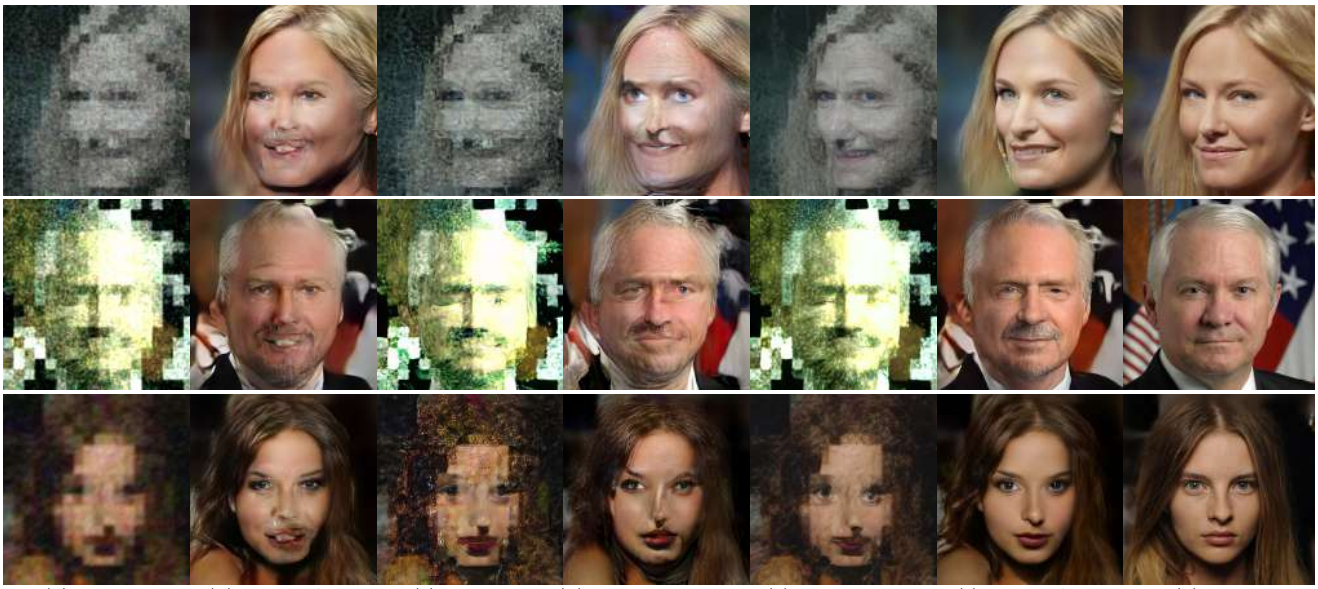
This CVPR workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

TFRGAN: Leveraging Text Information for Blind Face Restoration with Extreme Degradation

Chengxing Xie * Qian Ning * Weisheng Dong † Guangming Shi
Xidian University
xiechengxing34@gmail.com ningqian@stu.xidian.edu.cn
wsdong@mail.xidian.edu.cn gmshi@xidian.edu.cn



(a) Input (b) GFPGAN* (c) VQFR (d) RestoreFormer* (e) CodeFormer (f) TFRGAN (g) GT

Figure 1. Visual comparisons with state-of-the-art face restoration methods. The results of proposed TFRGAN contain more texture details and complete face structures, which are the most natural and realistic. $*$ denotes that the model is fine-tuned in our training set.

Abstract

Blind face restoration aims to recover high-quality face images from unknown degraded low-quality images. Previous works that are based on geometric or generative priors have achieved impressive performance, but the task remains challenging, particularly when it comes to restoring severely degraded faces. To address this issue, we propose a novel approach TFRGAN, that leverages textual information to improve the restoration of extremely degraded face images. Specifically, we propose to generate a better and more accurate latent code for StyleGAN2 prior via fusing the text and image information in the latent code space. Besides, extracted textual features are used to modulate the decoding features to obtain more realistic and natural facial images with more reasonable details. Experimental results demonstrate the superiority of the proposed method for restoring severely degraded face images.

1. Introduction

The goal of blind face restoration is to restore high-quality face images from corresponding low-quality ones. Many different degradation factors such as noise [49], blur [17,31], and downsampling [5,6] cause low-quality face images. Many different face-specific priors [3, 38, 43, 47, 52] have been used in previous works. For example, the geometric priors including facial landmarks [2, 3], heat maps [47], and parsing maps [1,43] provide a general outline and shape of the face for blind face restoration task. However,

*Equal Contribution
†Corresponding Author

the accuracy of these geometric priors is greatly affected by the degradation of low-quality face images, resulting in unsatisfactory guidance for blind face restoration.

More recently, many works have been proposed for blind face restoration via leveraging the generative priors [9, 38, 41, 52] to enhance the quality of generated results. These approaches typically involve some pre-trained face generative models, such as StyleGAN2 [15] and VQGAN [7], to incorporate rich texture details in the recovered faces. These methods can be divided into two categories: one that utilizes continuous feature space decoding for image generation [38], and another that generates faces from fixed, discrete codebook representations [9, 41, 52]. However, the discrete codebooks would limit the model's representational capability, resulting in less variety in the generated images compared to those decoded from the continuous space. Several works [9, 19, 38, 41] are exploring the projection of information from degraded images to potential vector representation in the latent space.

In cases of facial images with severe degradation, significant facial feature information from the original image may be lost during the degradation process, resulting in the extracted latent code that may not be sufficient or accurate for reconstruction of the real image when using generative priors such as StyleGAN2 [15] or Codebook [7]. Obtaining a more accurate latent code corresponding to a high-quality image will lead to a better restoration. Therefore, we propose a novel approach called TFRGAN that leverages text information to facilitate the restoration of extremely degraded facial images. However, the fusion of image and text information that belong to different two modalities is quite challenging due to the differences between these two modalities. To address this challenge, we process two different approaches to integrate textual information into the restoration process, both in the latent code space and decoding process. Firstly, we map the extracted textual features into text latent code and then fuse it with image latent code via the proposed text-image fusion block (TIFB). This process obtain an improved latent code, comparing to the original image latent code that is extracted from the extremely degraded facial image. Additionally, we propose a text-guided decoder that modulates the image decoding features with the extracted text feature maps to produce more realistic results.

The main contributions of this paper are summarized as follows:

- To the best of our knowledge, we are the first to use textual information to facilitating the restoration of extremely degraded facial images and achieving promising results.
- We propose two modules to fully leverage textual information in the restoration of facial images. Specifically, we propose the text-image fusion block to fuse text and image information in the latent space and the text-guided block to fuse them in the feature decoding space.
- Extensive experiments demonstrate the superiority of our proposed approach for the task of blind face restoration, especially in extremely degraded ones.

2. Related Works

2.1. Blind Face Restoration

The current state-of-the-art in face restoration primarily utilizes various forms of priors, such as geometric prior, generative prior, facial attributes, and identity information. These geometric prior include facial landmarks [2, 3], heat maps [47], and parsing maps [1, 43], which are often difficult to extract and utilize effectively in highly degraded facial images.

In recent years, advances in blind face restoration [7, 14, 15, 28] have been made by leveraging the high-quality face generative priors that contains ample facial texture information. These methods [9, 21, 23, 38, 41, 52] map the low-quality face images into the latent space of the generative model such as StyleGAN2 [15] to realize high-quality facial images reconstruction. GFPGAN [38] aims to exploit the high-quality image generation capability of StyleGAN2 [15] and use spatial information to modulate the features of StyleGAN2 at multiple scales, recovering the facial images with more fidelity. The RestoreFormer [41] uses a pre-trained HQ image feature dictionary [21, 35, 45] as a correspondingly high-quality image generator to train a corresponding encoder to map the network from low-quality image features to high-quality image dictionary locations. VQFR [9] introduces high-quality codebooks in it to obtain better restoration results. CodeFormer [52] uses a transformer network to predict the index of high-quality codebooks from low-quality image features to improve face image quality.

However, these methods rely solely on the high-quality images priors, which may not be sufficient for extremely degraded images. To address this issue, we propose to use textual information to improve the restoration results with extreme degradation in this paper. A more comprehensive understanding of the severely degraded images can be obtained by incorporating additional textual information that is extracted from a short text description, leading to better restoration results.

2.2. Multi Modalities Learning

Recently, there has been a growing body of researches in the field of multi-modality in image-text matching and text-image generation [25–27, 37, 46]. One notable approach,

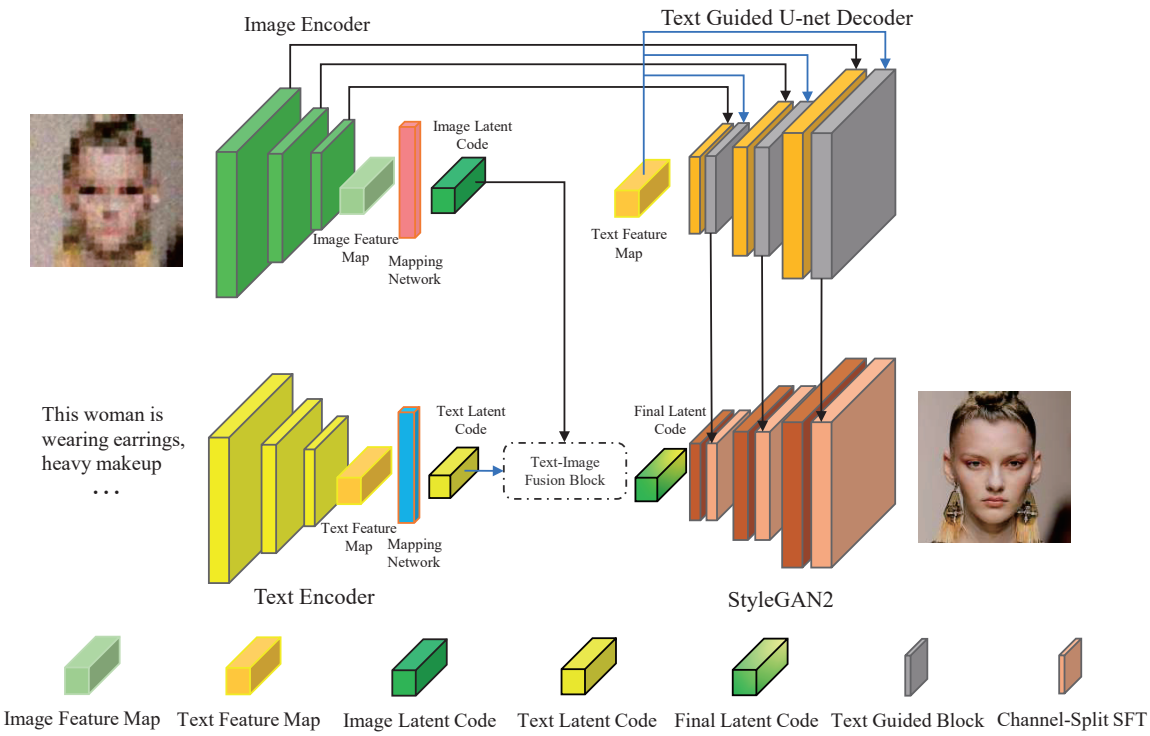


Figure 2. **Overview of TFRGAN:** The TFRGAN model's architecture includes an image encoder, a text encoder, a text-guided U-net decoder, and a StyleGAN2-based decoder. In the latent code space, we fuse the Text Latent Code and Image Latent Code using the text-image fusion block. This fused latent code is then passed to the StyleGAN2-based decoder to recover the final face image. To further enhance the quality of the generated image, a text-guided decoder is proposed to leverage the extracted text features to provide text guidance information to the StyleGAN2-based decoder.

CLIP [25], utilizes different encoders for texts and images respectively to create feature vectors that have high similarity for paired text and images, and low similarity between images and unpaired text by encoding both the text and images into the same feature space. Derivative works, such as DALLE [27] and DALLE 2 [26], have also been developed via utilizing the powerful capabilities of CLIP.

Another direction in this field involves utilizing a single encoder for both text and image modalities, which allows the same model to process data from both modalities. Examples of this approach include VL-BERT [33], which employs a unified transformer [36] encoder to process data from both visual and textual modalities. Additionally, there are models that integrate the features of both modalities, such as ALBEF [20]. ALBEF first encodes the visual and textual information separately using separate transformer encoders and then performs alignment operations on the outputs. The outputs from the first stage models are then concatenated and fed into a unified transformer encoder for feature fusion.

Although the application of multi-modal techniques into the task of image restoration is an area that has yet to be explored, incorporating text information as prior knowledge into the restoration networks through multi-modal fusion has the potential to improve the quality of restored images.

3. Method

In this paper, we propose a novel method called TFRGAN that leverages the text information to boost the restoration of extremely degraded facial images. As shown in Fig. 2, the TFRGAN mainly consists of an image encoder, a text encoder, a text-image fusion block, a text-guided U-net decoder, and a StyleGAN2-based decoder. Specifically, we utilize a U-net [29] like network as image encoder to extract facial feature maps from low-quality facial images, a BERT [4] like network as text encoder to embedding text descriptions which are paired with the low-quality images. Then, the extracted image and text feature maps are mapped into image and text latent code via two different mapping networks respectively. The proposed text-image fusion block fuse extracted image and text latent code to generate an improved latent code, which will be il-

lustrated in Sec. 3.2. In the decoding part of the TFRGAN model, we employ two decoders: a pre-trained StyleGAN2 [15] decoder and a text-guided U-net decoder as shown in Fig. 2. The fused latent code is passed to the StyleGAN2-based decoder to recover the final face image. To further enhance the quality of the generated image, a text-guided decoder is proposed to provide text guidance information to the StyleGAN2-based decoder by leveraging the extracted text features.

3.1. Text Encoder

To leverage the text information to facilitate the restoration of facial images that are extremely degraded, the text information will first be encoded and then mapped into the same latent space as the face images. In this paper, we propose to utilize DistilBERT [30] as the backbone of Text Encoder. The DistilBERT [30] is a distilled version of the BERT [4] model, which has been trained to maintain the majority of its original performance while being significantly smaller and faster than BERT [4]. This allows for the efficient encoding of text features, making it suitable for our image restoration task.

The DistilBERT [30] model encodes each word in an image caption into a text vector $V_{word} \in \mathbb{R}^{768}$. With a maximum caption length of 30 words, the final caption features are represented as $V_{caption} \in \mathbb{R}^{768\times30}$. To extract the most relevant information for blind face restoration, we introduce a mapping network to condense the caption features into a single vector, which can be denoted as $z'_{text} \in \mathbb{R}^{512\times1}$. To keep the same size of image latent code z_{img}, the final text latent code $z_{text} \in \mathbb{R}^{512\times16}$ is obtained by concatenating the condensed text features z'_{text} from multiple captions or random copy from z'_{text} when the captions are less than 16. It should be noted that only the parameters of the mapping network are trained during the training process, while the parameters of the DistilBERT model remain frozen.

3.2. Text-Image Fusion Block

The quality of the recovered facial images is heavily influenced by the representation of the latent code. After mapping the extracted text feature maps into the text latent code, we propose a novel module called the text-image fusion block (TIFB) to fuse the text latent code with image latent code to generate an improved latent code. The overview of the text-image fusion block is shown in Fig. 3a. To effectively leverage the extracted text and image latent code, we first adopt separate embedding layers before seeding them into the transformer layers as shown in Fig. 3a. Let z_{img} and z_{text} denote image latent code and text latent code respectively. Then, the latent code that is fed into transformer layers, which can be formulated as

$$z_{concat} = \text{Concat}(\text{Em}_i(z_{img}), \text{Em}_t(z_{text})), \quad (1)$$

where Em_i and Em_t denote the image and text embedding layers, respectively. Then, the latent code that is fused via multiple transformer layers can be represented as z_{trans}. The transformer block can effectively fuse two domain latent codes and extract the better latent code z_{trans}. The output z_{trans} of this block is then seamlessly integrated into the StyleGAN2-based decoder to restore the final face images with significant improvement via improved latent code.

3.3. Text Guided U-net Decoder

The text-guided U-net decoder is a modified U-net architecture, which takes the text feature maps F_{text} and extracted multi-scale image feature maps F^l_{img} as input, where l denotes scale index. The text information is incorporated into the decoder via modulating the corresponding feature maps of the U-net decoder. This allows the network to focus on restoring facial details that correspond to the encoded text descriptions. As shown in Fig. 2, the basic unit of a text-guided U-net decoder is U-net convolution layers and a text-guided block.

The architecture of text-guided block (TGB) is shown in Fig. 3b. The TGB takes the smallest scale text feature maps F_{text} and the multi-scale image feature maps $F^k_{img} \in \mathbb{R}^{C^k\times H^k\times W^k}$ as input, where k denotes scale index. In order to better preserve facial characters, we propose to use spatial feature transform (SFT) [39] to modulate the text information into the image decoding features for better restoration of facial images within the decoding process.

In each text-guided block, we generate a pair of affine transformation parameters T_{sc}, T_{sh} for U-net convolution layers from input text features via mapping network, which consist of several convolution layers. Specifically, the modulation process of U-net convolution layers can be formulated as

$$\begin{aligned} T_{sc}, T_{sh} &= \text{MappingNet}(F_{text}), \\ F_{U-SFT} &= T_{sc} \times F_{img} + T_{sh}, \end{aligned} \quad (2)$$

where T_{sc}, T_{sh} denote scale and shift parameters respectively. By incorporating text information in the image restoration process, the proposed method aims to produce higher-quality results that are more consistent with the given textual input. It is worth to mention that we generate images $I^k_{out\_Unet}$ at each resolution scale of the text-guided U-net decoder and constrain them to closely resemble the pyramid of the ground-truth image.

3.4. StyleGAN2 based Decoder

StyleGAN2 [15] is an impressive generative model that has proven to be highly effective in producing high-quality images of human faces with realistic details and textures. StyleGAN2 first generates a latent code from a given random noise via a mapping network. Then the latent code is

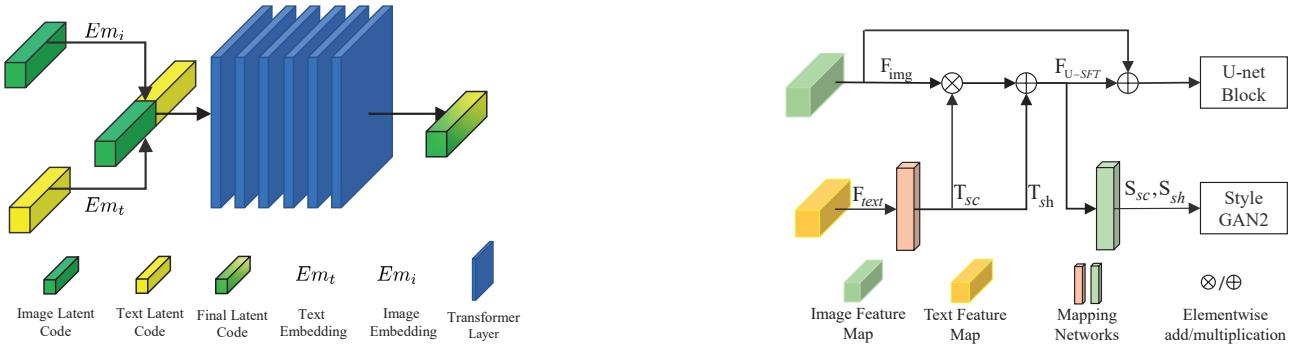


Figure 3. (a) The overview of Text-Image Fusion Block. (b) The overview of Text-Guided Block.

used to generate the final facial image via the trained StyleGAN2 network. StyleGAN2 also incorporates an adaptive instance normalization (AdaIN) operation, which provides fine-grained control over the generated images via changing the latent code.

In the proposed method, we utilize StyleGAN2 as a generative prior to produce high-quality restored images. The output of the text-image fusion block z_{trans} which fuses the text and image information, is used as the latent code to generate the final images. By leveraging the pre-trained generator prior, we are able to generate realistic and high-quality images with improved latent code z_{trans}. In order to better modulate the features of StyleGAN2 F_{GAN} with text information, we use the CS-SFT module proposed in GFP-GAN [38] model. Unlike GFPGAN which only use spatial features to obtain affine transformation parameters, we use the features F_{U-SFT} that incorporate the text and image before to obtain the modulation parameters. This process can be formulated as:

$$\begin{aligned} S_{sc}, S_{sh} &= \text{MappingNet}(F_{U-SFT}), \\ F_{GAN-SFT} &= S_{sc} \times F_{GAN} + S_{sh}, \end{aligned} \tag{3}$$

where S_{sc}, S_{sh} denote scale and shift parameters respectively, and the mapping networks consist of several convolution layers. This modulation enables our model to better capture the underlying features of the input image with the hint of text description and effectively restore the facial details. With these techniques, we are able to produce highly realistic and visually pleasing restored facial images.

3.5. Training Loss

To restore realistic facial images with fidelity, multiple loss functions are used to train the proposed network. The overall loss functions consist of reconstruction loss, adversarial loss [8, 15], perceptual loss [12, 50], identity preserving loss [11], and Pyramid Restoration Loss. Let I_{out} denote the output of the StyleGAN network and I_{gt} denotes the target images.

We adopt $\mathcal{L}_1$ as our reconstruction loss to measure the difference between the restored facial images and the target images.

$$\mathcal{L}_{rec} = \lambda_{rec} \| I_{out} - I_{gt} \|_1, \tag{4}$$

where λ_{rec} represents the weight of reconstruction loss.

In addition to the reconstruction loss, we also incorporate a perceptual loss, which is calculated using a pre-trained VGG [32] network, which can be formulated as

$$\mathcal{L}_{per} = \lambda_{per} \sum_{i}^{N} || g_i(I_{out}) - g_i(I_{gt}) ||_1, \tag{5}$$

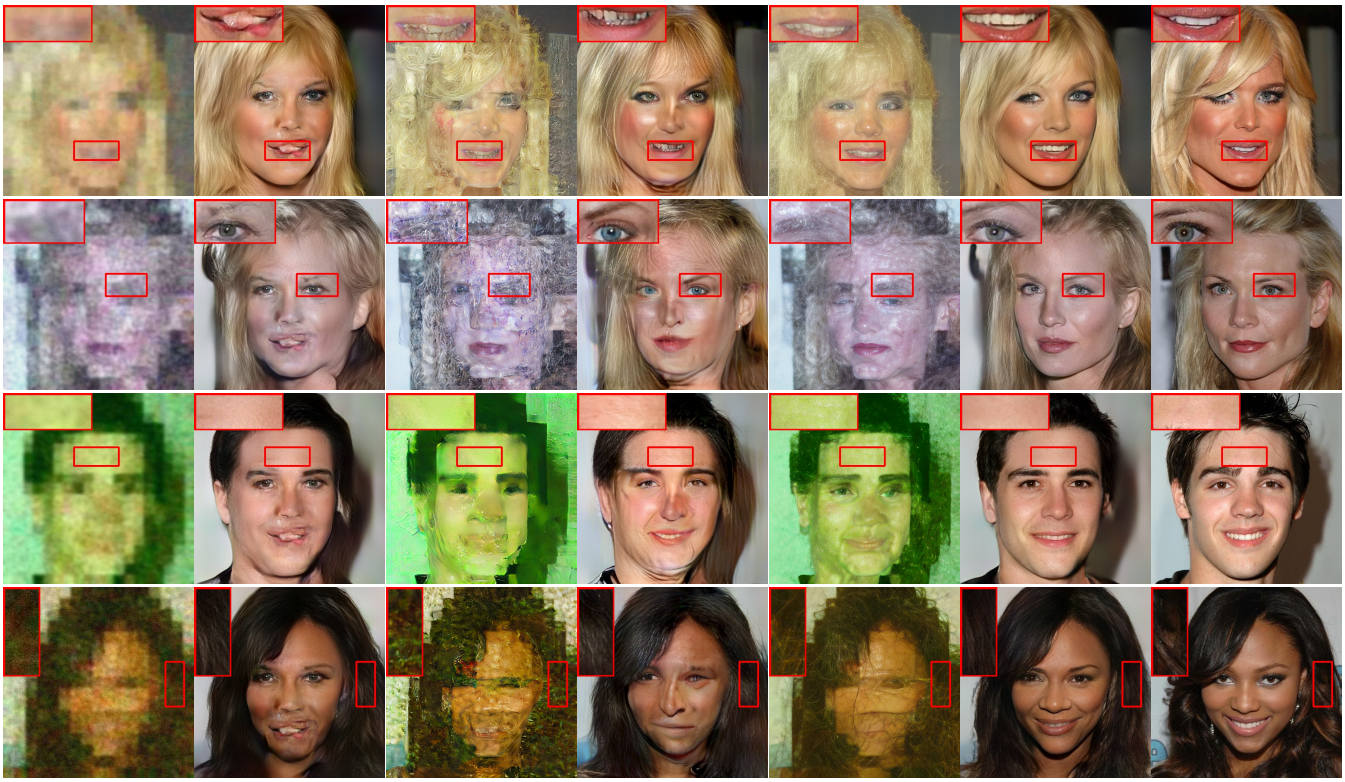
where g_i denotes the i-th layer feature maps of a pre-trained VGG network, λ_{per} represents the weight of perceptual loss. In our implementation, we compute the perceptual loss using the first five layers of VGG feature maps.

To further improve the visual realism of the restored image, we incorporate an adversarial loss, similar to the loss that are used in StyleGAN2. This loss function encourages the model to produce perceptually realistic images that are indistinguishable from high-quality images.

$$\mathcal{L}_{adv} = -\lambda_{adv} \mathbb{E}_{I_{out}} \, \text{softplus}(D(I_{out})), \tag{6}$$

where λ_{adv} represents the weight of adversarial loss, and D represents discriminator.

The Pyramid Restoration Loss (PRL) is also used. In this loss function, the ground-truth (GT) image is first down-sampled into several scales I^k_{gt}, and then the images $I^k_{out\_Unet}$ are generated from each resolution scale of the text-guided U-net decoder. A reconstruction loss is computed at each scale to measure the similarity between the generated images and the corresponding down-sampled GT image. By evaluating the difference at different levels of resolution, the PRL aims to provide a more comprehensive and robust evaluation of the generated image quality. Math-



(a) Input (b) GFPGAN* (c) VQFR (d) RestoreFormer* (e) CodeFormer (f) **TFRGAN** (g) GT

Figure 4. Qualitative comparison on the CelebAMask-Test for blind face restoration. Our TFRGAN recovers better results with more details in hair, skin, mouth, and eyes. The symbol $*$ denotes that the model is fine-tuned in our training set.

ematically, the PRL can be formulated as:

$$\mathcal{L}_{PRL} = \lambda_{PRL} \sum_{k=1}^{K} \lambda_k ||(I_{gt}^k, I_{out\_Unet}^k)||_1, \tag{7}$$

where λ_k is a weighting factor that controls the contribution of each scale to the final loss, and K is the total number of scales.

In the end, the total loss can be formulated as :

$$\mathcal{L} = \mathcal{L}_{rec} + \mathcal{L}_{per} + \mathcal{L}_{adv} + \mathcal{L}_{PRL} \tag{8}$$

4. Experiments

4.1. Datasets, Settings, and Metrics

Training Datasets: In our training set, we utilize 24000 images from the CelebAMask-HQ dataset [18], where text captions of both CelebAText-HQ [34] and Multi-Modal CelebA-HQ [42] are used. During the training process, the input image size is 512×512. We use a multi-step degradation function to attain extreme degradation, which can be formulated as:

$$I_{deg} = \text{Deg}(\text{Deg}(I_{gt})), \tag{9}$$

where I_{gt} means Ground-Truth Image, I_{deg} means degraded Image, Deg is formulated in Eq. (10).

The Deg function comprises several operations to degrade an image. Firstly, a Gaussian blur kernel is applied to create a blurred image, which is then down-sampled r times. Next, Gaussian white noise n_δ and JPEG compression with a quality factor of j are added to the image, and a random color dithering process is applied. Finally, the BSRGAN [48] degradation function [22, 44] is applied for an additional two image degradation operations. The full formula is expressed as follows:

$$\text{Deg}(I_{gt}) = M(B(B(\left\{\left[(I_{gt} \otimes k_G)_{\downarrow_r} + n_\delta\right]_{\text{JPEG}_j}\right\}_{\uparrow_r} + C))) \tag{10}$$

where k_G represents the Gaussian blur kernel, r represents the downsampling scale, n represents the added Gaussian noise, JPEG_j represents the JPEG compression quality factor, C represents the color jittering factor, and B denotes BSRGAN [48] degradation. We randomly sample G, r, δ, j from $\{0.1,10\}$, $\{4,5\}$, $\{20,40\}$, $\{30,50\}$. We implement the M by first downsampling the image to a resolution of $\{15,20\}$ and then upsampling it back to 512×512, using nearest neighbor sampling for both the downsampling and upsampling operations.

Testing Datasets: In order to test the effectiveness of the proposed method that uses textual prior, a dataset consisting of degraded images and corresponding captions is used. The test dataset is CelebAMask-HQ Test, which contains 6000 high-quality face images selected from the CelebA [13] dataset. To generate the degraded images, the same image degradation function used in the training dataset is applied to the test dataset. The corresponding textual tags are obtained by combining the text descriptions from CelebAText-HQ [34] and Multi-Modal CelebA-HQ [42] for each image. It is important to note that there is no overlap between the test dataset and the training dataset to prevent any data leakage.

Settings: In the proposed method, a U-net like framework is utilized for image restoration, consisting of a seven-layer down-sampling and seven-layer up-sampling. The up-sampling layer includes two additional convolution layers that serve as guiding biases for text information, while the prior spatial information is extracted from the up-sample feature map by two convolution layers. The employed generative model is StyleGAN2 [15], which takes advantage of the rich face information of StyleGAN2 to generate high-quality images. The parameters of the generative network and DistilBERT are fixed while other parameters of the proposed TFRGAN are trained. The batch size is set to 8, using the Adam [16] optimizer and the learning rate is $1e^{-4}$, which is halved every $10k$ iterations. In our experiments, the loss weight is $\lambda_{rec} = 0.1, \lambda_{per} = 1, \lambda_{adv} = 0.1, \lambda_{PRL} = 1$ The experiments are conducted using four NVIDIA RTX 3090 Ti GPUs.

Metrics: To evaluate the performance of the proposed method, we will use non-reference and reference metrics. The non-reference metrics include the Frechet Inception Distance (FID) [10] and the Natural Image Quality Evaluator (NIQE) [24]. The reference metrics include the Peak Signal-to-Noise Ratio (PSNR), the structural similarity index (SSIM) [40] and Learned Perceptual Image Patch Similarity (LPIPS) [51].

4.2. Comparisons with State-of-the Art Methods

We have compared the proposed TFRGAN with several current state-of-the-art blind face restoration methods: VQFR [9], CodeFormer [52], RestoreFormer [41], and GFPGAN [38]. Officially released pre-trained models of these methods are used to restore final results. In order to obtain a more fair comparison, the RestoreFormer [41] and GFPGAN [38] are fine-tuned on our training dataset. The results are shown in Tab. 1, Fig. 1 and Fig. 4 with symbol $*$. Due to the training code of CodeFormer [52] has not been released, we just used officially released pre-trained model for inference. It should be noted that although the proposed method utilizes text information, the size of the training set used is smaller than these of the state-of-the-art methods, may causing an unfair comparison. Same degradation in Eq. (10) are adopted to obtain the degraded testing image, substantially reducing the image quality without changing the resolution. The text captions for this data are from CelebAText-HQ and Multi-Modal CelebA-HQ, respectively. All models are tested in our testing dataset and relevant metrics are calculated.

Methods	PSNR↑	SSIM↑	FID↓	LPIPS↓	NIQE↓
Input	12.55	0.3829	274.8	0.7679	7.927
GFPGAN	13.58	0.4101	122.1	0.6228	**3.016**
GFPGAN*	14.90	<u>0.4649</u>	86.6	<u>0.5262</u>	4.322
CodeFormer	13.63	0.3717	102.3	0.6881	4.507
RestoreForm.	13.57	0.3417	152.7	0.6885	4.004
RestoreForm.*	<u>15.12</u>	0.4396	**61.2**	0.5277	3.509
VQFR	12.09	0.3021	148.1	0.7023	<u>3.268</u>
TFRGAN	**15.36**	**0.4856**	<u>69.8</u>	**0.2943**	4.351
GT	∞	1	49.2	0	3.777

Table 1. The comparison between current state-of-the-art methods and proposed approach. The best performance is shown in bold and the second-best performance is shown by underline. The symbol $*$ denotes that the model is fine-tuned in our training set.

We present the quantitative comparison results in Tab. 1, where our proposed TFRGAN demonstrates consistently competitive performance in comparison to state-of-the-art methods. Notably, our model achieves the highest scores in both the PSNR and SSIM metrics, indicating that our restored results are most similar to the original image in terms of pixel-level accuracy and structural similarity. These high scores suggest that our model outperforms other methods in terms of pixel-level quality. Furthermore, our model achieves the best score in the LPIPS metric, which measures consistency with human perception. The FID metric also indicates that our model's reconstruction results are highly similar to the real image in terms of whole dataset. As for NIQE metric, our model performs slightly inferior to other competing methods. The visual comparisons are shown in Fig. 1 and Fig. 4. The visual results demonstrate that our proposed method outperforms other methods, particularly in extremely degraded facial images.

4.3. Ablation Study

We have conducted ablation studies to verify the effectiveness of the proposed text-image fusion block (TIFB) and Text-Guided Block (TGB). The quantitative results in terms of FID, NIQE, PSNR, SSIM, and LPIPS are shown in Tab. 2. In Tab. 2, the model without TIFB and TGB is denoted as Baseline, the model with TIFB and without TGB is denoted as TIFB, and the model without TIFB and with TGB is denoted as TGB. The visual results are shown in Fig. 5.

Text-Image Fusion Block: The comparison between the model with and without the proposed text-image fusion

(a) Text (b) Input (c) Baseline (d) With TIFB (e) With TGB (f) **TFRGAN** (g) GT

Figure 5. Visual results of ablation experiments on the effectiveness of proposed TIFB module and the TGB module. The model without TIFB and TGB is denoted as Baseline, the model with TIFB and without TGB is denoted as TIFB, the model without TIFB and with TGB is denoted as TGB, and the model with TIFB and TGB is denoted as TFRGAN.

Methods	PSNR↑	SSIM↑	FID↓	LPIPS↓	NIQE↓
Baseline	15.03	0.4846	91.93	0.3020	4.471
With TIFB	15.29	0.4836	73.03	0.2962	4.381
With TGB	15.26	0.4826	72.21	0.2972	4.377
TFRGAN	**15.36**	**0.4856**	**69.79**	**0.2943**	**4.3511**

Table 2. Ablation results on CelebAMask-HQ. The best performance is shown in bold.

block (TIFB) is presented in Tab. 2, which are denotes as *With TIFB* and *Baseline* respectively. The results clearly indicate that the proposed TIFB module leads to a significant improvement in performance, as demonstrated by widely used image quality metrics such as PSNR and FID. Additionally, the visual comparison between the model with and without the proposed text-image fusion block (TIFB) is shown in Fig. 5 (c) and (d). Specifically, the restored images by the model with TIFB exhibit lower distortion and higher quality than those generated by the baseline model. For instance, the bright color of the lips with traces of lipstick is more pronounced in the images produced with TIFB.

Text-Guided Block: We conducted experiments to investigate the effect of proposed TGB module in the text guided U-net decoder. The results are presented in Tab. 2, which are denoted as *Baseline* (without TGB) and *With TGB* (with TGB) respectively. The results in Tab. 2 indicate that integrating a TGB module leads to a significant decrease in the FID metric. We obtained similar findings when comparing *TIFB* (with TGB) and *TFRGAN* (without TGB), as reported in Tab. 2. These results suggest that incorporating additional textual information into the auxiliary network (U-net decoder) guides the generator to generate high-quality images. Moreover, we the visual comparison between the model with and without the proposed text-guided block (TGB) is shown in Fig. 5 (c) and (e). These visual results further support the conclusion that the addition of textual information to the generator improves its performance. The proposed TFRGAN that integrates both TIFB and TGB modules achieves the best performance quantitatively and qualitatively as shown in Fig. 5 and Tab. 2.

4.4. Limitations and Future Work

Although our proposed model has shown promising results in restoring extremely degraded facial images with some text descriptions, we acknowledge that there are some limitations of datasets used in this study. First, the size of both our training and testing sets is relatively small, which could potentially limit the performance of our model. To address this limitation, we plan to expand our dataset by including more diverse and representative facial images, along with their corresponding text descriptions. Moreover, the quality of the text annotations in our current dataset is not optimal since it was built for other tasks, which could negatively affect the performance of our model. To improve the accuracy of our text information, we aim to incorporate higher-quality text annotations in future experiments. Besides, we plan to further enhance the performance of our model by utilizing other more advanced generative networks as prior knowledge for restoring degraded facial images using text information.

5. Conclusion

In this paper, we present the TFRGAN model, a novel approach for blind face restoration that incorporates textual information to enhance the restoration of extremely degraded face images. To achieve this, we introduce two distinct modules that fully exploit textual information in facial image restoration. The first module involves mapping the extracted textual features into text latent code, which is then fused with the image latent code via the proposed text-image fusion block. Additionally, we propose a text-guided decoder that modulates the image feature with the extracted text feature map, leading to more realistic outcomes. Our extensive experiments demonstrate that TFRGAN outperforms existing methods of blind face restoration, producing highly detailed results for severely degraded face images.

Acknowledgments.

This work was supported by the Natural Science Foundation of China under Grant 61991451 and Grant 61836008.

References

[1] Qingxing Cao, Liang Lin, Yukai Shi, Xiaodan Liang, and Guanbin Li. Attention-aware face hallucination via deep reinforcement learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 690–698, 2017. 1, 2

[2] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11896–11905, 2021. 1, 2

[3] Yu Chen, Ying Tai, Xiaoming Liu, Chunhua Shen, and Jian Yang. Fsrnet: End-to-end learning face super-resolution with facial priors. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2492–2501, 2018. 1, 2

[4] 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. 3, 4

[5] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV 13*, pages 184–199. Springer, 2014. 1

[6] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV 13*, pages 184–199. Springer, 2014. 1

[7] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 2

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 5

[9] Yuchao Gu, Xintao Wang, Liangbin Xie, Chao Dong, Gen Li, Ying Shan, and Ming-Ming Cheng. Vqfr: Blind face restoration with vector-quantized dictionary and parallel decoder. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVIII*, pages 126–143. Springer, 2022. 2, 7

[10] 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. *Advances in neural information processing systems*, 30, 2017. 7

[11] Rui Huang, Shu Zhang, Tianyu Li, and Ran He. Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis. In *Proceedings of the IEEE international conference on computer vision*, pages 2439–2448, 2017. 5

[12] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 694–711. Springer, 2016. 5

[13] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 7

[14] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 2

[15] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 2, 4, 5, 7

[16] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 7

[17] Orest Kupyn, Volodymyr Budzan, Mykola Mykhailych, Dmytro Mishkin, and Jiří Matas. Deblurgan: Blind motion deblurring using conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8183–8192, 2018. 1

[18] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 6

[19] Aijin Li, Gen Li, Lei Sun, and Xintao Wang. Faceformer: Scale-aware blind face restoration with transformers. *arXiv preprint arXiv:2207.09790*, 2022. 2

[20] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021. 3

[21] Xiaoming Li, Chaofeng Chen, Shangchen Zhou, Xianhui Lin, Wangmeng Zuo, and Lei Zhang. Blind face restoration via deep multi-scale component dictionaries. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pages 399–415. Springer, 2020. 2

[22] Xiaoming Li, Chaofeng Chen, Shangchen Zhou, Xianhui Lin, Wangmeng Zuo, and Lei Zhang. Blind face restoration via deep multi-scale component dictionaries. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pages 399–415. Springer, 2020. 6

[23] Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. Pulse: Self-supervised photo upsampling via latent space exploration of generative models. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 2437–2445, 2020. 2

[24] Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2013. 7

[25] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 3

[26] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 2, 3

[27] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 2, 3

[28] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32, 2019. 2

[29] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 3

[30] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108, 2019. 4

[31] Ziyi Shen, Wei-Sheng Lai, Tingfa Xu, Jan Kautz, and Ming-Hsuan Yang. Deep semantic face deblurring. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8260–8269, 2018. 1

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

[33] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019. 3

[34] Jianxin Sun, Qi Li, Weining Wang, Jian Zhao, and Zhenan Sun. Multi-caption text-to-face synthesis: Dataset and algorithm. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 2290–2298, 2021. 6, 7

[35] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 2

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3

[37] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*, 2022. 2

[38] Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. Towards real-world blind face restoration with generative facial prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9168–9178, 2021. 1, 2, 5, 7

[39] Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 606–615, 2018. 4

[40] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 7

[41] Zhouxia Wang, Jiawei Zhang, Runjian Chen, Wenping Wang, and Ping Luo. Restoreformer: High-quality blind face restoration from undegraded key-value pairs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17512–17521, 2022. 2, 7

[42] Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. Tedigan: Text-guided diverse face image generation and manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2256–2265, 2021. 6, 7

[43] Lingbo Yang, Shanshe Wang, Siwei Ma, Wen Gao, Chang Liu, Pan Wang, and Peiran Ren. Hifacegan: Face renovation via collaborative suppression and replenishment. In *Proceedings of the 28th ACM international conference on multimedia*, pages 1551–1560, 2020. 1, 2

[44] Tao Yang, Peiran Ren, Xuansong Xie, and Lei Zhang. Gan prior embedded network for blind face restoration in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 672–681, 2021. 6

[45] Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, and Yonghui Wu. Vector-quantized image modeling with improved vqgan. *arXiv preprint arXiv:2110.04627*, 2021. 2

[46] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 2

[47] Xin Yu, Basura Fernando, Bernard Ghanem, Fatih Porikli, and Richard Hartley. Face super-resolution guided by facial component heatmaps. In *Proceedings of the European conference on computer vision (ECCV)*, pages 217–233, 2018. 1, 2

[48] Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4791–4800, 2021. 6

[49] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE transactions on image processing*, 26(7):3142–3155, 2017. 1

[50] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 5

[51] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 7

[52] Shangchen Zhou, Kelvin CK Chan, Chongyi Li, and Chen Change Loy. Towards robust blind face restoration with codebook lookup transformer. *arXiv preprint arXiv:2206.11253*, 2022. 1, 2, 7