ENTED: Enhanced Neural Texture Extraction and Distribution for Reference-based Blind Face Restoration

Yuen-Fui Lau
HKUST
yflauad@connect.ust.hk

Tianjia Zhang
HKUST
tzhangbl@connect.ust.hk

Zhefan Rao
HKUST
zraoac@connect.ust.hk

Qifeng Chen
HKUST
cqf@ust.hk

Abstract

We present ENTED, a new framework for blind face restoration that aims to restore high-quality and realistic portrait images. Our method involves repairing a single degraded input image using a high-quality reference image. We utilize a texture extraction and distribution framework to transfer high-quality texture features between the degraded input and reference image. However, the StyleGAN-like architecture in our framework requires high-quality latent codes to generate realistic images. The latent code extracted from the degraded input image often contains corrupted features, making it difficult to align the semantic information from the input with the high-quality textures from the reference. To overcome this challenge, we employ two special techniques. The first technique, inspired by vector quantization, replaces corrupted semantic features with high-quality code words. The second technique generates style codes that carry photorealistic texture information from a more informative latent space developed using the high-quality features in the reference image's manifold. Extensive experiments conducted on synthetic and real-world datasets demonstrate that our method produces results with more realistic contextual details and outperforms state-of-the-art methods. A thorough ablation study confirms the effectiveness of each proposed module.

1. Introduction

Blind face restoration (BFR) is a technique in computational photography that focuses on transforming low-quality facial images into high-quality ones, even when the type of degradation is unknown. This process is crucial for those who wish to improve the quality of their facial images. However, the task of accurately reconstructing facial

details from a low-quality image can be quite challenging due to the loss of specific identity information. To tackle this issue, we consider the use of a high-quality reference image of the same individual. This approach leads us to explore the concept of reference-based blind face restoration, where our goal is to utilize the information from a high-quality reference image to enhance the restoration process.

The technique of reference-based super-resolution (RefSR) has been gaining interest recently. It enhances the quality of a low-resolution input by incorporating high-quality semantic details from a reference image. However, if the features of the reference image are not managed correctly, it can result in under-utilization or incorrect use of the reference. To overcome these challenges, we employ a texture extraction and distribution framework that can be trained with attention reconstruction loss to improve the accuracy and the use of HQ features throughout the texture transfer process. The texture extraction and distribution framework was successfully applied to the controllable human image synthesis task [26], and we extend this concept to our reference-based blind face restoration framework.

The latent representation of low-quality (LQ) input often contains incorrect information during the texture distribution process. Simply applying the texture extraction and distribution framework is not enough to produce high-fidelity images. To overcome this challenge, we utilize a vector-quantization (VQ) technique. This technique entails replacing degraded latent features with high-quality latent codes obtained from a high-quality (HQ) dictionary. By directly substituting these codes, we are able to close the disparity between low and high-quality latent codes, thereby offering suitable semantic guidance for texture distribution.

As shown in table 3, we notice that modulated convolutions aid in enhancing the realism of facial details during the restoration process. However, they necessitate high-quality style code representation for superior image restoration.

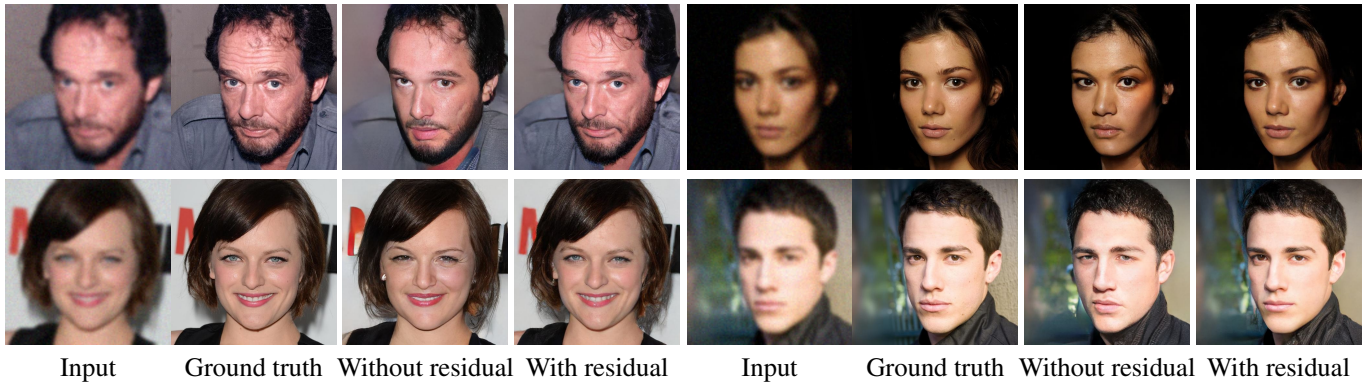| Input | Ground truth | Without residual | With residual | Input | Ground truth | Without residual | With residual |

Figure 1. When each face restoration is performed using the same reference image and experimental setup, the output without a residual connection tends to degrade the facial identity. However, when a residual connection is used, the result exhibits superior facial details that align more closely with the original image.

Generating meaningful style codes from a high-dimensional latent space is challenging [30], particularly with an imperfect and noisy latent representation of the degraded input image. To tackle this, we refine the latent space using the cross-attention technique [4] and information from the reference prior. This cross-comparison and refinement of the low-quality latent features with the HQ latent codes from the reference prior yield an enriched latent space. This assists in creating HQ style codes with more significant texture detail. Different from previous works [24, 27] for blind face restoration, which typically infers the style code from a single information source, we utilize a reference-based method, using the HQ features from the reference image to refine the latent space. This results in a higher quality style code generation with fewer flawed semantic features and biases.

We also recognize that a degraded input image loses not just texture information but essential structural content information, crucial for maintaining fidelity. As demonstrated in figure 1, insufficient fidelity information or incomplete inference for fidelity features leads to a significant loss of facial identity. To counteract this, we add residual connections to enhance the flow of fidelity information between the decoder and content encoder, thereby preserving identity.

Our contribution can be summarized as follows:

- We apply a latent space refinement technique that can utilize the information from the high-quality reference prior for generating style code that carries high-quality semantic details.

- To facilitate the texture distribution process, we modify the texture extraction and distribution framework by adopting the vector-quantization technique.

- Extensive experiments show that our model advances the state-of-the-art method in blind face restoration. It

can handle badly corrupted face images from both real-world and synthetic data.

2. Related Work

2.1. Blind Face Restoration

The goal of Blind Face Restoration (BFR) is to restore high-quality faces that have suffered from complicated and unknown degeneration. In recent years, it has gotten a lot of attention. Since Human perceptions are more sensitive to facial images than other image domains, hence necessitating more solid and thorough image generation manipulation. Most high-resolution face restoration methods [9, 27, 34] use maximum likelihood estimation to rebuild realistic facial features and adversarial learning to produce a natural image distribution. Prior face restoration methods often make use of face-specific priors, such as generative priors [8, 24, 27, 32, 33]. These techniques typically use a pre-trained face generative adversarial network such as StyleGAN [13, 14] and embed the latent representation of the degraded image into the GAN latent space, which encapsulates rich and diverse high-quality features like facial texture and facial geometry, etc. Some of them are done by projecting the degraded face straight into the latent space [32, 33] or exploring a latent vector with a costly target-specific optimization [24]. Directly embedding the latent representation of a degraded input image into the latent space of a pre-trained GAN model, however, tends to generate overly smooth face images and inaccurate facial details because the latent representation of a degraded input image carries noise and leads to the formation of imperfect latent codes with erroneous information. Another approach explores geometric priors, for which facial landmarks [3, 15], face parsing maps [1, 28], and facial component heat-maps [39] are commonly employed in blind face restoration algorithms. Because the bulk of the geometric
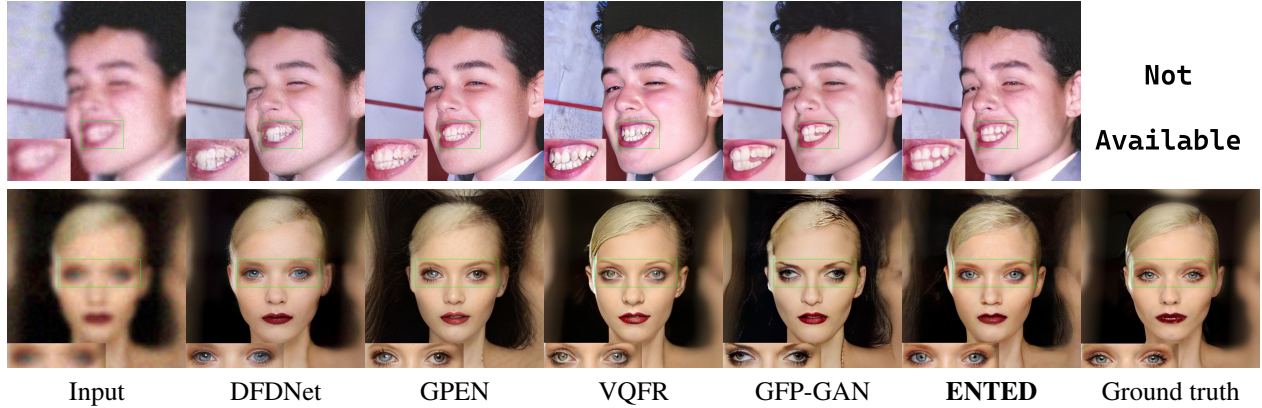
Figure 2. Visual comparison with the state-of-the-art blind face restoration methods. The first row demonstrates the visual comparisons of real-world data and the second row shows the visual comparisons of synthesized data.

information is approximated from distorted faces, and the geometric information becomes imprecise for substantially degraded input images, the restoration performance is easily influenced by the input's quality. It does not function well for generating realistic facial details in extreme cases.

2.2. Reference-based Restoration

As corrupted images often lack high-quality details, the reference-based image restoration approach seeks to extract high-quality features from the reference image to be transferred to the low-quality image. Unlike Single Image Super-Resolution (SISR), which does not include extra information, the Reference-based Image Super-Resolution (RefSR) task restores damaged input images by transferring high-quality features from reference images. In SRNTT [41], the correspondence matching was done based on the extracted features from the pre-trained VGG network. Xie et al. [36] and Yang et al. [35] employed learnable feature extractors that undergo end-to-end training with the primary restoration network. However, their correspondences were calculated only using contents and appearances. Recent works like Dogan et al. [6] employ a warper sub-network to align the content of the input and reference images in order to perform feature fusion. Li et al. [21] proposed an integrated framework for face restoration, which includes a warping subnetwork to predict flow fields for transferring features and an extra reconstruction subnetwork to utilize the warped features and then recover the clear images. Subsequently, Li et al. [20] increase the restoration quality by using more reference face images and an adaptive mechanism to fuse the high-quality guidance features and low-quality features. Recently Li et al. [22] turn the high-quality features into two dictionaries encoding generic high-quality prior information and person-specific features, respectively. Texture mapping methods [23, 36] maps the high-quality texture from the reference image to the corrupted input image. In order to align the features of the input and reference images, Yang et al. [36] and Zhang et al. [41] adopt the Patch-Matching technique. However, there are transformation gaps and resolution gaps between the input and reference images. [11] performs correspondence matching in a different way. It employs a contrastive correspondence network and knowledge distillation to bridge the gap.

3. Method

3.1. Neural Texture Extraction and Distribution

One of the most difficult aspects of texture transfer is precisely reassembling the texture extracted from the reference images and properly mapping them to the degraded input images. To address this issue, neural texture extraction and distribution [26] based on the attention mechanism is introduced. The benefit of using [26] for texture transfer is that it helps to learn specific extraction and distribution kernels to map appropriate semantic texture components between the input and reference images. During the extraction process, high-quality semantic textures extracted from the reference image act as keys and values, while the features obtained from the low-quality input image work as queries, providing clues that guide appropriate high-quality semantic textures to enhance the corrupted textures in the input image.

Texture Extraction Process. Firstly, a texture encoder E_{ref} is employed to extract a hierarchy level of feature maps from the reference image I_{ref}, denoted as F_{ref}^i, where i represents the layer index. Then we utilize a set of texture extraction kernels W_e^i to project the feature maps into latent texture codes, $C_e^i = W_e^i F_{ref}^i$, where $W_e^i \in \mathbb{R}^{k \times c}$ and $F_{ref}^i \in \mathbb{R}^{c \times hw}$. Here we compress the height and width dimensions into one spatial dimension to better apply matrix multiplication. Softmax is applied on C_e^i along the spatial dimension to obtain semantic texture extraction matrix
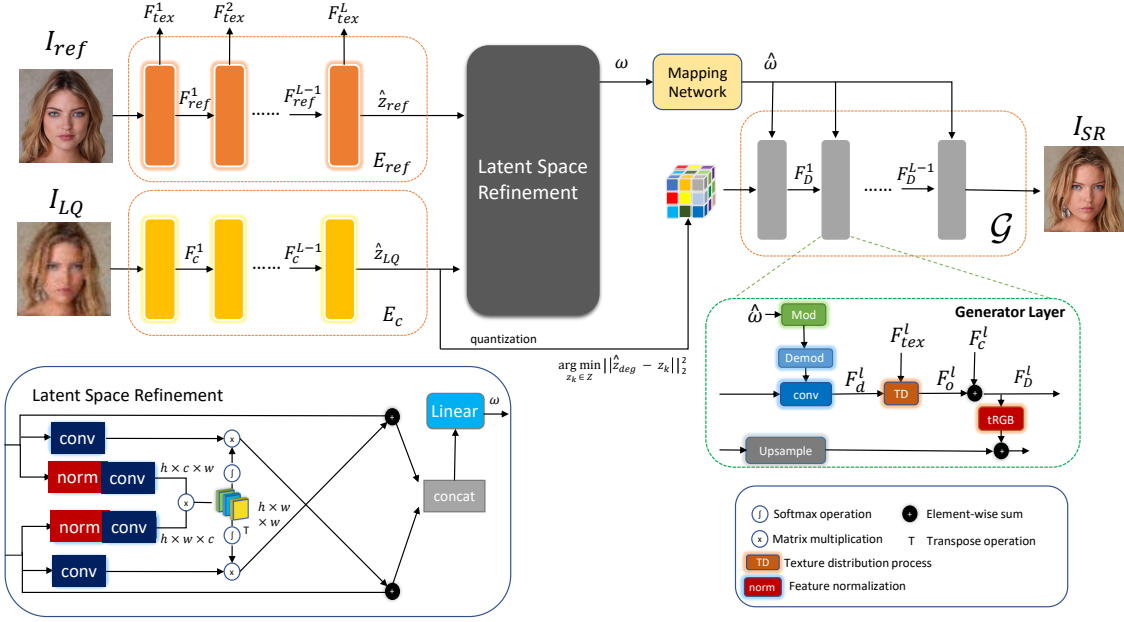
Figure 3. A summary of our pipeline. Using reference features, we construct a high-quality image by transferring high-quality reference details (Neural Texture Extraction and Distribution Process) and repairing (Application of VQ Dictionary and Latent Space Refinement) distorted semantic information in degraded input images.

$\tilde{C}_e^i$. The neural texture $F_{tex}^i \in \mathbb{R}^{k \times c}$ is obtained as:

$$\tilde{C}_e^i(p, q) = \frac{exp(C_e^i(p, q))}{\sum_{q=1}^{hw} exp(C_e^i(p, q))},$$
$$F_{tex}^i = (\tilde{C}_e^i)(f_{1 \times 1}(F_{ref}^i))^\mathsf{T}, \quad (1)$$

where $f_{1 \times 1}(\cdot)$ indicates a 1×1 convolution, p and q are the indices along the channel and spatial dimensions, respectively.

Texture Distribution Process. To distribute the extracted semantic texture to the low-quality input image $I_{LQ} \in \mathbb{R}^{3 \times H \times W}$, a content encoder E_c is first applied to I_{LQ} to get the high-level semantic feature map of the input image. The final output of E_c is then directly passed into a decoder G. Let F_d^i denote the output feature map from the ith layer of the decoder G. The distribution process starts in the same manner as the extraction operation. The latent codes $C_d^i = W_d^i F_d^i$ is obtained by applying texture distribution kernels $W_d^i \in \mathbb{R}^{k \times c}$ on $F_d^i \in \mathbb{R}^{c \times hw}$. The resulting feature map $F_o^i \in \mathbb{R}^{c \times hw}$ is obtained by first adopting softmax function on C_d^i along the channel dimension and then multiplying the transpose of extracted neural texture F_{tex} with the semantic texture distribution matrix $\tilde{C}_d^i$. The overall distribution process is formulated as follows:

$$\tilde{C}_d^i(p, q) = \frac{exp(C_d^i(p, q))}{\sum_{p=1}^{k} exp(C_d^i(p, q))},$$
$$F_o^i = (F_{tex}^i)^\mathsf{T}(\tilde{C}_d^i). \quad (2)$$

3.2. Vector-Quantized Dictionary

Some prior works [9, 19, 34] have introduced a vector quantized dictionary to replace faulty features in the latent space for blind face restoration. They consider the dictionary a high-quality low-level feature bank that stores meaningful facial representations. Different from prior work [2,9,34,42,43] that consider the dictionary as a feature bank for facial representation, we treat the HQ dictionary as a reservoir of semantic queries that can provide comprehensive guidance during the texture distribution process. We define the HQ dictionary with K discrete tokens as $\mathcal{Z} = \{z_k\}_{k=1}^K$. The feature map of the last layer obtained from the content encoder $Z_{LQ} = E_c(I_{LQ}) \in \mathbb{R}^{c \times h \times w}$ could be considered as a set of vectors of length cc expanded along the spatial dimension, denoted by $\{z_{LQ}^{i,j}\}_{i=1,j=1}^{h,w}$. Subsequently, each vector can be enhanced by replacing it with the nearest entry $\hat{z}_k \in \mathbb{R}^c$ in the HQ dictionary $\mathcal{Z}$, which has the minimum l_2 distance with the feature vector $\hat{z}_{LQ}^{i,j} \in \mathbb{R}^c$ located at spatial position (i, j), The overall quantization process $q(\cdot)$ is defined as

$$z_q^{i,j} = q(z_{LQ}^{i,j}, \hat{z}_k) = \arg\min_{z_k \in \mathcal{Z}} \|z_{LQ}^{i,j} - z_k\|_2^2, \quad (3)$$

where $z_q^{ij} \in \mathbb{R}^c$ is the quantized feature vector, and all the quantized vectors could be rearranged back as a refined feature map $\hat{Z}_{LQ}$.

Since the quantization process is non-differentiable, we follow a similar optimization strategy as VQ-GAN [7]. A

straight-through gradient estimator is employed to simply copies the gradients from the decoder to the content encoder and optimize via the quantization loss $\mathcal{L}_q$:

$$\mathcal{L}_q = \sum_{i=1}^{h}\sum_{j=1}^{w}\|sg[z_{LQ}^{i,j}] - z_q^{i,j}\|_2^2 + \beta\|sg[z_q^{i,j}] - z_{LQ}^{i,j}\|_2^2, \quad (4)$$

where β is the commitment weight and $sg[\cdot]$ denotes the stop gradient operator. Throughout all the experiments, the commitment weight is set to 0.25.

Given a degraded input image $I_{LQ} \in \mathbb{R}^{3 \times H \times W}$, the degraded latent representation $\hat{z}_{LQ} = E_c(I_{LQ}) \in \mathbb{R}^{c \times h \times w}$, where $h = H/s$, $w = W/s$, and s is the down-sampling factor. Following previous works [7, 9], we choose $s = 32$ for 512×512 input. Moreover, in order to maximize the usage of the HQ dictionary, we adopt [18] and periodically re-initialize the dictionary with the k-means algorithm as what it did in [9]. The number of clusters k is chosen to be 1024.

3.3. Latent Space Refinement

Our method applies style code and modulated convolution proposed in the StyleGAN [13] architecture for restoration. However, generating the style code $\omega \in \mathcal{W} \subseteq \mathbb{R}^c$ using only high-quality features from the reference image leads to the loss of fidelity information, which is essential for retaining facial identity. Nonetheless, generating the style code using only latent features from degraded input introduces erroneous information, which is harmful to the perceptual quality of the final restored image. To minimize the fidelity change while maintaining high perceptual quality, we generate the style code ω from the refined latent space of $\hat{z}_{LQ}$ and $\hat{z}_{ref}$, which is achieved by exchanging information between them through a cross attention technique. We follow a similar formulation as [4] with a bit of modification on the normalization method, which results in more stable training. We define the scalar dot product attention [31] as ϕ with given query Q, key K and value V:

$$\phi_\lambda(Q, K, V) = Softmax(\frac{QK^\intercal}{\sqrt{d}})V, \quad (5)$$

where λ indicates applying the Softmax function along λ axis and d is a normalization constant. Given an input image $I_{LQ} \in \mathbb{R}^{3 \times H \times W}$ and a reference image $I_{ref} \in \mathbb{R}^{3 \times H \times W}$, we pass them into the content encoder E_c and the texture encoder E_{ref} separately to get the semantic latent representation $\hat{Z}_{LQ} = E_c(I_{LQ}) \in \mathbb{R}^{c \times h \times w}$ and $\hat{Z}_{ref} = E_{ref}(I_{ref}) \in \mathbb{R}^{c \times h \times w}$. Then, we apply feature normalization [26] FN on $\hat{z}_{LQ}$ and $\hat{z}_{ref}$ for stable training. The feature normalization is formulated as follows:

$$FN_\Lambda(z) = \frac{x - \mathbb{E}_\Lambda(x)}{\sqrt{\sum_{i=1}^{\Lambda} x_i^2 + \epsilon}}, \quad (6)$$

where Λ indicates the operation along the Λ axis (e.g., the channel axis), ϵ is a constant to avoid numeric issues, and $\mathbb{E}_\Lambda(x)$ represents the expectation value of x along the Λ axis. After that, the quantized feature maps are processed by 1x1 convolution and reshaped along the spatial dimension to get the feature matrices $\tilde{Z}_{LQ}, \tilde{Z}_{ref} \in \mathbb{R}^{c \times hw}$ for attention calculation:

$$\begin{aligned} \hat{V}_{LQ} &= \phi_{hw}(\tilde{Z}_{LQ}, \tilde{Z}_{ref}, \tilde{Z}_{ref}), \\ \hat{V}_{ref} &= \phi_{hw}(\tilde{Z}_{ref}, \tilde{Z}_{LQ}, \tilde{Z}_{LQ}), \end{aligned} \quad (7)$$

where $\hat{V}_{LQ}, \hat{V}_{ref}$ are the refined feature maps encoding exchanged information.

$\hat{V}_{LQ}, \hat{V}_{ref}$ are then reshaped to normal size with $c \times h \times w$ and re-scaled by learn-able scale factors $\gamma_{LQ} \in \mathbb{R}^c$ and $\gamma_{ref} \in \mathbb{R}^c$ respectively. To make the best use of high-quality information in the reference image, we concatenate both the input and reference's refined latent representations along the feature channel c to achieve a more informative and complete inference for the style code generation. The style code $\omega \in \mathbb{R}^c$ is then obtained by embedding refined latent representations as

$$\omega = \Psi(concat[\gamma_{LQ}\hat{V}_{LQ} + \hat{Z}_{LQ}, \gamma_{ref}\hat{V}_{ref} + \hat{Z}_{ref}]). \quad (8)$$

where $\Psi(\cdot)$ is a linear layer.

3.4. Loss Functions

To train our model, we employ non-saturating adversarial loss [14] for restoring photo-realistic images, perceptual loss [12] to evaluate the style difference, quantization loss [7] as discussed in Section 3.2 and attention reconstruction loss [26] to supervise the texture extraction and distribution process.

Adversarial loss $\mathcal{L}_{adv}$. [14] To motivate the generator to prioritize entries in the natural image manifold and generate realistic textures, we train the generator with non-saturating adversarial loss. Denote the discriminator as $\mathcal{D}$ and the generator as $\mathcal{G}$. We use χ to denote the set of degraded images. Given a low-quality input image I_{LQ} and a reference image I_{ref}, the adversarial loss is calculated as

$$\mathcal{L}_{adv} = \mathbb{E}_{I_{LQ}\sim\chi}log\Big(1+exp\Big(-\mathcal{D}\big(\mathcal{G}(I_{LQ}, I_{ref})\big)\Big)\Big). \quad (9)$$

Perceptual loss $\mathcal{L}_{percep}$. [12] Given a degraded image I_{LQ} and a reference image I_{ref}, the restoration model $\mathbb{G}$ generates a restored image $I_{SR} = \mathbb{G}(I_{LQ}, I_{ref})$. The restored image I_{SR} is then compared with corresponding ground truth image I_{gt} and evaluated by the l_1 difference between the pretrained VGG-19 [29] activations:

$$\mathcal{L}_{percep} = \sum_j \|\varphi_j(I_{SR}) - \varphi_j(I_{gt})\|_1, \quad (10)$$

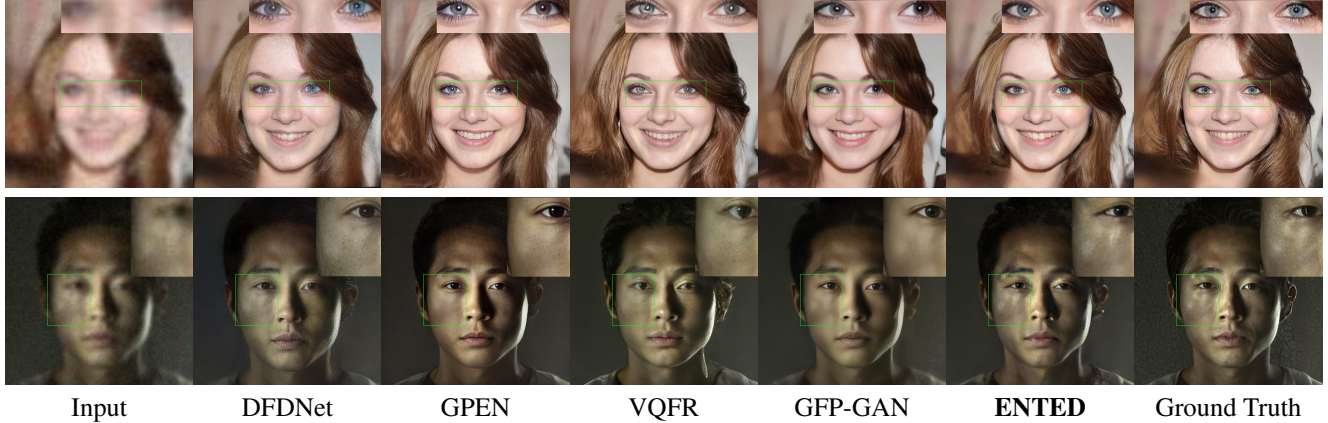| Input | DFDNet | GPEN | VQFR | GFP-GAN | **ENTED** | Ground Truth |

Figure 4. The first row demonstrates a 8x blind face restoration. Our results display fewer distortions and align more closely with the original images, particularly in terms of the color of the pupils. The second row shows a 4x blind face restoration. Our approach reveals a skin texture that is more detailed and refined compared to what is achieved by current state-of-the-art methods.

where φ_j indicates the jth activation map of VGG-19.

Attention reconstruction loss $\mathcal{L}_{att}$. We use attention reconstruction loss [26] to penalize inaccurate texture extracted from the reference image to facilitate the performance of the texture extraction and distribution process. As mentioned in Section 3.1, at each layer l of the encoders, we can obtain an extraction matrix $\tilde{C}_e^l \in \mathbb{R}^{k \times HW/s^2}$ and a distribution matrix $\tilde{C}_d^l \in \mathbb{R}^{k \times HW/s^2}$, where H and W are the initial height and width of the input image and s is the size downsample ratio at the l-th layer. Therefore the attention reconstruction loss is calculated as:

$$\mathcal{L}_{att} = \sum_l \|I_{gt}^{l\downarrow} - I_{ref}^{l\downarrow}(\tilde{C}_e^l)^\intercal \tilde{C}_d^l\|_1, \qquad (11)$$

where $I_{gt}^{l\downarrow_s}$, $I_{ref}^{l\downarrow_s} \in \mathbb{R}^{3 \times HW/s^2}$ represent the bilinearly downsampled and dimension-compressed ground truth image and reference image corresponding to the lth layer, respectively.

Total loss $\mathcal{L}_{total}$. The overall loss is the sum of the losses mentioned above:

$$\mathcal{L}_{total} = \lambda_{adv}\mathcal{L}_{adv} + \lambda_{percep}\mathcal{L}_{percep} + \lambda_q\mathcal{L}_q + \lambda_{att}\mathcal{L}_{att}, \qquad (12)$$

where λ_{adv}, λ_{percep}, λ_q and λ_{att} are weighting factors corresponding to adversarial loss, perceptual loss, quantization loss, and attention reconstruction loss respectively.

4. Experiments

4.1. Experimental Setup

We train our model from scratch on the FFHQ dataset [13], which contains 70,000 high-resolution images. As for the reference base blind face restoration, it is hard to find a corresponding reference image for each photo in a huge dataset. Therefore, we adopt [30] to generate reference images during training by altering pose and age. All images are resized to 512×512 during training. Additionally, we employ the degradation model used in [33] to generate low-quality input images. The degradation model simulates low-quality images found in the actual world as a result of noises, defocus, jpeg compression, long-range sensing, and their combinations. As we propose a reference-based method for blind face restoration, which requires high-quality reference images, we use the CelebA-HQ test set and select the relevant image from real data distribution as the reference image for each individual test image in the dataset. There is no overlap between test data, including the reference images, and our training data. Table 1 quantitatively demonstrates the 4x blind face restoration results conducted on this modified CelebA-HQ test set which consists of 2398 image pairs with corresponding reference images for quantitative and visual evaluation. As demonstrated in Table 2, we also evaluate our model performance quantitatively on real-world data. For **LFW-Test**, we carefully select 1,680 images with corresponding relevance from their own data distribution. For **CelebChild-Test**, which consists of 180 child images, we take the adult image as the reference image for each corresponding child image. For **WebPhoto-Test**, there are a total of 407 images in this dataset. As it is a dataset constructed by crawling images from the internet, it is difficult to find a corresponding reference image for each individual test data from the dataset. Therefore, we employ Tov et al. [30] to generate reference images for each test image, the same as what we have done for the FFHQ training set.

Implementation Setting. The model generates images of 512×512 resolution with a given degraded input of size 512×512. The training batch size is set to 4. We augment

Table 1. Quantitative comparison with state-of-the-art blind face restoration methods on the modified CelebA-HQ test set. **Red** indicates the best performance while **Blue** indicates the second best performance

Method	LPIPS ↓	NIQE ↓	FID ↓	Deg. ↓	PSNR ↑	SSIM ↑
HiFaceGAN [37]	0.3789	4.96	40.99	35.13	**25.57**	0.6503
PSFRGAN [1]	0.2779	4.15	24.04	36.12	25.11	0.6272
DFDNet [19]	0.2910	5.41	19.47	30.48	24.69	0.6375
GPEN [38]	0.2439	4.08	13.79	**28.65**	25.11	0.6313
VQFR [9]	0.2274	**3.68**	**12.02**	32.91	24.25	0.6371
GFP-GAN [33]	**0.2224**	4.23	13.63	32.84	25.34	**0.6694**
ENTED (Ours)	**0.2156**	**3.60**	**12.80**	**27.77**	**25.87**	**0.6664**

the data by applying the degradation model. We employ the pretrained vector-quantized (VQ) dictionary from Gu et al. [9]. The VQ dictionary contains 1024 entries with a code length of 256 for each latent code. The VQ dictionary is re-estimated via k-means clustering for every 2000 iterations. We also adopt 10,000 warm-up iterations before starting to adopt the VQ dictionary. Our model was trained using the Adam optimizer [16] for 400,000 iterations, using a learning rate of 2×10^{-3} for the generator and 1.882×10^{-3} for the discriminator. The weighting factor λ_{adv}, λ_{percep}, λ_q, λ_{att} are set to 1.5, 1.0, 1.0 and 15.0 respectively.

Evaluation Metrics. We qualitatively evaluate our models based on the Peak Signal-to-Noise Ratio (PSNR), the Structural Similarity Index Measure (SSIM), the Learned Perceptual Image Patch Similarity (LPIPS) [40], the Fréchet Inception Distance (FID) [10], the Natural Image Quality Evaluator (NIQE) [25] and the identity metric indicated as Deg., which calculates the face embedding angle between the restored face and the ground truth face using the embedding angle of ArcFace [5]. For the LPIPS evaluation model, we adopt pretrained AlexNet [17] to evaluate the LPIPS value for all our experiments.

4.2. Comparison with Blind Face Restoration Baselines

Our model is compared to several state-of-the-art blind face restoration baselines. We compare our model with GFP-GAN [33], VQFR [9], GPEN [38], DFDNet [19]. PSFRGAN [1] and HiFaceGAN [37]. We adopt the official code and pre-trained weight for all the qualitative and quantitative comparisons.

4.2.1 Comparison on Modified CelebA-HQ Test

The experimental results are the outcomes of 4x blind face restoration. As seen in Table 1, **ENTED** has the best LPIPS and NIQE. The improvement in LPIPS demonstrates that our method generates images that are perceptually nearer to ground truth than other cutting-edge methods. Furthermore,

our model achieves the best NIQE, suggesting that our approach can generate face images with better high-quality details than other state-of-the-art approaches. The restored image is closer to the natural image distribution. Besides, our method preserves richer identity information from the given input, as indicated by the smallest face feature embedding angle degree. The best PSNR value in Table 1 also suggests that our results have a closer pixel-wise value with the ground truth.

4.2.2 Comparison on Real-world Datasets

We conduct experiments on Modified **_LFW-Test_**, **_CelebChild-Test_** and **_WebPhoto-Test_** separately. To compare our method to state-of-the-art methods in the extreme scenario for real-world data, we impose a bit of deterioration on the input data using the degradation model [33]. According to the quantitative findings in Table 2, **ENTED** has relatively better FID results over three real-world datasets, indicating that our method can recover images near the true face distribution. On the other hand, **ENTED** outperforms the majority of other state-of-the-art methods across three real-world datasets in terms of NIQE.

4.3. Ablation Study

4.3.1 Effect of Residual Connections

The residual connection plays a crucial role in transmitting fidelity information from the content encoder to the decoder, ensuring the preservation of important identity details in the final restored image. By comparing Model 1 and Model 4 in Table 3, we observe that Model 1 performs better in terms of LPIPS and FID. This indicates that the output from the model without the residual connection is perceptually distant from the ground truth and true face distribution. The absence of the residual connection leads to a significant decrease in fidelity. However, by incorporating the residual connection into the decoder, the content encoder is able to provide complete fidelity information, resulting in facial details that are consistent with the ground truth distribution.

Table 2. Quantitative Comparison with state-of-the-art blind face restoration methods on modified real-world LFW, CelebChild, WebPhoto dataset. **Red** indicates the best performance while **Blue** indicates the second best performance

Dataset	LFW-Test		CelebChlid		WebPhoto	
Method	FID ↓	NIQE ↓	FID ↓	NIQE ↓	FID ↓	NIQE ↓
HiFaceGAN [37]	95.32	4.94	130.62	4.99	123.59	5.48
PSFRGAN [1]	70.57	4.89	114.95	4.80	90.19	4.34
DFDNet [19]	75.11	5.81	111.09	5.81	105.30	5.70
GPEN [38]	70.74	4.62	110.70	4.33	95.60	4.78
VQFR [9]	**70.21**	**3.98**	**106.18**	**3.74**	**78.68**	**3.77**
GFP-GAN [33]	**69.04**	4.77	113.30	4.55	86.79	4.40
ENTED (Ours)	77.76	**3.78**	**104.37**	**3.79**	**76.24**	**3.91**

Table 3. Ablation study results on Modified CelebA-HQ test set. *Skip*: applying skip connection; *Style*: applying modulated convolution; *VQ*: applying vector-quantized dictionary; *Refine*: applying latent space refinement. **Red** indicates the best performance while **Blue** indicates the second best performance

Models	Configuration				CelebA-Test		
	Skip	Style	VQ	Refine	LPIPS ↓	FID ↓	NIQE ↓
Model 1		✓	✓		0.2762	16.59	3.64
Model 2	✓		✓		0.2185	**12.20**	3.73
Model 3	✓	✓			**0.2121**	12.89	3.81
Model 4	✓	✓	✓		0.2178	13.08	**3.63**
ENTED	✓	✓	✓	✓	**0.2156**	**12.80**	**3.60**

4.3.2 Effect of Applying Modulated Convolution

When comparing Model 2 and Model 4, we observe that Model 4 achieves lower LPIPS and NIQE values. This suggests that Model 4 is perceptually closer to the ground truth distribution and exhibits more high-quality details. The utilization of modulated convolution promotes the generation of high-frequency details, resulting in more photo-realistic images. However, in Model 4, there is no refinement of the latent space for degraded features, and the style code is directly generated from the degraded input's latent space. As a result, any defective features present in the input affect all modulated features, leading to inaccurate inference for fidelity details. This discrepancy contributes to a slightly higher FID in Model 4, indicating that the output of Model 4 deviates from the true face features.

4.3.3 Effect of VQ Dictionary

Since we are utilizing the high-quality details from the reference image, the effectiveness of the texture transfer process is vital for generating high-quality facial details in the restored image. The VQ dictionary provides high-quality semantic queries to help in texture distribution, resulting in effective texture transfer. When Model 3 and Model 4 are

compared in Table 3, it is shown that the performance in NIQE improves when the VQ dictionary is included, indicating that the existence of VQ dictionary helps to produce more high-quality details.

4.3.4 Effect of Latent Space Refinement

The quality of the final restored image is closely related to the quality of the generated style code. As shown in Table 3, Model 4 does not employ the latent space refinement procedure and instead generates the style code straight from the latent space of degraded input. Erroneous information in the degraded input's latent space impacts the quality of the style code unless the input's latent space is rectified using high-quality features from the reference prior. Refining the Latent space using reference prior aids in the generation of high-quality style code, resulting in photorealistic images with more high-quality details (e.g., best NIQE value, which favors high-quality details) and hence improved performance.

5. Conclusion

In this study, we have introduced a reference-based approach for blind face restoration. To address the blind face restoration problem, we have expanded and adapted the texture extraction and distribution framework proposed in controllable human image synthesis task. We have explored the network architecture and made modifications by incorporating a skip connection and utilizing modulated convolution. Additionally, we have integrated the VQ dictionary to facilitate the texture distribution process. To enhance the image restoration capabilities of the model, we have incorporated two types of attention mechanisms. A double attention mechanism has been employed for the texture transfer process, while the cross-attention technique has been utilized to refine the latent space and generate high-quality style codes.

References

[1] 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*, pages 11896–11905, 2021. 2, 7, 8

[2] Chaofeng Chen, Xinyu Shi, Yipeng Qin, Xiaoming Li, Xi-aoguang Han, Tao Yang, and Shihui Guo. Blind image super resolution with semantic-aware quantized texture prior. *arXiv preprint arXiv:2202.13142*, 2022. 4

[3] Yu Chen, Ying Tai, Xiaoming Liu, Chunhua Shen, and Jian Yang. Fsrnet: End-to-end learning face super-resolution with facial priors. In *CVPR*, pages 2492–2501, 2018. 2

[4] Xiaojie Chu, Liangyu Chen, and Wenqing Yu. Nafssr: Stereo image super-resolution using nafnet. In *CVPR*, pages 1239–1248, 2022. 2, 5

[5] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, pages 4690–4699, 2019. 7

[6] Berk Dogan, Shuhang Gu, and Radu Timofte. Exemplar guided face image super-resolution without facial landmarks. In *CVPRW*, pages 0–0, 2019. 3

[7] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, pages 12873–12883, 2021. 4, 5

[8] Jinjin Gu, Yujun Shen, and Bolei Zhou. Image processing using multi-code gan prior. In *CVPR*, pages 3012–3021, 2020. 2

[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. *arXiv preprint arXiv:2205.06803*, 2022. 2, 4, 5, 7, 8

[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. *NIPS*, 30, 2017. 7

[11] Yuming Jiang, Kelvin CK Chan, Xintao Wang, Chen Change Loy, and Ziwei Liu. Robust reference-based super-resolution via c2-matching. In *CVPR*, pages 2103–2112, 2021. 3

[12] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, pages 694–711. Springer, 2016. 5

[13] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, pages 4401–4410, 2019. 2, 5, 6

[14] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, pages 8110–8119, 2020. 2, 5

[15] 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. 2

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

[17] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017. 7

[18] Adrian Łańcucki, Jan Chorowski, Guillaume Sanchez, Ricard Marxer, Nanxin Chen, Hans JGA Dolfing, Sameer Khurana, Tanel Alumäe, and Antoine Laurent. Robust training of vector quantized bottleneck models. In *IJCNN*, pages 1–7. IEEE, 2020. 5

[19] Xiaoming Li, Chaofeng Chen, Shangchen Zhou, Xianhui Lin, Wangmeng Zuo, and Lei Zhang. Blind face restoration via deep multi-scale component dictionaries. In *ECCV*, pages 399–415. Springer, 2020. 4, 7, 8

[20] 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*, June 2020. 3

[21] Xiaoming Li, Ming Liu, Yuting Ye, Wangmeng Zuo, Liang Lin, and Ruigang Yang. Learning warped guidance for blind face restoration. In *ECCV*, September 2018. 3

[22] Xiaoming Li, Shiguang Zhang, Shangchen Zhou, Lei Zhang, and Wangmeng Zuo. Learning dual memory dictionaries for blind face restoration. *TPAMI*, 2022. 3

[23] Liying Lu, Wenbo Li, Xin Tao, Jiangbo Lu, and Jiaya Jia. Masa-sr: Matching acceleration and spatial adaptation for reference-based image super-resolution. In *CVPR*, pages 6368–6377, 2021. 3

[24] 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*, pages 2437–2445, 2020. 2

[25] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012. 7

[26] Yurui Ren, Xiaoqing Fan, Ge Li, Shan Liu, and Thomas H Li. Neural texture extraction and distribution for controllable person image synthesis. In *CVPR*, pages 13535–13544, 2022. 1, 3, 5, 6

[27] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *CVPR*, pages 2287–2296, 2021. 2

[28] Ziyi Shen, Wei-Sheng Lai, Tingfa Xu, Jan Kautz, and Ming-Hsuan Yang. Deep semantic face deblurring. In *CVPR*, pages 8260–8269, 2018. 2

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

[30] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. *ACM Transactions on Graphics (TOG)*, 40(4):1–14, 2021. 2, 6

[31] 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. 5

[32] Ziyu Wan, Bo Zhang, Dongdong Chen, Pan Zhang, Dong Chen, Jing Liao, and Fang Wen. Bringing old photos back to life. In *CVPR*, pages 2747–2757, 2020. 2

[33] Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. To-wards real-world blind face restoration with generative facial prior. In *CVPR*, pages 9168–9178, 2021. 2, 6, 7, 8

[34] Zhouxia Wang, Jiawei Zhang, Runjian Chen, Wenping Wang, and Ping Luo. Restoreformer: High-quality blind face restoration from undegraded key-value pairs. In *CVPR*, pages 17512–17521, 2022. 2, 4

[35] Yanchun Xie, Jimin Xiao, Mingjie Sun, Chao Yao, and Kaizhu Huang. Feature representation matters: End-to-end learning for reference-based image super-resolution. In *ECCV*, pages 230–245. Springer, 2020. 3

[36] Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Bain-ing Guo. Learning texture transformer network for image super-resolution. In *CVPR*, pages 5791–5800, 2020. 3

[37] 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. 7, 8

[38] Tao Yang, Peiran Ren, Xuansong Xie, and Lei Zhang. Gan prior embedded network for blind face restoration in the wild. In *CVPR*, pages 672–681, 2021. 7, 8

[39] Xin Yu, Basura Fernando, Bernard Ghanem, Fatih Porikli, and Richard Hartley. Face super-resolution guided by facial component heatmaps. In *ECCV*, pages 217–233, 2018. 2

[40] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018. 7

[41] Zhifei Zhang, Zhaowen Wang, Zhe Lin, and Hairong Qi. Image super-resolution by neural texture transfer. In *CVPR*, pages 7982–7991, 2019. 3

[42] Yang Zhao, Yu-Chuan Su, Chun-Te Chu, Yandong Li, Mar-ius Renn, Yukun Zhu, Changyou Chen, and Xuhui Jia. Re-thinking deep face restoration. In *CVPR*, pages 7652–7661, 2022. 4

[43] Shangchen Zhou, Kelvin CK Chan, Chongyi Li, and Chen Change Loy. Towards robust blind face restora-tion with codebook lookup transformer. *arXiv preprint arXiv:2206.11253*, 2022. 4