Copy or Not? Reference-Based Face Image Restoration with Fine Details

Min Jin Chong[1], Dejia Xu[2], Yi Zhang[3], Zhangyang Wang[2], David Forsyth[1],
Gurunandan Krishnan[4], Yicheng Wu[4], Jian Wang[4] ✉
[1]UIUC, [2]UT Austin, [3]CUHK, [4]Snap Inc.

Abstract

*Reference-guided face restoration can have better identity preservation than non-reference-based methods. However, existing methods can (a) easily produce artifacts, possibly attributable to inefficient facial priors and (b) do not well preserve fine-grained facial details crucial for identity, such as freckles, tattoos, and scars. In this work, we propose solutions for these problems. (1) We incorporate a stronger facial prior, generative facial prior (GFP), for reference-based face image restoration. (2) We identify an ambiguity and point out that traditional loss prevents the network from heavily copying facial features from the reference. To address this, we set a new goal and come up with a new loss to realize the new goal. More specifically, when the ground truth and reference are different (*e.g.*, differences in wrinkles, makeup, facial hair,* etc.*), which one should the output look like? As a simple example, ground truth does not have a mole while reference has one. Traditional loss chose the ground truth, which seems natural, but then the network also learns to ignore reference's facial features; during testing, the network often hesitates. Our new goal is to copy features from the reference as much as possible while maintaining semantic consistency with the degraded input. We propose to use spatial minimum loss and cycle consistency loss to realize the new goal and make the network copy features without hesitation. Using only a single reference image, our proposed method is able to restore highly degraded images while accurately capturing fine-grained facial details. To our knowledge, we are the first face restoration framework that is able to restore faces at this granularity. Code and data are available at* https://github.com/RefineFIR/RefineFIR.

1. Introduction

State-of-the-art face restoration methods [36,40,47] typically involve using pre-trained face GANs as a prior to produce sharp and realistic-looking faces. However, the main focus of these recent works is on unconditional face restoration. This works well for instances where the degradation is mild, and most of the facial features are visible in the input. In cases with severe degradations, while these methods can recover a high-quality face, the identity of the original face will likely be lost. At the same time, fine-grained facial details such as freckles, eye colors, tattoos, scars, *etc*. will also be lost in the restoration. It is important to get these facial details right. In many cases, they are unique to an

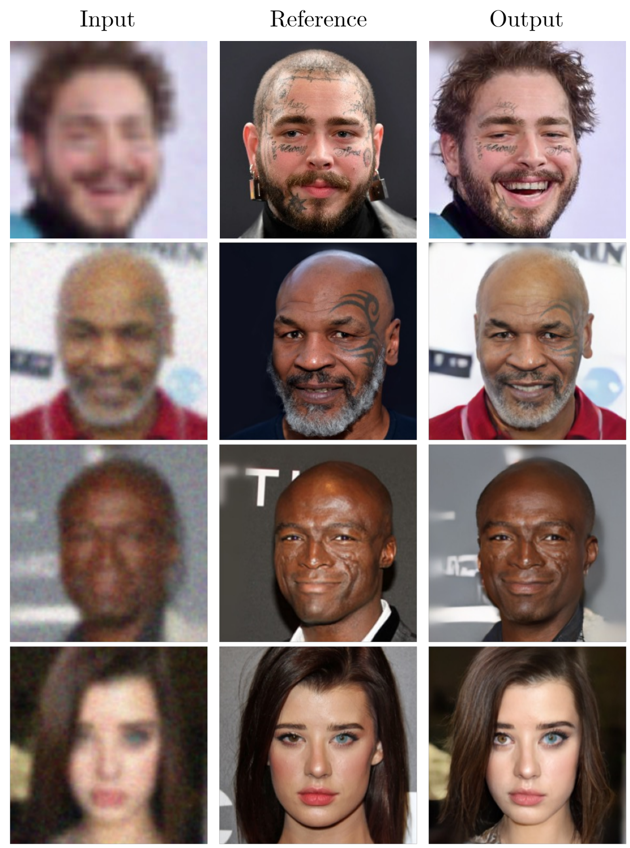

Figure 1. Given a reference image, our method ReFine can restore a severely degraded image while preserving identity and fine-grained details. Notice the highly detailed tattoos in rows 1 and 2, the scar under different lightings in row 3, and the different eye colors in row 4. "ReFine" is an acronym derived from "Reference-based Face Restoration with Focus on Fine Details."

individual and are distinctively tied to their identities.

There have been several works that propose to alleviate these issues by using a high-quality reference image in addition to the degraded image [10, 25]. The identity and facial features are borrowed from the reference image in order to guide the restoration. However, as our experiments in Figure 6 show, their image quality is subpar compared to SOTA blind face restoration works, and they are also unable to capture fine-grained facial details well.

In this work, we revisit reference image-guided face image restoration by incorporating the diverse and rich generative facial priors (GFP), which outperform existing reference-based methods in overall quality. More importantly, we deeply analyze the task of reference-based face image restoration, challenge the loss design of all traditional methods, and propose new loss functions.

All traditional reference-based face image restoration methods use the loss measuring the difference between the output and the ground truth (the clean version of the input). At first glance, this seems reasonable and natural, but it actually hinders the algorithm from copying features from the reference image. Why is that? In daily life, there are often cases where the ground truth and the reference image look different (like makeup (influencing moles, freckles, eyebrows, eyelashes, tattoos), wrinkles, accessories, hair, and beard, *etc.*). **Let us take a simple example: the reference image has a mole, but the ground truth does not (perhaps it is covered by makeup). Should the output have a mole?** The answer is: if you cannot tell from the input whether there is a mole or not, then the output should have one. If you can clearly see from the input that there is definitely no mole, then the output should not have one. If the traditional loss is used, the network would know to ignore the mole in the reference when it sees such training data. However, during testing, when there is a mole in the reference, copy or not? Basically, you do not know which features in the reference should be copied and which should not, leading to hesitation in the network, which is reluctant to copy features, resulting in unclear and faint or even loss of fine-grained facial details. We propose to use spatial minimum loss (either close to the ground truth or the reference) and cycle consistency loss (maintaining similarity to the degraded input). If our proposed loss is used, the desired effect can be achieved. The network is clear when to copy features from the reference. As can be seen from Figure 1, our method can heavily copy features from the reference while maintaining consistency with the input; it restores severely degraded faces at unprecedented granularity.

We summarize our contributions in the following:

1. We identify an ambiguity in reference-guided image restoration. We set a new goal to borrow high-quality details from the reference as much as possible while maintaining coherence with the degraded input.
2. The new goal is realized by a combination of losses (spatial minimum loss, cycle consistency loss, and other assistant losses) which are to ensure our face restoration borrows features heavily from the reference but remains semantically consistent with the input.
3. We incorporate generative facial priors for the reference-guided face image restoration task. The pipeline improves the overall quality and enlarges the task's scope so that the input could be more degraded, and the reference image could be very different from the ground truth.
4. Extensive experiments show that our method demonstrates a significant performance superiority over both reference-based and non-reference-based in terms of preserving identity and fine facial features. Our method is the first to restore faces at this granularity.

2. Related work

2.1. Blind face restoration

Restoring a degraded face image is a challenging task because of the unknown degradation process and severe information loss. A face prior is thus usually exploited for this task. Based on how the face prior is used, previous works can be divided into three categories. 1) Geometric prior – facial landmarks [1, 6, 21], face parsing maps [4, 34, 39], facial component heatmaps [42], or 3D shapes [14] are included into the network design. Such priors however cannot provide fine-grained facial details for high-quality image restoration. 2) Dictionary prior – a dictionary is learned from face images, where each word in the dictionary contains rich face details in the feature domain. The LQ image is then reconstructed by these words. These methods [12, 23, 37, 47, 48] can recover better details than methods in 1). 3) Generative facial prior (GFP) – in recent years, pre-trained StyleGANs [17–19] have shown powerful face generation capability and are employed in image restoration. These include early exploration by GAN inversion [11, 29, 33] and recent success by fusing input's structural information and the face generator for superior fidelity [2, 28, 36, 40]. Diffusion model-based methods like DiffFace [43], DiffBIR [27], and DR2 [38] can also achieve great performance (although still slightly worse than StyleGAN methods in terms of identity preservation and speed). However, when the input face image is severely degraded, or the face has unique details (like freckles, wrinkles, and tattoos), the restored images by all these methods, although still a face, do not match the original identity and do not include individualistic details because the algorithms have no way to get such information.

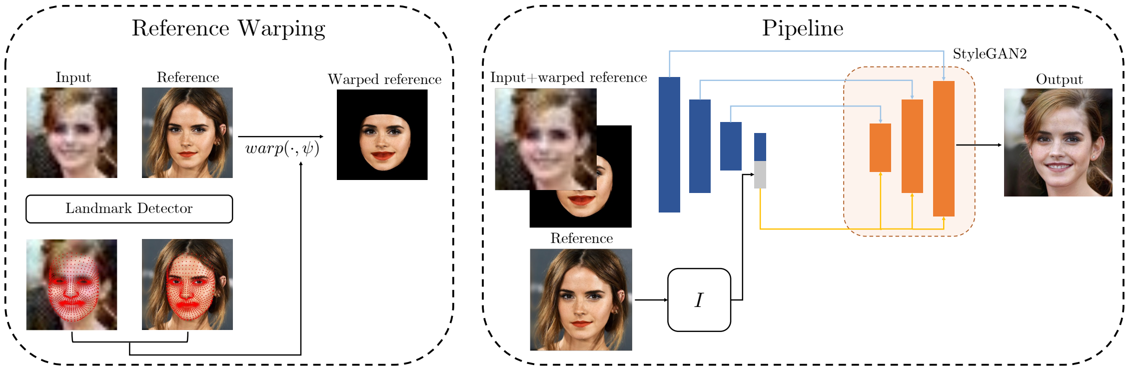

Figure 2. Overview of our ReFine network. We first warp the reference face image towards the input face image based on the facial landmarks detected by fine-tuned landmark detection algorithm. Then the input and the warped reference images are sent to an encoder. The facial structural information extracted in different levels and the identity feature vector extracted from the reference image are used to modulate the pre-trained face generator network StyleGAN2 to generate the output. During training, we employ spatial minimum loss, cycle consistency loss, adversarial loss, and identity loss.

2.2. Reference-guided face restoration

A reference image of the same identity can provide fine-grained facial details to guide the restoration. Depending on the number of reference images being used, these methods have two categories.

Single reference Single reference-guided face restoration has been explored in the literature, with GFRNet [25] and GWAINet [10] being two contemporary methods that are most relevant to our work. GFRNet learns two UNet-like subnetworks, one for warping the reference image, the other for restoring image restoration. GWAINet is composed of a localization network for warping and a generator network for fusing the features from the two inputs and generating the image. GWAINet does not require facial landmarks during training. It is worth mentioning that GFRNet and GWAINet often perform worse than SOTA non-reference-based methods. In contrast, our method ReFine warps the reference using keypoints from a finetuned facial landmark detector. More importantly, ReFine takes advantage of the powerful generation capability of a pretrained StyleGAN to ensure a realistic output.

Multiple references Instead of relying on a single reference image, ASFFNet [24] selects the optimal guidance based on landmark locations, while Wang *et al.* [35] used pixel-wise weights on the multiple exemplars for face image denoising and super-resolution. More recent methods use advanced dictionary learning or StyleGAN, *e.g.*, DMDNet [26] utilized a dictionary which is constructed by tens of reference images, and MyStyle [32] fine-tuned a pretrained StyleGAN face generator by ~ 100 same identity images to personalize the generator. Although multiple reference image-based methods achieve good performance, the need for multiple references restricts their use. Although ASFFNet and DMDNet can also accept a single reference, the performance would drop considerably. In contrast, we design an algorithm that only needs *a single reference*, making it more accessible to users.

Common to all existing single (or several) reference(s)-guided face restoration approaches, they might generate obvious artifacts (see Figure 6, which we show can be fixed by using GFP) and generate unclear fine-grained facial details or even lose them. We go through in detail why this happens in Section 3.2.

3. Methodology

3.1. Overview

Given an input face image with an unknown degradation and a reference image, our goal is to restore the face image by copying the features and identity from the reference.

Network As shown in Figure 2, our network is an encoder-decoder architecture. The decoder is a pre-trained StyleGAN2 [19] to provide a generative facial prior. We first warp the reference y towards the degraded input x_d (whose ground truth is x) to get warped reference y_w, and then feed x_d, y_w, and identity embedding of the reference into the network $\mathcal{G}$, and we have

$$\hat{x} = \mathcal{G}(x_d, y_w, I(y)) \tag{1}$$

where $\hat{x}$ is the output and $I(\cdot)$ is a pretrained identity embedding network by ArcFace [8].

Our network architecture is similar to GPEN [40], with a convolutional encoder that encodes the input image and warped reference into intermediate feature stacks and a style code, which are then fed to the StyleGAN2 prior.

There are two architectural differences between ReFine and GPEN. Firstly, we have an additional identity embedding as an input. Let s be the style code from our encoder and $I(y)$ as our identity embedding of the reference image. We concatenate s and $I(y)$ vectors and pass it through four

MLP layers of dimension 512 with Leaky ReLU activations of slope 0.2 to obtain our final style code s_y. s_y is then fed into the Modulated Convolution blocks in StyleGAN2. The second difference compared to GPEN is how we fuse features from the encoder and features from the StyleGAN2 prior. We fuse the two features by simple interpolation with a predicted mask while GPEN concatenates them. More details can be found in the supp. Section 2.

The network is trained on synthetic data triplets (x, x_d, y) where $x_d = T(x, \theta)$ and T is a differentiable parametric degradation function. We employ a combination of losses which will be detailed in Section 3.3. The StyleGAN2 prior is finetuned during training.

Warping It is important for reference face y to be spatially aligned to x_d so that $\mathcal{G}$ can easily copy facial details. Unlike previous work that used complicated methods such as deep features warping [24] or learning entirely new warping subnets [10, 25], we simply find correspondences between x_d and y and warp them in the image space. We observed that by finetuning existing face landmark detectors (like [7,15,20]; we use a commercial implementation of [7]) on low-quality data with a fixed range of degradations, the facial landmarks on low-quality inputs can be detected with high accuracy (see Figure s2 in supp.). More details of the warping can be found in the supp. Section 3.

Working scope and limitation We target restoring face images with middle to severe degradation levels [5]. For mild to middle degradations where the input still contains facial details, existing methods like GFP-GAN [36], GPEN [40], and CodeFormer [48] can work reasonably well, and there is no need to use a reference. For very severe degradations, the finetuned landmark detection network might not work, and then our method will fail. With the use of GFP and well-engineered landmark detection, we pushed the boundary of the working scope of existing reference-based methods in terms of degradation levels.

3.2. Ambiguity in reference-guided face image restoration and the proposed goals

We identify a major ambiguity in reference-based face image restoration which causes a dilemma. When the facial details of the ground truth and the reference perfectly match, there is no doubt that the output's details should look like the ground truth. But when they have differences in facial details, what the output's details should look like (remember the input is highly degraded)? Still the ground truth? Somewhere in between (but how)? Or the reference?

All previous reference(s)-guided approaches GFRNet [25], GWAINet [10], ASFFNet [24], Wang *et al*. [35] and Li *et al*. [26] exploited different ways to fuse the input and the reference(s) and trained the network to minimize the distance between the output $\hat{x}$ and ground truth x and do not include reference y in their training objective. Thus, **their goal** is

1. Restore back the ground truth

In an ideal scenario where x perfectly matches y, by learning to reconstruct x from x_d, the network also learns to copy features directly from y. However, suppose y does not perfectly match x, the network has to ignore the unmatching parts of y while "hallucinating" details to reconstruct x. Details from y are then likely to be lost and it is unclear what parts of y the model are supposed to preserve.

Figure 3. We visualize images from the same person in the CelebA-HQ dataset. In many cases, the same person can look very different in different images. For example, we have images taken at different ages, under various lightings, with and without sunglasses. When trained on such datasets, the reference image we choose can look *very different* from the input image. This presents a dilemma for our optimization; see Section 3.2.

There are many cases where x does not match y because people can look *very different* under different lightings, makeups, postures, seasons, or even due to accessories such as glasses. This is very evident if we look at images from the same person in dataset such as CelebaA-HQ [16] (see Figure 3). There are various pictures taken a long time apart, under various different conditions, making it somewhat difficult to even match them to the same identity. It is virtually impossible to obtain a perfect dataset where all the images perfectly match. Suppose that this dataset does exist and we train under traditional objective (only making $\hat{x}$ look like x). During training, the model will only see reference images that very closely resemble x (in terms of facial details). It is then unlikely it will generalize and perform well during test time where users will likely upload a reference image that does not closely resemble x.

In contrast, we think, on a high level, reference-guided face restoration should have four goals, which are:

1. **Copy features from reference as much as possible.**
2. Generate details not available from reference.
3. **Output semantically consistent with degraded input.**
4. Face must be realistic.

Goal 1 is especially important because the main reason we use a reference is to borrow facial details that will be otherwise unavailable in blind restoration.

Our goals are more inclusive. It goes back to the traditional goal if there is no mismatch in facial details between

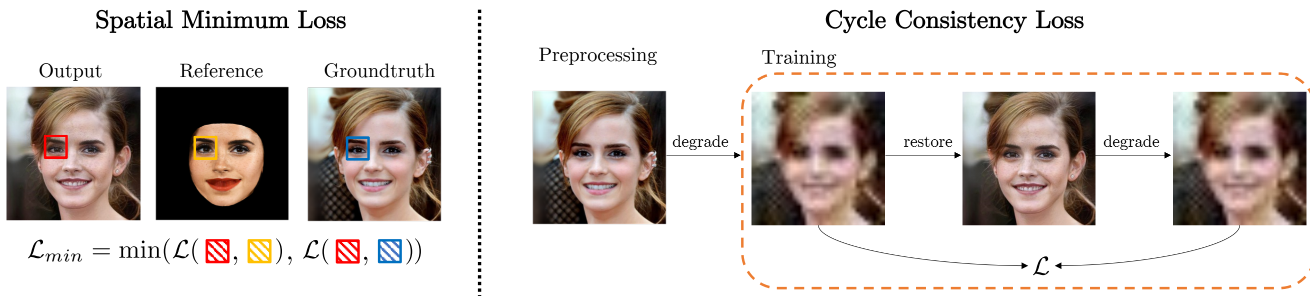

Figure 4. We use a combination of losses to realize the proposed inclusive goal for reference-guided face restoration. **Left:** Spatial minimum loss ensures that our output, at each spatial location, is close to either the warped reference image or the ground truth image. **Right:** In principle, assuming faithful restoration, the degradation and restoration steps should cancel out each other. Cycle consistency loss is thus the loss between the input and the degraded output. The cycle consistency constraint ensures that our output preserves the semantics of the input.

the ground truth and the reference. It can handle more general cases where there is a mismatch, like in the training dataset or during the test time. Looking from an extreme case, suppose x_d is very degraded with identity details missing and y looks very different from the original x. Following goal 1, the expected behavior for the model is to take identity details from y to restore x_d. The expected output $\hat{x}$ should thus look like y instead of x, while of course semantically matching x_d (goal 3). This cannot be realized by the traditional training objective of reconstructing x.

Based on the above discussion, we argue that training the model to make the output look like x only will likely cause issues with preserving fine-grained details from the reference image; in order to preserve sharp facial details and realize the more inclusive goal, we need a new objective design.

3.3. Objective functions

The common objective for face restoration is simply an L2 or perceptual loss between $\hat{x}$ and x [10, 24–26, 35]. As discussed in Section 3.2, using x as the only target might be inappropriate and not inclusive as x can look very different from reference y. We instead formulate a combination of losses (see Figure 4) in order to tackle all the goals of reference-guided face restoration as outlined in Section 3.2.

Spatial minimum loss Following goals 1 and 2, we want $\hat{x}$ to have details from both x and y. We thus introduce spatial minimum loss, which essentially says that at each spatial location, $\hat{x}$ should either be close to y_w or x. Specifically, let $\mathcal{L}_a = \mathcal{L}(\hat{x}, x)$ and $\mathcal{L}_b = \mathcal{L}(\hat{x}, y_w)$ where $\mathcal{L}(\cdot, \cdot)$ is an elementwise distance function (we use LPIPS [45]), our spatial minimum loss is thus

$$\mathcal{L}_{min} = \mathbb{E}[\min(\mathcal{L}_a, \mathcal{L}_b)] \tag{2}$$

Intuitively, pixels that are closer to x should be pushed closer to x, and vice versa. However, using $\mathcal{L}_{min}$ naively in training results in $\hat{x} = y_w$, where $\mathcal{G}$ simply copies the warped reference. This is because y_w is an input to $\mathcal{G}$ and it is easier to learn an identity function than to hallucinate new details to match x. As a result $\mathcal{L}_{min}$ will always be minimized w.r.t. y_w. $\hat{x}$ will thus look like y_w which obviously does not stay faithful to x_d.

Cycle consistency loss To prevent this from happening and ensure that the output semantics match the input's (goal 3), we formulate the cycle consistency loss inspired by CycleGAN [49]. When we restore a degraded image to a clean image, we expect that performing the same degradation on the cleaned image will return us to the degraded image. Specifically, recall $x_d = T(x, \theta)$ that we degraded image x with differentiable function with parameter θ. Using the cycle consistency argument, our cycle loss is

$$\mathcal{L}_{cycle} = \mathcal{L}(T(\hat{x}, \theta), x_d) \tag{3}$$

With $\mathcal{L}_{cycle}$, the restored face will be faithful to the degraded input, which now prevents the network from solely copying details from y_w.

Adversarial loss In addition to the two above losses, we also have an adversarial loss to ensure the face stays realistic (goal 4). We need an adversarial loss because $\mathcal{L}_{min}$ and $\mathcal{L}_{cycle}$ encourage $\mathcal{G}$ to produce an image close to x or y_w at each spatial location. This does not guarantee that $\hat{x}$ will be a coherent and natural-looking face. Adversarial loss penalizes when $\mathcal{G}$ produces unnatural blending artifacts.

Identity loss Lastly, there is no guarantee that y_w perfectly preserves the original identity of y after warping. Thus, by copying features from y_w, $\hat{x}$ still might not look like the original reference y. We thus include an identity loss between $\hat{x}$ and y. Specifically,

$$\mathcal{L}_{id} = 1 - \frac{I(\hat{x}) \cdot I(y)}{||I(\hat{x})|| ||I(y)||} \tag{4}$$

These 4 losses ensure that 1) the restored face preserves details from both the degraded face and reference face; 2) the overall looking of the output and the degraded face is the same; 3) the face looks natural.

Figure 5. Visual comparison results for face super-resolution ($\times 8$). Ours performs much better than all existing reference-based methods whose development timeline is shown in the upper left. Notice the five moles, the nose stud, and the tattoos; best viewed when zoomed in.

Table 1. Quantitative results for face image restoration tested on CASIA-WebFace dataset. Our PSNR/SSIM/LPIPS are comparable or even better than other methods although our method does not directly optimize the output to be close to the ground truth. Our method is SOTA in terms of image quality measured by NIQE and FID while also having the best Identity Preservation Score (IPS).

Method	GWAINet	ASFFNet	Codeformer	GFPGAN	Ours
PSNR ↑	22.91	23.72	23.76	23.63	**23.77**
SSIM ↑	0.782	0.814	**0.815**	0.802	0.803
LPIPS ↓	0.383	0.161	0.176	0.203	**0.158**
NIQE ↓	4.854	4.214	4.074	4.364	**4.011**
FID ↓	73.09	58.61	64.22	51.06	**49.46**
IPS ↑	0.2662	0.3660	N/A	N/A	**0.4649**

4. Experiments

We perform quantitative and qualitative comparisons between ReFine and several SOTA blind and reference-guided face restoration approaches. For blind face restoration, we compare with two SOTA methods, GFPGAN [36] and CodeFormer [48]. GFPGAN is a widely used face restoration method based on StyleGAN2 prior; CodeFormer relies on a learned discrete codebook prior. For reference-based methods, we compare with the publicly available GWAINet [10], ASFFNet [24], and DMDNet [26]. GWAINet uses a single reference image and while ASFFNet requires multiple reference images, it only selects one as the input to the pipeline. Thus, it is still suitable for our comparison. DMDNet is a upgraded version of ASFFNet (see Figure 5 upper left for the timeline of the methods). GWAINet can only do super-resolution (SR) while others can do both SR and restoration.

Degradation synthesis For $\times 8$ SR task, we downsize the image and then upscale it by bicubic method. For restoration task, we follow a similar degradation procedure as [36] to generate our training data. The degradation parameters are selected to result in severe degradations. More details can be found in the supplementary.

Dataset Our model is trained on CelebRef-HQ dataset [26], which contains celebrity face images from Bing. This dataset contains 10,555 images with 1,005 identities. Each person has $3 \sim 21$ high-quality images.

Training process We train the model for 800k iterations with a batch size of 4 using Adam optimizern [22], and set the base learning rate to 2×10^{-3} with betas= $(0.5, 0.99)$. The StyleGAN2 prior is finetuned at a learning rate of 4×10^{-4}.

We evaluate our models on three datasets, (1) CASIA-WebFace dataset [41], from which we select 19,557 good-quality image pairs with 10,575 identities based on the BRISQUE metric [30], (2) our new ReFine dataset where we collected 34 celebrities with unique facial features from the internet ourselves, and (3) self-collected real-world data from acquaintances (70 low-quality images from 14 people). We use (1) for quantitative comparison, (2) for qualitative comparison, and (3) for user study.

Quantitative metrics We emphasis that ReFine is focused on details and identity preservation, and some quantitative evaluations do not capture our strength. Thus, we performed extensive qualitative evaluations and user study later.

Restoration quality As pointed out in Section 3.2, we do not directly optimize the output towards the ground truth like other methods [10, 24–26, 35]. However, our metrics in PSNR, SSIM, and LPIPS, which compare the output with the ground truth, are comparable or better than other methods (see Table 1). We also adopt Naturalness Image Quality Evaluator (NIQE) [31] and Fréchet Inception Distance (FID) [13] to measure how closely the restored image distribution matches real scene/face image distribution. While this does not tell us how accurate the model is at preserving identities, it gives us an idea how good the restoration quality is. From Table 1, ReFine compares very favorably compared to previous reference-guided approaches and matches or exceeds the quality of blind restoration approaches.

Restoration accuracy For reference-guided face restoration, it is important for the method to preserve the identity from the reference. We can measure this accuracy by simply computing the cosine similarity between the output image and the reference image and averaging this over the entire testset. We refer to this as Identity Preservation Score (IPS). Because a reference is not available for blind face restoration, we only compute IPS for the reference-guided approaches. From Table 1, ReFine has a substantially better IPS compared to other approaches. That is to say, ReFine is better at copying identity features from the reference, validating our approach.

Qualitative comparisons We provide qualitative comparisons in Figures 5 and 6. ReFine produces natural and accurate images, preserving fine details at an unprecedented

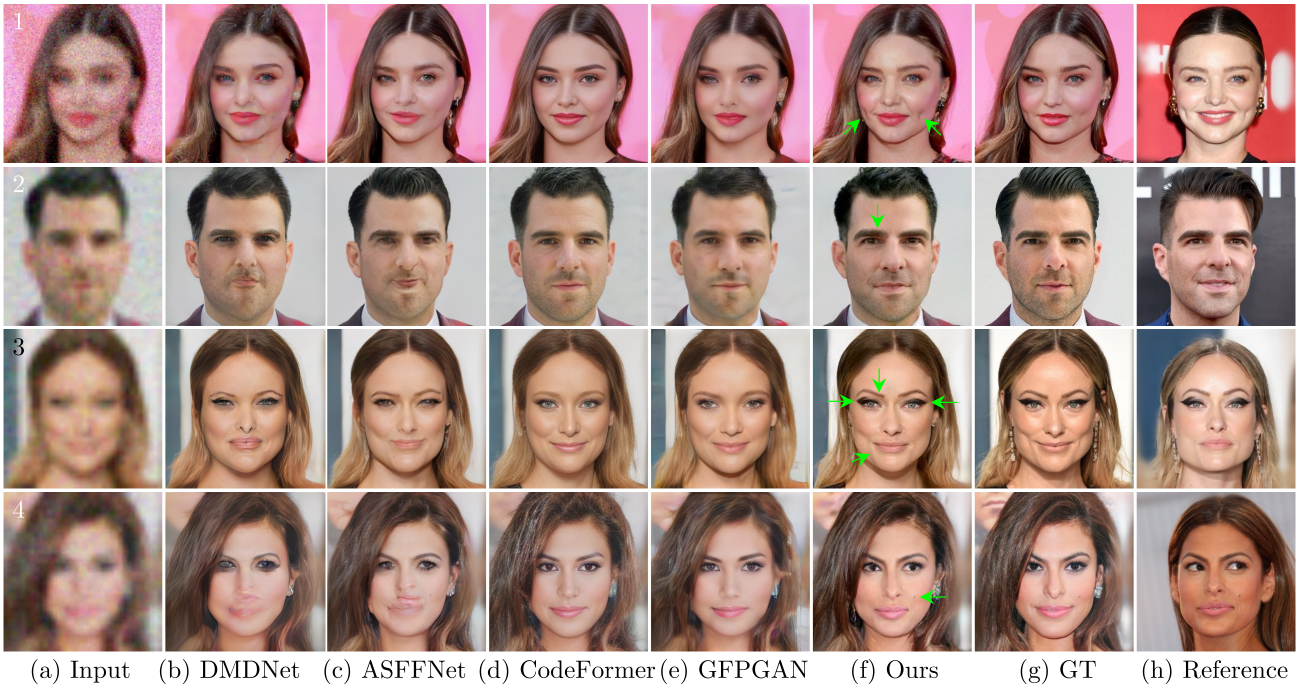

Figure 6. Visual comparison results for face image restoration. Our method is best at preserving identity and facial features (*e.g.* dimples in row 1, eyebrow in 2, eyebrow, eye's makeup and jawline in 3, moles in 4) compared with other works. Please zoom in to see the details.

level, while other approaches either struggle with realism or accuracy. Notably, for restoration task as shown in Figure 6, the reference-guided DMDNet and ASFFNet have a difficult time producing realistic images. This is likely because they do not capitalize on a GFP which has been shown to produce very realistic images. ASFFNet in particular, does have some limited success in preserving some fine-grained facial details, like the eyebrow in row 2, but it is not consistent, often missing out on important facial details, see dimples in row 1 and eye's makeup in row 3. On the other hand, SOTA blind face restoration methods like CodeFormer and GFPGAN produce consistently better quality results compared to their reference-guided counterparts. However, since the input images are heavily degraded, they are unable to preserve the identity and recover the facial details. More results are in the supplementary.

Ablation study If we only include the standard perceptual and adversarial loss like [10, 24–26, 35], it is hard to teach the model whether to "copy" features from the reference or "hallucinate" details directly. In this ablation study, we show the importance of our used losses – spatial minimum loss L_{min} and cycle consistency loss L_{cycle}. L_{min} encourages the network to borrow fine-grained features from the reference image while L_{cycle} prevents the network from copying features indiscriminately from the reference and causing artifacts. Figures 8 and 9 show that ours (GFP + our loss) can better preserve the fine facial details like moles, lips, beard, *etc*. than the baseline method (GFP + traditional loss, where the network is the same except the loss).

In order to understand the design of our network, we did ablation study on limited guidance information or limited loss. We show them in Figure 7. The results demonstrate that each loss and each piece of guidance information play a crucial role in determining the network's final performance.

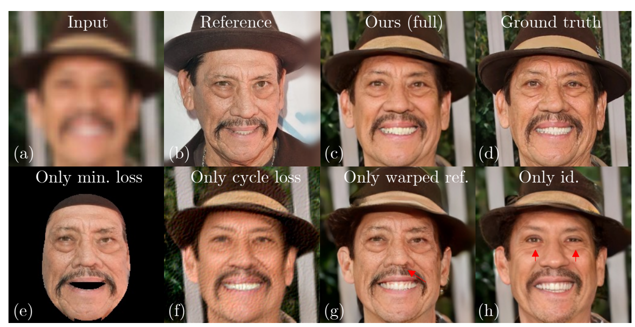

Figure 7. Ablation studies on limited loss or limited guidance information. In the second row, from left to right are results (e) using spatial minimum loss only, (f) cycle consistency loss only, (g) only warped reference (no reference id) for guidance, (h) only reference id (no warped reference) for guidance, respectively. Notice that (g)'s nasal tip is completely wrong, and (h) does not have fine facial details. Please zoom in to see the details.

Changing identities As a consequence of our method, we can achieve identity swapping since our training loss encourages copying from the reference image. Figure 8 shows a few examples with different identities as the reference. ReFine is able to copy the identity from the reference and

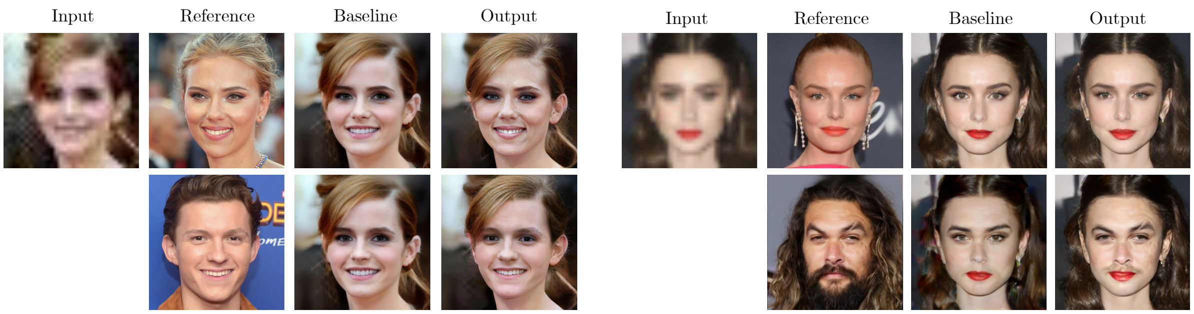

Figure 8. Changing identities when reference image is from a different person. Here “Baseline” is our method but with traditional loss.

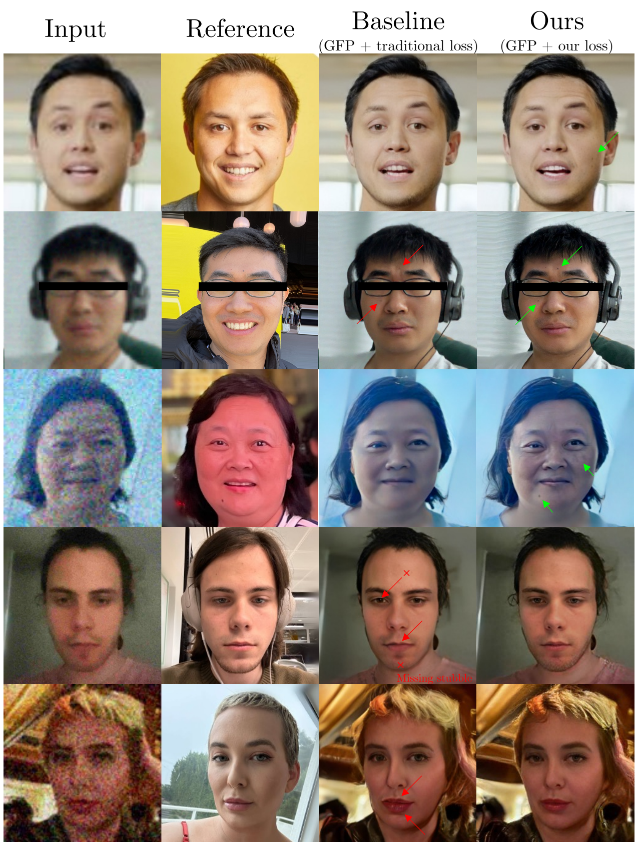

Figure 9. Results on real-world data. The user of the same person can notice the subtle difference easily. **Please zoom in.**

details such as the cheekbone and eyebrow. At the same time, even with such a drastic change in identity (such as using a different gender), the output is still semantically consistent with the input degraded image, *i.e.* it is a reasonable restoration.

Real-world results and user study We show real-world results in Figure 9. Notice the moles in the first three rows, the facial patches in the third row, the eye color, upper lip and beard in the fourth row, and the lips in the fifth row. The user of the same person will notice the subtle difference easily. Because of this, Table 2 introduces a new evaluation strategy: does the same user think the restored image is their own face, or someone else's face? After restoring one's image by several methods, we ask the same user to select their preference. The results show that (1) our method is predominately preferred over all existing methods (Table 2 second row), and (2) our proposed loss is highly effective (Table 2 third row).

Table 2. User study (14 people) for face image restoration tested on real-world data. After restoring one user's images by different methods, we ask the same user to select their preference because they can easily judge the faithfulness. **'Baseline' is the same network but with the traditional loss.** 'NI' means 'not included in the comparison'.

Method	DMDNet	ASFFNet	Codeformer	GFPGAN	Baseline	Ours
User preference ↑	0%	0%	11%	7%	NI	**82%**
User preference ↑	NI	NI	NI	NI	10%	**90%**

5. Conclusion and future work

In this paper, we propose a pipeline which exploits generative facial prior for reference-guided face image restoration. We identify an ambiguity for the output when facial details of the ground truth and the reference do not match, and the ambiguity and traditional loss design can negatively impact restoration quality and accuracy. We then set a new goal to resolve the ambiguity and propose a combination of losses to realize this goal. Our approach ReFine can restore a severely degraded image and preserve identity and fine-grained facial features (like freckles, tattoos, wrinkles, eye color, *etc.*). Extensive experiments show that unprecedented detail preservation is achieved by our method. To our knowledge, ReFine is the first restoration method that works at such granularity, outperforming previous art by a large margin.

Given the proficiency of GANs in handling face domain effectively and their swift processing speed, our experiments are conducted using GAN-based methods. However, the proposed techniques hold the potential for extension to other generative approaches, such as diffusion models [3, 9, 44, 46]. Exploring these alternative methods constitutes a compelling direction for future research.

References

[1] Adrian Bulat and Georgios Tzimiropoulos. Super-fan: Integrated facial landmark localization and super-resolution of real-world low resolution faces in arbitrary poses with gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 109–117, 2018. 2

[2] Kelvin CK Chan, Xintao Wang, Xiangyu Xu, Jinwei Gu, and Chen Change Loy. Glean: Generative latent bank for large-factor image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14245–14254, 2021. 2

[3] Pradyumna Chari, Sizhuo Ma, Daniil Ostashev, Achuta Kadambi, Gurunandan Krishnan, Jian Wang, and Kfir Aberman. Personalized restoration via dual-pivot tuning. *arXiv preprint arXiv:2312.17234*, 2023. 8

[4] 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. 2

[5] Wei-Ting Chen, Gurunandan Krishnan, Qiang Gao, Sy-Yen Kuo, Sizhou Ma, and Jian Wang. Dsl-fiqa: Assessing facial image quality via dual-set degradation learning and landmark-guided transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2931–2941, 2024. 4

[6] 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. 2

[7] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5203–5212, 2020. 4

[8] 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. 3

[9] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 8

[10] Berk Dogan, Shuhang Gu, and Radu Timofte. Exemplar guided face image super-resolution without facial landmarks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019. 2, 3, 4, 5, 6, 7

[11] Jinjin Gu, Yujun Shen, and Bolei Zhou. Image processing using multi-code gan prior. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3012–3021, 2020. 2

[12] 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

[13] 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. 6

[14] Xiaobin Hu, Wenqi Ren, John LaMaster, Xiaochun Cao, Xiaoming Li, Zechao Li, Bjoern Menze, and Wei Liu. Face super-resolution guided by 3d facial priors. In *European Conference on Computer Vision*, pages 763–780. Springer, 2020. 2

[15] Haibo Jin, Shengcai Liao, and Ling Shao. Pixel-in-pixel net: Towards efficient facial landmark detection in the wild. *International Journal of Computer Vision*, 129(12):3174–3194, 2021. 4

[16] 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. 4

[17] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, 34:852–863, 2021. 2

[18] 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

[19] 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, 3

[20] Yury Kartynnik, Artsiom Ablavatski, Ivan Grishchenko, and Matthias Grundmann. Real-time facial surface geometry from monocular video on mobile gpus. *arXiv preprint arXiv:1907.06724*, 2019. 4

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

[22] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[23] 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. 2

[24] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2706–2715, 2020. 3, 4, 5, 6, 7

[25] Xiaoming Li, Ming Liu, Yuting Ye, Wangmeng Zuo, Liang Lin, and Ruigang Yang. Learning warped guidance for blind face restoration. In *The European Conference on Computer Vision (ECCV)*, September 2018. 2, 3, 4, 5, 6, 7

[26] Xiaoming Li, Shiguang Zhang, Shangchen Zhou, Lei Zhang, and Wangmeng Zuo. Learning dual memory dictionaries for

blind face restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 3, 4, 5, 6, 7

[27] Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Ben Fei, Bo Dai, Wanli Ouyang, Yu Qiao, and Chao Dong. Diffbir: Towards blind image restoration with generative diffusion prior. *arXiv preprint arXiv:2308.15070*, 2023. 2

[28] Xuan Luo, Xuaner Zhang, Paul Yoo, Ricardo Martin-Brualla, Jason Lawrence, and Steven M Seitz. Time-travel rephotography. *ACM Transactions on Graphics (TOG)*, 40(6):1–12, 2021. 2

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

[30] Anish Mittal, Anush K Moorthy, and Alan C Bovik. Blind/referenceless image spatial quality evaluator. In *2011 conference record of the forty fifth asilomar conference on signals, systems and computers (ASILOMAR)*, pages 723–727. IEEE, 2011. 6

[31] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012. 6

[32] Yotam Nitzan, Kfir Aberman, Qiurui He, Orly Liba, Michal Yarom, Yossi Gandelsman, Inbar Mosseri, Yael Pritch, and Daniel Cohen-Or. Mystyle: A personalized generative prior. *arXiv preprint arXiv:2203.17272*, 2022. 3

[33] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2287–2296, 2021. 2

[34] 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

[35] Kaili Wang, Jose Oramas, and Tinne Tuytelaars. Multiple exemplars-based hallucination for face super-resolution and editing. In *Proceedings of the Asian Conference on Computer Vision*, 2020. 3, 4, 5, 6, 7

[36] Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. Towards real-world blind face restoration with generative facial prior. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, 2, 4, 6

[37] 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

[38] Zhixin Wang, Ziying Zhang, Xiaoyun Zhang, Huangjie Zheng, Mingyuan Zhou, Ya Zhang, and Yanfeng Wang. Dr2: Diffusion-based robust degradation remover for blind face restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1704–1713, 2023. 2

[39] 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. 2

[40] 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. 1, 2, 3, 4

[41] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014. 6

[42] 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

[43] Zongsheng Yue and Chen Change Loy. Difface: Blind face restoration with diffused error contraction. *arXiv preprint arXiv:2212.06512*, 2022. 2

[44] Howard Chenyang Zhang, Yuval Alaluf, Sizhuo Ma, Achuta Kadambi, Jian Wang, and Kfir Aberman. InstantRestore: Single-step personalized face restoration with shared-image attention. *arXiv preprint*, 2024. 8

[45] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 5

[46] Yi Zhang, Xiaoyu Shi, Dasong Li, Xiaogang Wang, Jian Wang, and Hongsheng Li. A unified conditional framework for diffusion-based image restoration. *Advances in Neural Information Processing Systems*, 36, 2024. 8

[47] Yang Zhao, Yu-Chuan Su, Chun-Te Chu, Yandong Li, Marius Renn, Yukun Zhu, Changyou Chen, and Xuhui Jia. Rethinking deep face restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7652–7661, 2022. 1, 2

[48] Shangchen Zhou, Kelvin Chan, Chongyi Li, and Chen Change Loy. Towards robust blind face restoration with codebook lookup transformer. *Advances in Neural Information Processing Systems*, 35:30599–30611, 2022. 2, 4, 6

[49] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 5