Unpaired Face Restoration via Learnable Cross-Quality Shift

Yangyi Dong[1*], Xiaoyun Zhang[1*†], Zhixin Wang[1], Ya Zhang[1,2], Siheng Chen[1,2], Yanfeng Wang[1,2 †]
[1]Cooperative Medianet Innovation Center, Shanghai Jiao Tong University, [2]Shanghai AI Laboratory
{twinplant,xiaoyun.zhang,dedsec_z,ya_zhang,sihengc,wangyanfeng}@sjtu.edu.cn

Abstract

Face restoration aims to recover high-quality (HQ) face images from low-quality (LQ) ones with various unknown degradations. Unpaired face restoration approaches focus on the adaptation to unseen degradations, which is a more challenging setting. Recently, generative facial priors of StyleGAN are used to improve the restoration capability of paired face restoration methods. For unpaired methods, however, using face priors is a challenge due to the lack of paired supervision. To address this issue, we take advantage of the editing capabilities of StyleGAN's latent code and propose a novel learnable cross-quality shift. The proposed learnable cross-quality shift not only introduces the generative facial priors into the unpaired framework, but also enables the straight-forward addition/subtraction in the latent feature space to achieve quality conversion. Furthermore, we design a two-branch framework with the proposed cross-quality shift to deal with unpaired data and improve the fidelity of restoration. With the unpaired framework, our method can be fine-tuned on images with unseen degradation. Experimental results show that (i) compared to state-of-the-art methods, our method improves performances under moderate and severe degradation situations; and (ii) both the proposed learnable cross-quality shift and the two-branch framework benefit the restoration performance.

1. Introduction

Face restoration aims to restore LQ images with different degradations, such as low resolution [4, 15], noise [25], and blur [20]. It has many applications, such as video surveillance [6] and face recognition [16]. Face restoration usually includes two approaches: paired or unpaired, according to whether it needs the associated pairs of HQ and LQ images or not.

In most paired face restoration studies [11, 15, 25], the training dataset is created by using a specific degradation operation (*e.g.*, bicubic, Gaussian noise, Gaussian blur, JPEG compression) on high-quality images. Recent methods [3, 11, 23] randomly use various degradation combinations on high-quality images to generate training pairs. However, these methods are still unable to handle various degradations beyond the training set because they need paired data for training.

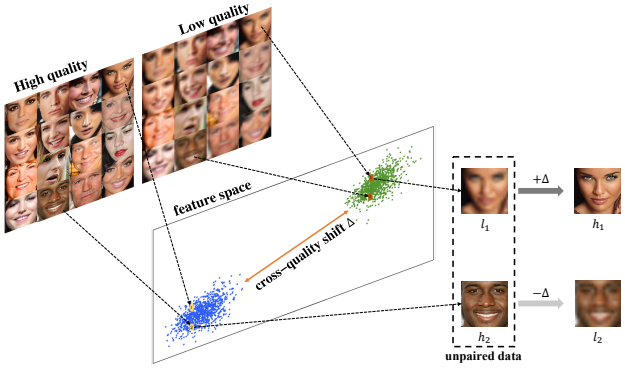

Figure 1. Low and high-quality facial images are mapped to a latent space of StyleGAN and naturally grouped into two subspaces. With a trainable cross-quality shift Δ, the restoration from an LQ facial image to its corresponding HQ version is converted to a simple addition operation in the latent space and the degradation from an HQ facial image to its corresponding LQ version is equivalent to a subtraction operation in the same latent space.

The unpaired approaches [2, 14] do not need paired data, which enables them to adapt to unseen degradation. For example, Bulat et al. [2] design two GANs to learn low-quality and high-quality images respectively. Lu et al. [14] use a KL divergence loss to regularize the distribution of degradation features. However, these methods ignore using facial priors for face restoration tasks, which limits the restoration capability for severely degraded images. How to utilize face priors in unpaired frameworks remains a challenge. Inspired by the finding that images are clustered according to their degradation in feature space [13], we consider that the essence of facial restoration is to find a transformation from the subspace of the low quality to the subspace of the high quality. Therefore, we propose a learnable cross-quality shift in the latent space of StyleGAN, which introduces the

*Equal contribution(co-first authors).
†Corresponding author.

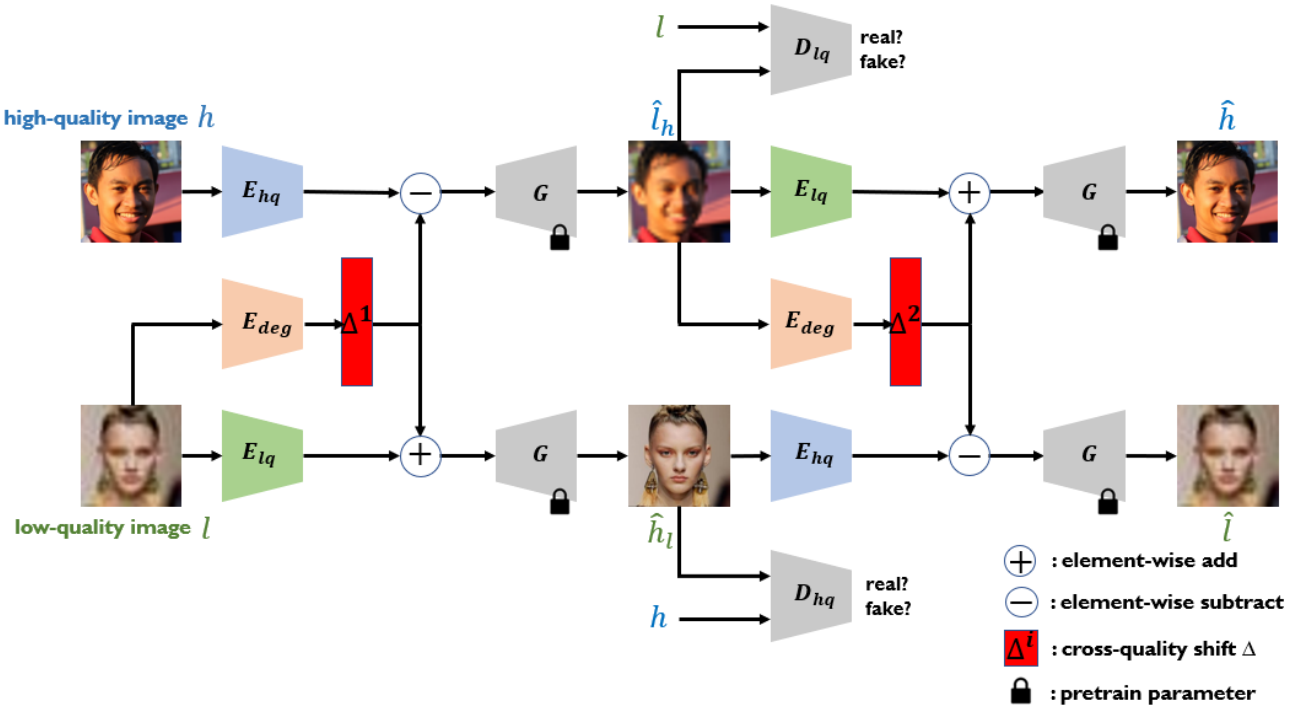

Figure 2. Overview of the proposed architecture. The data flow of HQ sample h (LQ sample l) is the upper (lower) branch. Three encoders E_{hq}, E_{lq} and E_{deg} for HQ and LQ images encode two input images and estimate the degradation to the shared latent space of a pre-trained StyleGAN. The learned cross-quality shift Δ^i is used to transfer representations from one quality level to the other. Two GAN losses distinguish $\hat{l}_h$ from LQ images, and $\hat{h}_l$ from HQ images. Other losses have not been shown in this figure.

generative facial priors into the unpaired framework.

As shown in Fig. 1, we can transform the LQ/HQ image into a corresponding HQ/LQ image through straightforward addition/subtraction in the latent space by learning a cross-quality shift. Thus, the restoration task can be solved by mapping the shifted latent codes back to images. Compared to the previous unpaired methods, the proposed learnable cross-quality shift not only introduces the generative facial priors into the unpaired framework, but also explicitly estimates the degradation in the latent space, which allows a user to tweak the shifting scale to adjust the restoration level of the restored image, alleviating the blur or oversharp problem.

To cope with unpaired data, a two-branch framework is designed. As shown in Fig. 2, the upper branch takes HQ images and the lower branch takes LQ images. The two-branch framework allows us to add constraints between h and its reconstructed image $\hat{h}$, l and its reconstructed image $\hat{l}$, which improves the fidelity of the restoration.

Extensive experiments are conducted to evaluate the performance of our proposed method. In CelebAHQ, compared to previous face restoration methods, our method improves by 2.0 % and 3.3 % in terms of the metric of LPIPS on downsampled with noise, or blur+noise, respectively.

The main contributions are summarized as follows:

- We propose a novel concept, learnable cross-quality shift, a unique translation operator that enables the conversion between two different quality levels in the latent space of StyleGAN. The proposed learnable cross-quality shift not only leverages the generative facial priors, but also allows a user to tweak the shifting scale to adjust the restoration level of the restored image.
- Based on the proposed learnable cross-quality shift, a two-branch framework is designed to deal with unpaired data and improve the fidelity of restoration.
- Extensive experiments are conducted to validate that the proposed unpaired face restoration method achieves higher perceptual quality on moderate and severe degradation images.

2. Related Works

2.1. Face Restoration

Relying on strong facial prior knowledge, face restoration methods achieve better performance than common image restoration methods. These facial priors include facial landmarks [4,9], face parsing maps [3,4,19] and facial component heatmaps [26].

To synthesize the training data that approximate the real LQ images in the wild, some face restoration methods [3, 11, 23] randomly use various degradation combinations on high-quality images to generate training pairs. PSFRGAN [3] progressively restores the input image through semantic-aware style transformation. GFPGAN [23] leverages generative facial prior provided by a pre-trained GAN. However, these methods still rely on supervision training, which means they cannot handle degradations not seen during training.

To deal with images with unseen degradation, unpaired face restoration methods are proposed. Bulat et al. [2] propose learning the degradation before face restoration from unpaired data. They design a high-to-low GAN to learn the real degradation processes from unpaired LQ and HQ images and a low-to-high GAN for face restoration. Lu et al. [14] disentangle the content and degradation features from low-quality images by using a KL divergence loss. These methods ignore using facial priors for face restoration tasks, which limits their restoration capability, especially when input images are degraded severely. In this work, we refer to the latent space editing methods and propose the cross-quality shift in the latent space of StyleGAN, which introduces the generative facial priors into the unpaired framework.

2.2. Latent Code Editing

The generator of an unconditional GAN learns the mapping $G : \mathcal{Z} \rightarrow \mathcal{X}$. When $z_1, z_2 \in \mathcal{Z}$ are close in the $\mathcal{Z}$ space, the corresponding images $x_1, x_2 \in \mathcal{X}$ are visually similar [24]. Latent code editing methods map an image x back to its latent representation z and then edit z as the vector arithmetic $z' = z + sn$, where s means the scale and n is the edit direction. Finally, the latent code z' is mapped to the image x', which has the desired style different from x. For instance, InterFaceGAN [18] employs some off-the-shelf classifiers to learn a hyperplane in the latent space serving as the separation boundary, which enables it to modify face attributes. Abdal et al. [1] learn a semantic mapping between the space $\mathcal{Z}$ and the space $\mathcal{W}$ of StyleGAN. In the face restoration task, PULSE [17] iteratively optimizes the latent code of StyleGAN and gets an HQ output in a self-supervised way. However, the optimization function of PULSE may be improper to images with other degradations than Bicubic. Unlike previous methods, the proposed method works in an unpaired framework, which makes it able to adapt to unseen degradation.

3. Methodology

Let $\mathcal{D}_{\text{pair}} = \{(I^i_{LQ}, I^i_{HQ})\}^L_{i=1}$ be a training dataset for paired face restoration, where each pair of the low-quality facial image I^i_{LQ} and the high-quality facial image I^i_{HQ} is well associated. The task of traditional paired face restoration is to train a conditional generating function $R_{\text{pair}} : \mathbb{R}^{m\times n} \rightarrow \mathbb{R}^{M\times N}$ based on the paired dataset $\mathcal{D}_{\text{pair}}$, so that the restored image

$$I_R = R_{\text{pair}}\left(I_{LQ}\right) \in \mathbb{R}^{M\times N} \tag{1}$$

can well approximate the ground-truth high-quality image I_{HQ}. However, a model trained by the paired dataset $\mathcal{D}_{\text{pair}}$ can usually only cope with low-quality images of the type in the dataset. Unpaired face restoration is proposed to address this issue. Let $\mathcal{D}_{\text{unpair}} = \{(I^i_{LQ}, I^{\pi_i}_{HQ})\}^L_{i=1}$ be a training dataset for unpaired face restoration, where the permutation π indicates that each pair of the low-quality facial image I^i_{LQ} and the high-quality facial image $I^{\pi_i}_{HQ}$ does not have any content correspondence. The goal of unpaired face restoration is to train a conditional generating functions $R_{\text{unpair}} : \mathbb{R}^{m\times n} \rightarrow \mathbb{R}^{M\times N}$ based on the unpaired dataset $\mathcal{D}_{\text{unpair}}$, so that the restored image

$$I_R = R_{\text{unpair}}\left(I_{LQ}\right) \in \mathbb{R}^{M\times N} \tag{2}$$

has the same facial content with I_{LQ} and has almost the same quality level as any high-quality image $I^{\pi_i}_{HQ}$. Unpaired face restoration is much more challenging than paired face restoration because there is no ground-truth high-quality image for supervision. How to utilize facial priors is thus important in unpaired face restoration tasks.

3.1. Learnable Cross-Quality Shift

We now present the proposed learnable cross-quality shift, which introduces the generative facial priors into the unpaired framework and models the degradation between the low and high qualities in the latent space of StyleGAN. Hypothetically, we consider that the essence of facial restoration is to find a transformation from the subspace of the low quality to the subspace of the high quality. In this work, we assume that with the powerful representation learning ability of deep neural networks, the intuition is similar to the one in Word2vec:

$$MAN - WOMAN = KING - QUEEN. \tag{3}$$

In other words, we could make an analogy that MAN and KING are from the LQ subspace; and WOMAN and QUEEN are from HQ space. Then, the degradation process from HQ quality level to a specific LQ quality level can be considered similar and close. Following this spirit, given the unpaired latent code w_h in the HQ representation domain $\mathcal{W}^H$ and w_l in the LQ representation domain $\mathcal{W}^L$, they can be shifted to the opposite domain via almost the same shift Δ:

$$\Delta = w_{h_l} - w_l \approx w_h - w_{l_h}, \tag{4}$$

where w_{h_l} is the HQ version of w_l and w_{l_h} is the LQ version of w_h. **We call Δ in (4) the learnable cross-quality shift.**

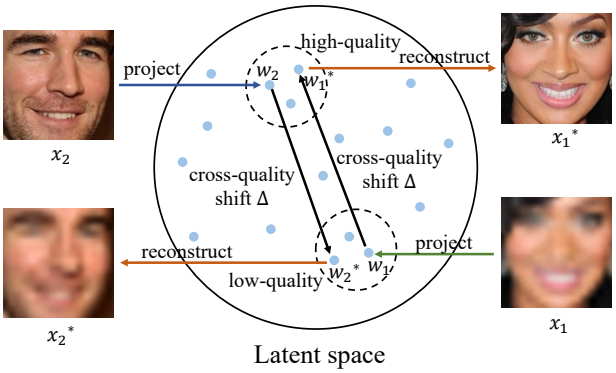

Figure 3. In the latent space, representations of the low quality form a subspace and representations of the high quality form another. With the learnable cross-quality shift Δ, low-quality representations can be transformed to the high-quality subspace and vice versa, which achieves restoration and degradation.

Fig. 3 shows how degradation and restoration work with the proposed learnable cross-quality shift Δ.

During the training phase, such a shift is trainable and we could get it via the following formulation:

$$w_{h_l} = w_l + \Delta \in \mathcal{W}^H, w_{l_h} = w_h - \Delta \in \mathcal{W}^L. \quad (5)$$

With two discriminators to enforce w_{h_l} and w_{l_h} to stay in HQ and LQ representation domains respectively, Δ can be learned in an unpaired framework.

During the inference phase, the cross-quality shift is applied to the representation of input image l:

$$w_{h_l} = w_l + \Delta. \quad (6)$$

Note that the value of Δ changes for different degradations. With the unpaired framework, Δ can be adjusted for an unseen degradation.

The proposed learnable cross-quality shift Δ in this work is similar to the edit direction in latent code editing methods. Given a Δ, which can be learned by the proposed framework, the editing direction n is available for control over restoration level in the latent space of StyleGAN:

$$n = \frac{\Delta}{\|\Delta\|_2}. \quad (7)$$

Then the latent code w can be edited as vector arithmetic with the editing direction n:

$$w' = w + \alpha n, \quad (8)$$

where α means the scale. The edited code w' is mapped back to the image space with the StyleGAN for the final result. In this way, the proposed learnable cross-quality shift Δ allows a user to tweak the shifting scale to adjust the restoration level of the restored image.

3.2. Two-branch Framework

Based on the proposed learnable cross-quality shift, a two-branch framework is designed to deal with unpaired data and add more constraints for fidelity.

A StyleGAN pre-trained on FFHQ [8] with 256×256 outputs is used for all our experiments. Different from other GANs, StyleGAN has two latent spaces: $\mathcal{Z}$ and $\mathcal{W}$. We choose to map the face representation into $\mathcal{W}$, as it is a more disentangled latent space than $\mathcal{Z}$ [18], thus more suitable for the learnable cross-quality shift. The latent codes $w \in \mathcal{W}$ have 14 channels, corresponding to the output size 256×256 of StyleGAN, each channel is a 512-dimensional vector. For the encoders E_{hq}, E_{lq} and E_{deg}, we use a ResNet-34 backbone and modify its output dimensionality to 14×512, which means the dimensionality of the learnable cross-quality shift Δ is also 14×512.

Given unpaired training sample h in the HQ image domain and l in the LQ image domain, the E_{hq} and E_{lq} extract latent codes from h and l separately:

$$w_h = E_{hq}(h), w_l = E_{lq}(l), \quad (9)$$

where w_h (w_l) is the representation of h (l) in latent space $\mathcal{W}$ of StyleGAN. The learnable cross-quality shift Δ^1 is extracted from l by the E_{deg}:

$$\Delta^1 = E_{deg}(l). \quad (10)$$

w_h (w_l) is converted into its LQ (HQ) version with the cross-quality shift Δ^1:

$$\hat{l}_h = G(w_h - \Delta^1), \hat{h}_l = G(w_l + \Delta^1), \quad (11)$$

where G is a pre-trained StyleGAN. $\hat{l}_h$ ($\hat{h}_l$) is the generated LQ (HQ) image corresponding to h (l). Then, Δ^2 is extracted from $\hat{l}_h$ by the E_{deg}:

$$\Delta^2 = E_{deg}(\hat{l}_h). \quad (12)$$

$\hat{l}_h$ and $\hat{h}_l$ are fed into E_{lq} and E_{hq} again to generate cycle results $\hat{h}$ and $\hat{l}$:

$$\hat{h} = G(E_{lq}(\hat{l}_h) + \Delta^2), \hat{l} = G(E_{hq}(\hat{h}_l) - \Delta^2). \quad (13)$$

Cycle results are used for adding constraints to ensure the fidelity. To keep w_h (w_l) as the representation of h (l), we map it back to the image h_{rec} (l_{rec}) through G, which makes adding constraint on them possible:

$$h_{\text{rec}} = G(w_h), l_{\text{rec}} = G(w_l). \quad (14)$$

Compared with previous methods [14], the learnable cross-quality shift Δ^i is estimated in the latent space of StyleGAN, which means that the degradation estimation takes advantage of the powerful representation learning and decoupling capabilities of the latent space $\mathcal{W}$. The learnable cross-quality shift Δ^i also enables a user to tweak the shifting scale to adjust the restoration level (See Fig. 6 in the section of Experiment Results).

3.3. Objective Function

Note that there is no ground-truth image as supervision, model objectives need to be designed carefully. The overall loss function includes the downscaled identity loss and the cycle-consistency loss for fidelity, as well as the adversarial loss and the perceptual loss for high quality.

Downscaled identity loss: Intuitively, when we downscale the input LQ image and the restored HQ images to the same resolution level, two downscaled images should be similar. To better preserve the facial identity, we use the identity constraints on the downscaled images. Specifically, a dedicated identity loss measuring the cosine similarity between the downscaled input LQ image and downscaled output restoration image. For the input LQ image l and the generated HQ image $\hat{h}_l$ corresponding to it, we introduce the downscaled identity loss:

$$\mathcal{L}_{\text{down}}(l, \hat{h}_l) = 1 - \langle R(l\downarrow), R(\hat{h}_l\downarrow)\rangle, \tag{15}$$

where R is a pre-trained ArcFace [5] network for face recognition, $\downarrow$ is the downsample function.

Cycle-consistency loss: The cycle-consistency loss is employed on both domains for preserving the content of input images:

$$\mathcal{L}_{cc} = \|h - \hat{h}\|_1 + \|l - \hat{l}\|_1. \tag{16}$$

Adversarial losses: To make the generated images look more realistic, the adversarial loss is applied on both domains:

$$\mathcal{L}_{\text{adv}} = -\log(D_{lq}(G(w_h - \Delta^1))) - \log(D_{hq}(G(w_l + \Delta^1))). \tag{17}$$

Perceptual losses: A perceptual loss is applied between the reconstructed HQ image $\hat{h}$ and the corresponding original HQ image h:

$$\mathcal{L}_{\text{perc}} = \|\phi(h) - \phi(\hat{h})\|_1, \tag{18}$$

where $\phi(\cdot)$ is the features of the pre-trained CNN. In our experiments, $conv_{5,4}$ layer of pre-trained VGG19 network [22] is employed to extract features from images.

Overall objective: The overall objective function for our network is:

$$\mathcal{L} = \lambda_{\text{cc}}\mathcal{L}_{\text{cc}} + \lambda_{\text{adv}}\mathcal{L}_{\text{adv}} + \lambda_{\text{perc}}\mathcal{L}_{\text{perc}} + \lambda_{\text{down}}\mathcal{L}_{\text{down}}, \tag{19}$$

where λ_{cc}, λ_{adv}, λ_{perc} and λ_{down} are the weights.

4. Experimental Results

4.1. Experimental Setup

Dataset: We use FFHQ dataset [8] as HQ dataset, which consists of 70,000 high-quality images at a size of 1024×1024. We use the first 7500 images for training and the last 1000 images for testing. We use CelebAHQ dataset [7] to generate the corresponding LQ images as LQ dataset, which has no identity intersection with FFHQ. We use images with numbers from 7500 to 15000 for training and the last 1000 images for testing. Before training and testing, images in CelebAHQ are cropped and aligned in the same manner as FFHQ. This step results in 7211 images left for training and 967 images left for testing.

Degradation models: To fully demonstrate the generalization of our proposed method on different degraded images, three degradation models are used for simulating LQ images. They can be formulated as follows [3, 11, 12, 23]:

$$I_{LQ} = [(I_{HQ} \circledast \boldsymbol{k}_\sigma)\downarrow_r + \boldsymbol{n}_\delta]_{\text{JPEG}_q}, \tag{20}$$

where $\circledast$ represents the convolution operation between the HQ image I_{HQ} and a blur kernel k_σ. $\downarrow_r$ is the downsampling operation with a scale factor r. n denotes the additive white Gaussian noise (AWGN) with a noise level δ. $(\cdot)_{\text{JPEG}_q}$ indicates the JPEG compression operation with quality factor q. The degradation models are designed to simulate three different degradation levels: mild, moderate, and severe. Images are first resized to 256×256 in all three degradation models. The first one is to downsample with scaling factor 8 by Bicubic, and then compress by JPEG with quality factor $q \in \{90 : 95\}$ (denote as BicC for short), which is to simulate mild level degradation. The second one is to downsample with scaling factor 8 by Bicubic, and then add Gaussian noise with covariance $\delta \in \{20 : 25\}$ (denote as BicN for short) for simulating moderate level degradation. The third one is to Gaussian blur with kernel standard deviation $\sigma \in \{5 : 7\}$, downsample with scaling factor 8 by Bilinear, and then add Gaussian noise with covariance $\delta \in \{10 : 15\}$ (denote as BBilN for short), which is to simulate severe level degradation. Note that this work can adapt to unseen degradations beyond these three degradation models since it is an unpaired method.

Training and testing details: We use Adam optimizer [10] to train our networks. We choose $\beta_1 = 0.5$, $\beta_2 = 0.999$, and set the learning rate of the generator and discriminator to 0.0001 and 0.0004 respectively. For hyper-parameters, we experimentally set: $\lambda_{\text{cc}} = 1$, $\lambda_{\text{adv}} = 0.1$, $\lambda_{\text{perc}} = 0.05$, and $\lambda_{\text{down}} = 0.1$. The training batch size is set to 4. All models were implemented by PyTorch and trained on a Tesla V100 GPU. Since projecting an image into latent space is a hard task, we pre-train the E_{lq} and E_{hq} with $\mathcal{L}_{rep}$:

$$\mathcal{L}_{rep} = \|l - G(E_{lq}(l))\|_1 + \|h - G(E_{hq}(h))\|_1. \tag{21}$$

We fix parameters of E_{lq}, E_{hq} to train the E_{deg} in the first 10k iterations, then we train the whole framework except for the pre-trained generator G in the next 50k iterations. When testing, E_{hq} is not required. Given a test LQ image l, the LQ encoder E_{lq} extracts its representation w_l and the

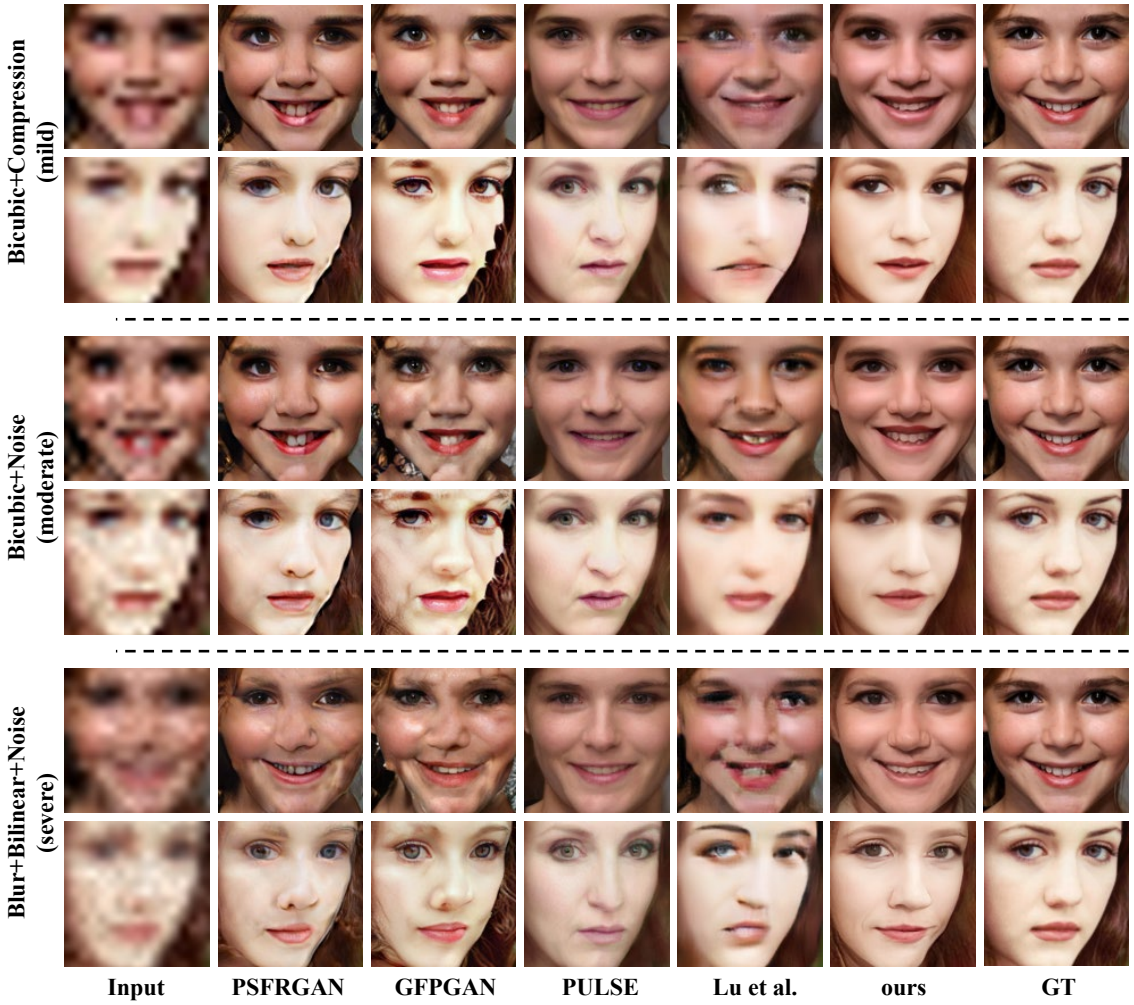

Figure 4. Visual evaluation under three different kinds and levels degradation, including bicubic plus compression, bicubic plus noise, and blur plus bilinear and noise. Our method produces the best visual results across these degradations.

degradation encoder E_{deg} estimates its degradation, which gets the cross-quaity shift. Then the learned shift is applied to w_l and the pre-trained generator G will generate the restored image h_l with (11).

4.2. Comparison with SOTA Methods

We compare our method with some state-of-the-art face restoration methods. For paired methods, we include PSFRGAN [3] and GFPGAN [23]. The latter also includes a pre-trained StyleGAN in its framework. For unpaired methods, Bulat et al. [2] and Lu et al. [14] are included. For other unsupervised methods, ZSSR [21] and PULSE [17] are included. We adopt official codes except for Bulat et al. and Lu et al., for which we use a re-implementation. We use the official pre-trained model of paired methods for two reasons: 1) They are designed for blind face restoration; 2) They cannot be trained on unpaired data with unseen degradation. We fine-tune the unpaired methods on our face training set for fair comparisons.

Only the face region is cropped for the evaluation. We employ pixel-wise metrics (SSIM) and the perceptual metric (LPIPS [27]) for restored images with the Ground-Truth (GT). Similar to GFPGAN, the identity similarity in the ArcFace [5] feature embedding is also measured. We calculate the inner product (IP) of the output vectors of the restored image and its GT, where bigger values indicate closer identity to the GT. Since PSNR and SSIM are known not to correlate very well with perceptual quality [27], we just list SSIM scores in the paper.

Tab. 1 summarizes quantitative results on three types of degraded data. 1) Our method achieves the lowest LPIPS on moderate and severe degraded data, indicating that our results are perceptually close to the GT. 2) Besides the perceptual performance, our method also retains a better iden-

Methods		Bicubic+Compression (mild)			Bicubic+Noise (moderate)			Blur+Bilinear+Noise (severe)		
		SSIM↑	LPIPS↓	IP.↑	SSIM↑	LPIPS↓	IP.↑	SSIM↑	LPIPS↓	IP.↑
Paired	PSFRGAN	0.680	**0.127**	0.703	0.598	0.153	0.655	0.581	0.152	0.646
	GFPGAN	0.671	0.137	**0.712**	0.474	0.251	0.616	0.523	0.184	0.644
Unsupervised	ZSSR	0.679	0.505	0.573	0.533	0.611	0.511	0.564	0.623	0.484
	PULSE	0.625	0.166	0.660	0.618	0.166	0.658	0.598	0.180	0.635
Unpaired	Bulat et al.	0.583	0.363	0.549	0.481	0.342	0.504	0.563	0.372	0.492
	Lu et al.	0.623	0.181	0.616	**0.676**	0.228	0.558	0.531	0.251	0.541
	Ours	**0.684**	0.139	0.711	0.643	**0.150**	**0.668**	**0.614**	**0.147**	**0.679**

Table 1. Quantitative comparison on three types of degraded data. The proposed metric, IP., represents the inner product of the feature vectors of the restored image and its GT, higher is better. On mild degraded data, the proposed method produces slightly worse results than paired methods. And on moderate and severe degraded data, the proposed method produces the best results in LPIPS.

Methods	SSIM↑	LPIPS↓	IP.↑
baseline	0.633	0.181	0.626
+ upper	0.623	0.177	0.641
+ upper + Δ^i	0.614	0.163	0.660
+ upper + Δ^i + $\mathcal{L}_{\text{down}}$	**0.643**	**0.150**	**0.668**

Table 2. Ablation study. For SSIM, higher is better; for LPIPS, lower is better; for IP., higher is better. Both adding upper branch and adding learnable cross-quality shift Δ^i benefit the restoration results in LPIPS and IP.. Adding downscaled identity loss $\mathcal{L}_{\text{down}}$ further improves the performance.

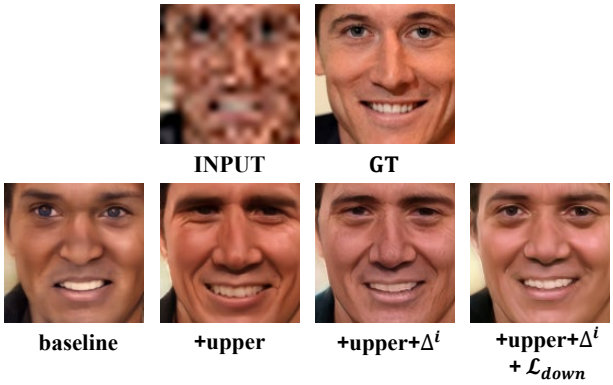

Figure 5. Ablation study. Adding upper branch, adding the cross-quality shift Δ and adding downscaled identity loss $\mathcal{L}_{down}$ can benefit the qualitative performance. With the combination of them, our proposed method achieves the best performance.

tity on two types of degraded data, indicated by the highest inner product of the feature vectors. 3) On mild degraded data, our method produces slightly worse results than paired methods. This is because the mild degraded data is similar to the training data of the paired methods. The GT images in the training data of the paired methods promote their good performance. However, when paired methods are faced with data that differs greatly from the training data (moderate and severe degraded data), their results are unsatisfactory.

Fig. 4 shows the qualitative comparisons. 1) Thanks to the powerful generative facial prior provided by the pre-trained StyleGAN, our method recovers high perceptual quality details in the eyes, teeth, etc. 2) Paired methods (PSFRGAN and GFPGAN) are unable to handle degradations beyond their training setting, as shown in the second and third columns of Fig. 4. 3) Although PULSE can also generate high perceptual quality results with the facial prior provided by the pre-trained StyleGAN, it can not retain the face identity, as shown in the fourth column of Fig. 4. 4) Previous unpaired methods (Lu et al.) fail to generate high perceptual quality results, which may be due to their lack of help from facial priors, as shown in the fifth column of Fig. 4. As shown in the sixth column of Fig. 4, our method can produce results with realness and fidelity on different types of degraded images.

4.3. Ablation Study

We perform an ablation study to analyze the effectiveness of each component or loss in the proposed framework. Both quantitative and qualitative results on BicN are reported for the four variants of our framework: 1) Baseline: only using the lower branch of the proposed architecture without the upper branch or the learnable cross-quality shift Δ^i or downscaled identity loss $\mathcal{L}_{\text{down}}$; 2) adding the upper branch; 3) adding the learnable cross-quality shift Δ^i; 4) adding downscaled identity loss $\mathcal{L}_{\text{down}}$.

We present the SSIM, LPIPS, and IP. for each variant in Tab. 2. Tab. 2 demonstrates that adding the upper branch, the learnable cross-quality shift Δ, and the $\mathcal{L}_{\text{down}}$ can benefit the restoration results. With the combination of them, our method can achieve the best performance. As shown in Fig. 5, we can observe that adding the upper branch makes it possible to add the cycle-consistency loss, which helps with the fidelity of the results. Adding the learnable cross-quality shift Δ enables w_l and w_h to stay in the LQ domain and the HQ domain respectively, which also benefits

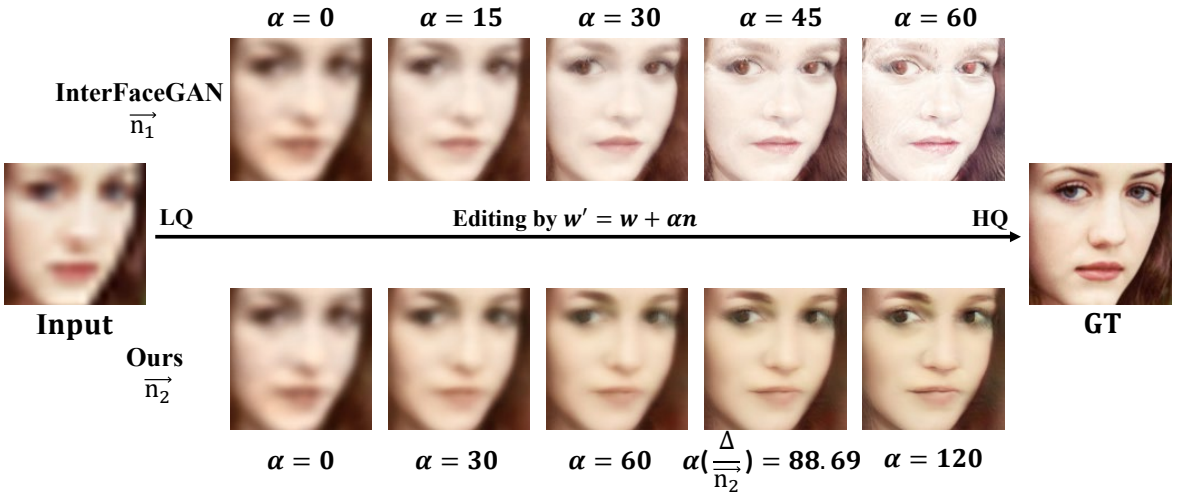

Figure 6. Visual evaluation under restoration level editing. Our method finds a better editing direction for adjusting the restoration level in the latent space. When manually set scale α changes, the results would be more blurry or sharper.

the performance. Adding $\mathcal{L}_{\text{down}}$ promotes the preservation of the facial identity, which further improves the fidelity of the results.

4.4. Control over Restoration Level

Here we compare with a popular unsupervised latent code editing method and display the ability to control the restoration level of the results.

InterFaceGAN [18] is a recent popular latent code editing method. Different from face restoration methods, InterFaceGAN learns the editing direction n in the latent space, which means that it requires a manually set scale α to decide the restoration level and generate the final result. InterFaceGAN predicts the same restoration direction for all images, while our method predicts a different restoration direction for each image. Note that our method can automatically learn the cross-quality shift Δ (product of α and n in EQ. (8)). The editing directions of the two methods are used for a fair comparison. For InterFaceGAN, we manually set five scale α. For our method, we manually set four scale α and use a α that is automatically learned by the learnable cross-quality shift Δ. Note that InterFaceGAN is not a face restoration method, so it is reasonable to change other attributes while adjusting the restoration level.

As shown in Fig. 6, with the increase of α, the results become sharper. Both methods enable a user to adjust the restoration level of the restored image, which alleviates the problem of blurry or over-sharp results. Our method finds a better editing direction for adjusting the restoration level in the latent space. InterFaceGAN turns images into a painting style during the adjustment of restoration level, which may be because improving image quality and changing image style are entangled. Our method is designed for image restoration, so the downscaled identity loss is beneficial for finding a restoration direction.

5. Conclusion

In this paper, we propose an unpaired method based on the learnable cross-quality shift. The proposed learnable cross-quality shift not only introduces the generative facial priors into the unpaired framework, but also explicitly models the degradation across two quality levels. The cross-quality shift also allows a user to tweak the shifting scale to adjust the restoration level of the restored image. Based on the learnable cross-quality shift, a two-branch framework is designed for unpaired data, which enables our method to be fine-tuned on images with unseen degradation. Experimental results show improved performance on moderate and severe degradation images compared with state-of-the-art methods.

Acknowledgements

This work was supported in part by Chinese National Key R&D Program (2019YFB1804304), State Key Laboratory of UHD Video and Audio Production and Presentation, BirenTech Research, Shanghai Key Laboratory of Digital Media Processing and Transmissions (STCSM 18DZ2270700) and 111 plan (BP0719010).

References

[1] Rameen Abdal, Peihao Zhu, Niloy J Mitra, and Peter Wonka. Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows. *ACM Transactions on Graphics (TOG)*, 40(3):1–21, 2021. 3

[2] Adrian Bulat, Jing Yang, and Georgios Tzimiropoulos. To learn image super-resolution, use a gan to learn how to do image degradation first. In *Proceedings of the European conference on computer vision (ECCV)*, pages 185–200, 2018. 1, 3, 6

[3] 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, 5, 6

[4] 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

[5] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019. 5, 6

[6] Muhammad Ali Farooq, Ammar Ali Khan, Ansar Ahmad, and Rana Hammad Raza. Effectiveness of state-of-the-art super resolution algorithms in surveillance environment. In *Conference on Multimedia, Interaction, Design and Innovation*, pages 79–88. Springer, 2020. 1

[7] 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. 5

[8] 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. 4, 5

[9] 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

[10] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[11] Xiaoming Li, Chaofeng Chen, Shangchen Zhou, Xianhui Lin, Wangmeng Zuo, and Lei Zhang. Blind face restoration via deep multi-scale component dictionaries. In *European Conference on Computer Vision*, pages 399–415. Springer, 2020. 1, 3, 5

[12] Xiaoming Li, Ming Liu, Yuting Ye, Wangmeng Zuo, Liang Lin, and Ruigang Yang. Learning warped guidance for blind face restoration. In *Proceedings of the European conference on computer vision (ECCV)*, pages 272–289, 2018. 5

[13] Yihao Liu, Anran Liu, Jinjin Gu, Zhipeng Zhang, Wenhao Wu, Yu Qiao, and Chao Dong. Discovering" semantics" in super-resolution networks. *arXiv preprint arXiv:2108.00406*, 2021. 1

[14] Boyu Lu, Jun-Cheng Chen, and Rama Chellappa. Unsupervised domain-specific deblurring via disentangled representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10225–10234, 2019. 1, 3, 4, 6

[15] Cheng Ma, Zhenyu Jiang, Yongming Rao, Jiwen Lu, and Jie Zhou. Deep face super-resolution with iterative collaboration between attentive recovery and landmark estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5569–5578, 2020. 1

[16] Fabio Valerio Massoli, Giuseppe Amato, and Fabrizio Falchi. Cross-resolution learning for face recognition. *Image and Vision Computing*, 99:103927, 2020. 1

[17] 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. 3, 6

[18] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9243–9252, 2020. 3, 4, 8

[19] 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. 2

[20] Ziyi Shen, Wei-Sheng Lai, Tingfa Xu, Jan Kautz, and Ming-Hsuan Yang. Exploiting semantics for face image deblurring. *International Journal of Computer Vision*, 128(7):1829–1846, 2020. 1

[21] Assaf Shocher, Nadav Cohen, and Michal Irani. "zero-shot" super-resolution using deep internal learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3118–3126, 2018. 6

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

[23] 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, 3, 5, 6

[24] Weihao Xia, Yulun Zhang, Yujiu Yang, Jing-Hao Xue, Bolei Zhou, and Ming-Hsuan Yang. Gan inversion: A survey. *arXiv preprint arXiv:2101.05278*, 2021. 3

[25] 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

[26] 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. 2

[27] 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. 6