reconstruct the eyebrow and eyeglasses better in Figure 4. Eyeglasses also cannot be restored by PSFGAN [5] in Figure 4. This is because its estimated heatmap (upper-right corner of PSFGAN [5]), from the degraded input, is inaccurate.

Real-world Datasets. We also apply our RestoreFormer on three real-world datasets: LFW-Test, CelebChild-Test, and WebPhoto-Test for evaluating the generalization of the proposed method. Their quantitative results are shown in Table. 2. Due to the reconstruction-oriented HQ Dictionary and powerful MHCA fusion block, our method performs better in all three real-world datasets based on FID. The visual results of the three real-world datasets shown in Figure 5 also show that RestoreFormer can also robustly restore faces with more details, fewer artifacts, and keep

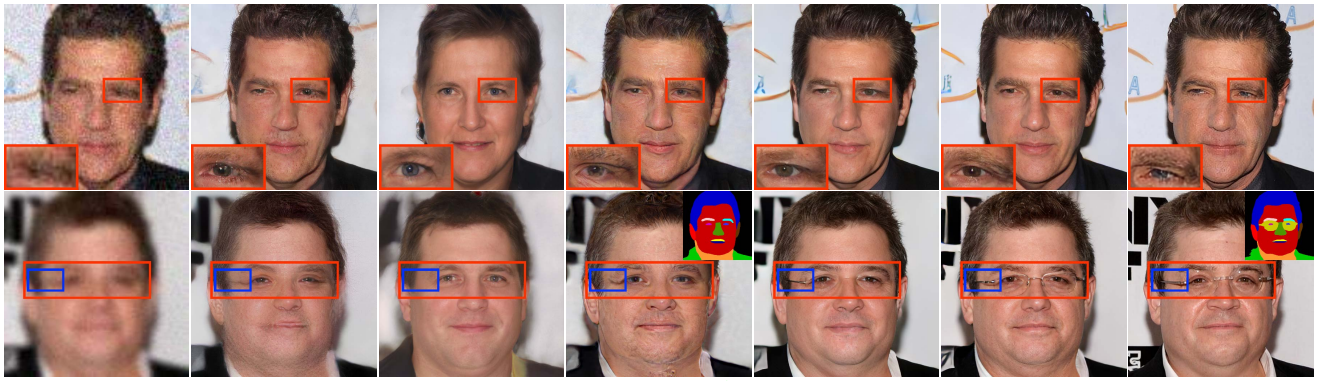


Figure 4. Qualitative comparison on the **CelebA-Test**. The results of our RestoreFormer have a more realistic overview and contain more details in eyes, mouth, and hair. **Zoom in for a better view and more results are shown in supplementary materials**.

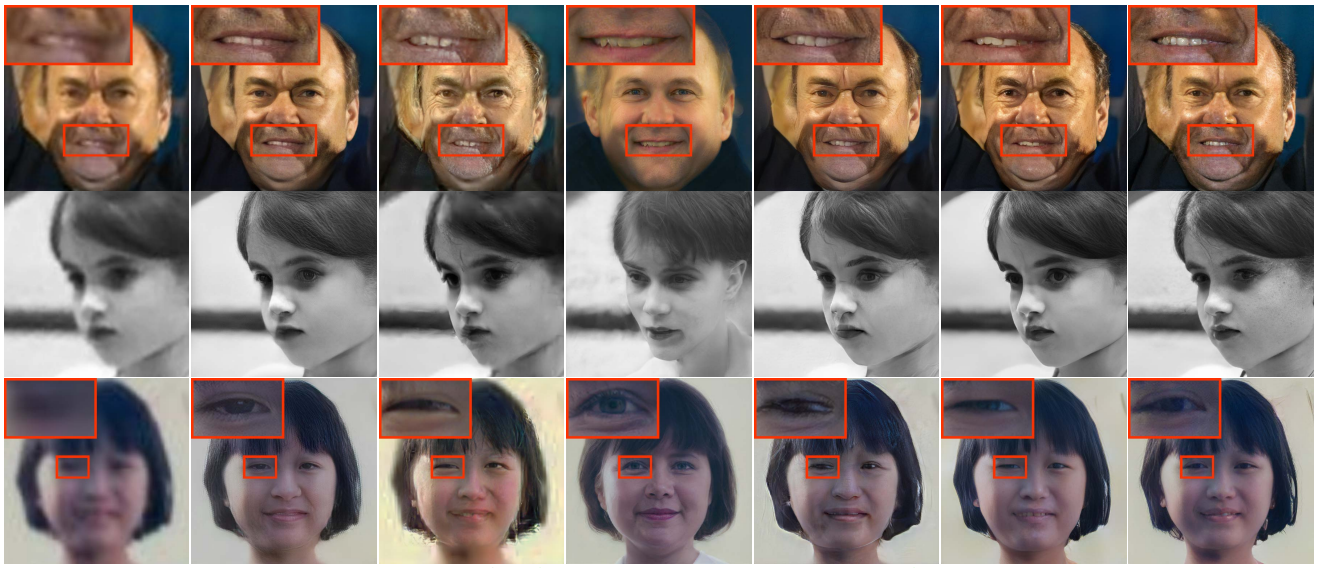


Figure 5. Qualitative comparison on the three **real-world** datasets: **LFW-Test**, **CelebChild-Test**, and **WebPhoto-Test** (from top to down, respectively). **Zoom in for a better view and more results are shown in supplementary materials**.

identity simultaneous relative to existing arts. Compare to the results of Wan *et al*. [35] and PULSE [29], which are based on generative priors without considering the identity information in the degraded faces, the results from RestoreFormer look more similar to the input. Besides, since the MHCA in RestoreFormer can utilize contextual information, the eyes of the third row in Figure 5 look more visually pleasant than [5, 24, 37].

To further evaluate the visual quality, we recruit 100 volunteers for a **user study** on 200 samples randomly selected from LFW-Test and WebPhoto-Test (each dataset provides 100 samples). We conduct pair-wise comparisons between RestoreFormer and three lately state-of-the-art methods: DFDNet [24], PSFRGAN [5], and GFP-GAN [37]. As shown in Table. 3, our RestoreFormer performs better than other methods with a higher percentage.

4.4. Ablation Study

According to the analysis above, RestoreFormer has several merits. First of all, the spatial attention mechanism is used to utilize the abundant contextual information in face images for restoration. In addition, the proposed method can properly utilize the identity information from the degraded face and high-quality facial details from priors. At last, Dictionary used in RestoreFormer is reconstruction-oriented other than the recognition-oriented one used in [24]. The above factors will be discussed in the following subsections and these networks[1] are trained by exactly the same settings as to RestoreFormer.

Spatial attention. In this subsection, variants of RestoreFormer without and with attention mechanism are compared. Both exp1 and exp2 in Table 4 only use degraded

[1]Please see the supplemental materials for the detailed structures for these networks.

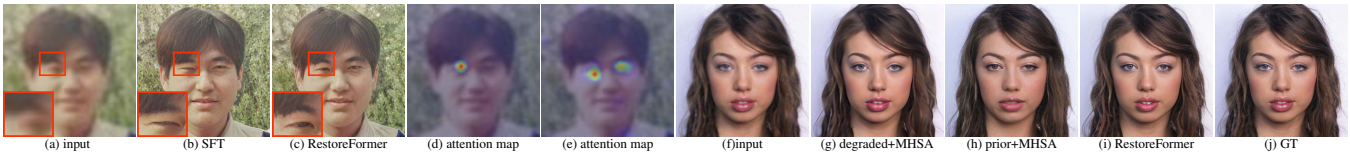


Figure 6. Ablation studies. (c) is the restored face of (a) by RestoreFormer. (b) replaces the MHCA with SFT to validate the effectiveness of the facial contextual information. (d) and (e) are two attention maps for the left eye in RestoreFormer. (f) to (j) are to validate the effectiveness of fusing the information from degraded image and prior. (g) and (h) use self-attention, *i.e.* MHSA, to process either degraded information from the input or prior information from HQ Dictionary. While our RestoreFormer can utilize these two sources of information to restore a face (i) that looks more visually pleasant than (g) and more similar to the ground truth (j) than (h). Please see the text for more details.

images in the network. By using self-attention and exploring contextual information, exp2 with MHSA has lower FID and IDD than exp1 which directly uses the features extracted from the degraded image. This conclusion is also valid when the networks consider information from both degraded image and dictionary prior in exp4 and RestoreFormer in Table 4. In exp4, MHCA is replaced by SFT in RestoreFormer to locally fuse the information. Without considering the global contextual information, the left eye seems strange in Figure 6 (b). As shown in Figure 6 (d) and (e), the multi-head attention maps of the left eye region have more weights for both two eyes in the RestoreFormer with MHCA. This means RestoreFormer with MHCA utilizes the information from both eyes to restore the left one and generates a more visually pleasant result in Figure 6 (c).

Degraded information and Prior. This subsection analyses the effect of degraded information from input images and priors from the HQ Dictionary. Similar to existing ViT methods which use self-attention(MHSA), all the query, key and value are either from features of degraded images (exp2) or priors (exp3) in Table 4. It shows that exp2 has a better average IDD score for keeping the identity of the faces and exp3 has a better average FID score for the realness of the results. By utilizing cross-attention (MHCA) in RestoreFormer to fuse these two sources of information, RestoreFormer is better than exp2 and exp3 for both IDD and FID. As to the visual result, Figure 6 (g) shows that 'degraded+MHSA' (exp2) can restore a face that looks more like the ground truth. However, its result contains fewer details relative to RestoreFormer in Figure 6 (i) which makes the face visually less pleasant. Even though the details in 'prior+MHSA' (exp3) look more natural in Figure 6 (h), the generated face looks like a different person relative to the ground truth, especially for the mouth. By fusing the information from degraded image and prior, RestoreFormer can restore face with more real details as well as maintaining identity shown in Figure 6 (i). According to Figure 2 (b) and Eq. 6, there is a skip connection between the attended feature $\boldsymbol{Z}_{mh}$ and prior $\boldsymbol{Z}_p$ in the RestoreFormer. This is because we experimentally find it performs better than adding $\boldsymbol{Z}_{mh}$ with the feature from degraded input $\boldsymbol{Z}_d$ denoted as exp5 in Table 4.

Reconstruction-oriented v.s. Recognition-oriented. To evaluate the effectiveness of the proposed reconstruction-oriented HQ Dictionary, we replace the encoder $\mathbf{E}_d$ and $\mathbf{E}_h$ with a well-trained VGG [33] which is used in [24] for face restoration and get a recognition-oriented HQ Dictionary in Restoreformer. When training this Restoreformer, the encoder is initialized by VGG and fixed similar to [24]. The experimental results in CelebA-Test show that the average FID and IDD of this Restoreformer variant are 61.43 and 1.1401 which are worse than the proposed one according to Table 4. And this demonstrates the effectiveness of the reconstruction-oriented dictionary.

	sources		methods					metrics	
No. of exp.	degraded	prior	none	MHSA	SFT	MHCA-D	MHCA-P	FID↓	IDD↓
exp1	✓		✓					50.68	0.6401
exp2	✓			✓				47.39	0.6284
exp3		✓		✓				45.83	0.7662
exp4	✓	✓			✓			41.47	0.6702
exp5	✓	✓				✓		42.00	0.5938
Ours	✓	✓					✓	**41.45**	**0.5650**

Table 4. Quantitative results of ablation studies on CelebA-Test. 'degraded' and 'prior' mean fusion information from degraded input and HQ Dictionary, respectively. 'none' and 'MHSA' mean the network uses either 'degraded' or 'prior' information without or with using self-attention mechanism, respectively. 'SFT', 'MHCA-D' and 'MHCA-P' use both 'degraded' and 'prior' information. 'SFT' uses SFT to fuse the information while 'MHCA-D' and 'MHCA-P' use multi-head cross attention. The difference between 'MHCA-D' and 'MHCA-P' is 'MHCA-D' fuses $\boldsymbol{Z}_{mh}$ with $\boldsymbol{Z}_d$ but 'MHCA-P' fuses $\boldsymbol{Z}_{mh}$ with $\boldsymbol{Z}_p$. The proposed RestoreFormer integrated with 'MHCA-P' performs the best relative to other variants.

5. Conclusion

This paper aims for blind face restoration with a RestoreFormer, which explores fully-spacial attentions to model contextual information with a multi-head cross-attention layer to learn spatial interaction between corrupted queries and high-quality key-value pairs. Especially, the high-quality key-value pairs are sampled from a reconstruction-oriented dictionary, whose elements are rich in high-quality facial features specifically aimed for face reconstruction. Extensive comparisons with state-of-the-art methods on several datasets demonstrate the superior capability of the proposed RestoreFormer.

References

[1] Yochai Blau, Roey Mechrez, Radu Timofte, Tomer Michaeli, and Lihi Zelnik-Manor. The 2018 pirm challenge on perceptual image super-resolution. In *ECCVW*, 2018. 6

[2] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020. 2

[3] Qingxing Cao, Liang Lin, Yukai Shi, Xiaodan Liang, and Guanbin Li. Attention-aware face hallucination via deep reinforcement learning. In *CVPR*, 2017. 1

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 2, 3

[5] Chaofeng Chen, Xiaoming Li, Lingbo Yang, Xianhui Lin, Lei Zhang, and Kwan-Yee K Wong. Progressive semantic-aware style transformation for blind face restoration. In *CVPR*, 2021. 1, 2, 3, 6, 7

[6] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *CVPR*, 2021. 2, 3

[7] Yu Chen, Ying Tai, Xiaoming Liu, Chunhua Shen, and Jian Yang. Fsrnet: End-to-end learning face super-resolution with facial priors. In *CVPR*, 2018. 1, 2

[8] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, 2019. 5, 6

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2

[10] Berk Dogan, Shuhang Gu, and Radu Timofte. Exemplar guided face image super-resolution without facial landmarks. In *CVPRW*, 2019. 1, 2

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2, 3

[12] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, 2021. 3

[13] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *CVPR*, pages 2414–2423, 2016. 4

[14] Jinjin Gu, Yujun Shen, and Bolei Zhou. Image processing using multi-code gan prior. In *CVPR*, 2020. 1, 2

[15] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 4

[16] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *NIPS*, 2017. 6

[17] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008. 5

[18] Huaibo Huang, Ran He, Zhenan Sun, and Tieniu Tan. Wavelet-srnet: A wavelet-based cnn for multi-scale face super resolution. In *ICCV*, 2017. 1

[19] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016. 4

[20] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 5

[21] Deokyun Kim, Minseon Kim, Gihyun Kwon, and Dae-Shik Kim. Progressive face super-resolution via attention to facial landmark. *arXiv preprint arXiv:1908.08239*, 2019. 1, 2

[22] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv*, 2014. 6

[23] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, 2017. 4, 6

[24] Xiaoming Li, Chaofeng Chen, Shangchen Zhou, Xianhui Lin, Wangmeng Zuo, and Lei Zhang. Blind face restoration via deep multi-scale component dictionaries. In *ECCV*, 2020. 1, 2, 3, 4, 5, 6, 7, 8

[25] Xiu Li, Guichun Duan, Zhouxia Wang, Jimmy Ren, Yongbing Zhang, Jiawei Zhang, and Kaixiang Song. Recovering extremely degraded faces by joint super-resolution and facial composite. In *ICTAI*, 2019. 2

[26] Xiaoming Li, Wenyu Li, Dongwei Ren, Hongzhi Zhang, Meng Wang, and Wangmeng Zuo. Enhanced blind face restoration with multi-exemplar images and adaptive spatial feature fusion. In *CVPR*, 2020. 1, 2, 5

[27] Xiaoming Li, Ming Liu, Yuting Ye, Wangmeng Zuo, Liang Lin, and Ruigang Yang. Learning warped guidance for blind face restoration. In *ECCV*, 2018. 1, 2, 5

[28] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, 2015. 5

[29] Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. Pulse: Self-supervised photo upsampling via latent space exploration of generative models. In *CVPR*, 2020. 1, 2, 6, 7

[30] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. *arXiv preprint arXiv:1711.00937*, 2017. 2, 3, 4, 5

[31] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. In *ICML*, 2018. 3

[32] Ziyi Shen, Wei-Sheng Lai, Tingfa Xu, Jan Kautz, and Ming-Hsuan Yang. Deep semantic face deblurring. In *CVPR*, 2018. 1, 2

[33] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2, 3, 4, 5, 8

[34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017. 2, 3

[35] Ziyu Wan, Bo Zhang, Dongdong Chen, Pan Zhang, Dong Chen, Jing Liao, and Fang Wen. Bringing old photos back to life. In *CVPR*, 2020. 1, 2, 6, 7

[36] Huiyu Wang, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Max-deeplab: End-to-end panoptic segmentation with mask transformers. In *CVPR*, 2021. 3

[37] Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. Towards real-world blind face restoration with generative facial prior. In *CVPR*, 2021. 1, 2, 3, 4, 5, 6, 7

[38] Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *CVPR*, 2018. 2

[39] Xiangyu Xu, Deqing Sun, Jinshan Pan, Yujin Zhang, Hanspeter Pfister, and Ming-Hsuan Yang. Learning to super-resolve blurry face and text images. In *ICCV*, 2017. 1

[40] Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo. Learning texture transformer network for image super-resolution. In *CVPR*, 2020. 3

[41] Xin Yu, Basura Fernando, Bernard Ghanem, Fatih Porikli, and Richard Hartley. Face super-resolution guided by facial component heatmaps. In *ECCV*, 2018. 1, 2

[42] Xin Yu, Basura Fernando, Richard Hartley, and Fatih Porikli. Super-resolving very low-resolution face images with supplementary attributes. In *CVPR*, 2018. 1, 2

[43] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 6

[44] Long Zhao, Zizhao Zhang, Ting Chen, Dimitris N Metaxas, and Han Zhang. Improved transformer for high-resolution gans. *arXiv preprint arXiv:2106.07631*, 2021. 3

[45] Mingrui Zhu, Changcheng Liang, Nannan Wang, Xiaoyu Wang, Zhifeng Li, and Xinbo Gao. A sketch-transformer network for face photo-sketch synthesis. *IJCAI*, 2021. 3

[46] Shizhan Zhu, Sifei Liu, Chen Change Loy, and Xiaoou Tang. Deep cascaded bi-network for face hallucination. In *ECCV*, 2016. 1, 2

[47] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 2, 3