Blemish-aware and Progressive Face Retouching with Limited Paired Data

Lianxin Xie[1], Wen Xue[1], Zhen Xu[1], Si Wu[1,2,3*], Zhiwen Yu[1], and Hau San Wong[4]
[1]School of Computer Science and Engineering, South China University of Technology
[2]Peng Cheng Laboratory
[3]PAZHOU LAB
[4]Department of Computer Science, City University of Hong Kong
{cslianxin.xie, csxuewen, csxuzhen}@mail.scut.edu.cn, {cswusi, zhwyu}@scut.edu.cn,
cshswong@cityu.edu.hk

Abstract

Face retouching aims to remove facial blemishes, while at the same time maintaining the textual details of a given input image. The main challenge lies in distinguishing blemishes from the facial characteristics, such as moles. Training an image-to-image translation network with pixel-wise supervision suffers from the problem of expensive paired training data, since professional retouching needs specialized experience and is time-consuming. In this paper, we propose a Blemish-aware and Progressive Face Retouching model, which is referred to as BPFRe. Our framework can be partitioned into two manageable stages to perform progressive blemish removal. Specifically, an encoder-decoder-based module learns to coarsely remove the blemishes at the first stage, and the resulting intermediate features are injected into a generator to enrich local detail at the second stage. We find that explicitly suppressing the blemishes can contribute to an effective collaboration among the components. Toward this end, we incorporate an attention module, which learns to infer a blemish-aware map and further determine the corresponding weights, which are then used to refine the intermediate features transferred from the encoder to the decoder, and from the decoder to the generator. Therefore, BPFRe is able to deliver significant performance gains on a wide range of face retouching tasks. It is worth noting that we reduce the dependence of BPFRe on paired training samples by imposing effective regularization on unpaired ones.

1. Introduction

With the development of social media, there is an increased demand for facial image beautification from selfies to portraits and beyond. Facial skin retouching aims to remove any unexpected blemishes from facial images, while preserving the stable characteristics that associate with face identity [2, 38, 41]. The main challenge is due to the wide range of blemishes including from small spots to severe acne. Conventional methods are based on blind smoothing, such that the facial characteristics, such as moles and freckles, may be removed. Professional face retouching can be expensive and needs specialized experience, which impedes the collection of large-scale paired data for model training.

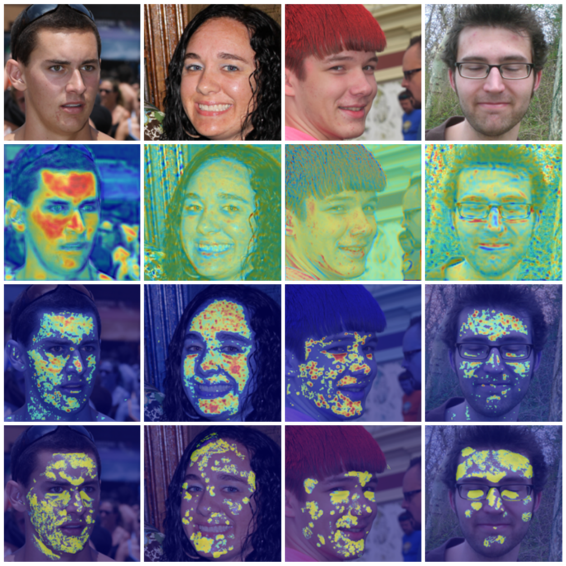

Figure 1. Visual comparison of the activation maps produced by a generic attention module [47] (*second row*) and the blemish-aware module used in BPFRe (*third row*), given a number of images of faces with blemishes (*top row*). BPFRe is capable of applying attention on the regions close to the manual retouching regions (*bottom row*).

Deep neural networks have been widely used for image-to-image translation, especially based on Generative Adver-

[*] Corresponding author.

sarial Networks (GANs) [6, 12, 22]. The translation performance has witnessed rapid progress in style transfer [7], image restoration [45, 48], image inpainting [47], and so on. The existing models are typically based on an encoder-decoder architecture. The source image is encoded into a latent representation, based on which a task-specific transformation is performed by the decoder. Different from the above image enhancement tasks, the regions needed to be retouched may be small, and most of the pixels are unchanged in this case. Generic encoder-decoder-based translation methods can preserve irrelevant content but tend to overlook large blemishes and produce over-smoothed images. Considering that StyleGAN-based methods have the capability of rendering the complex textual details [1, 11], we design a two-stage progressive face retouching framework to make use of the advantage of these types of architectures, and learn the blemish-aware attention (as shown in Figure 1) to guide the image rendering process.

More specifically, we propose a Blemish-aware Progressive Face Retouching model (BPFRe), which consists of two stages: An encoder-decoder architecture is applied at the first stage to perform coarse retouching. The intermediate features from the encoder are integrated into the decoder via skip connections for better reconstruction of image content. At the second stage, we modify the generator architecture of StyleGAN [22] to operate on the multi-scale intermediate features of the decoder and render an image with finer details. We consider that blemish removal cannot be effectively achieved by simply transferring the intermediate features between the components, since there is no mechanism to suppress the blemishes before being passed to the next components. To address this issue, we incorporate two blemish-aware attention modules between the encoder and decoder, and between the decoder and generator, respectively. This design enables progressive retouching by leveraging and refining the information from the previous components. In addition to the paired training images, we use the unpaired ones to optimize the discriminator, which in turn guides the generator to synthesize realistic details. We perform extensive experiments to qualitatively and quantitatively assess BPFRe on both standard benchmarks and data in the wild.

The main contributions of this work are summarized as follows: (a) To deal with a wide range of facial blemishes, we exploit the merits of both encoder-decoder and generator architectures by seamlessly integrating them into a unified framework to progressively remove blemishes. (b) A blemish-aware attention module is incorporated to enhance the collaboration between the components by refining the intermediate features that are transferred among the components. (c) We leverage unpaired training data to regularize the proposed framework, which effectively reduces the dependence on paired training data.

2. Related Work

2.1. Generic Image-to-Image Translation

The capability of Generative Adversarial Networks (GANs) [12] to synthesize high-fidelity image leads to considerable success in a variety of computer vision tasks, such as style transfer [7,8,28], image colorization [17,44], image inpainting [42,46,47], super-resolution [3,26], and so on.

As one of the earliest GAN-based image-to-image translation methods, Isola et al. [17] proposed a conditional GAN, Pix2Pix, to learn the mapping across different domains in a supervised manner. Pix2Pix was based on an encoder-decoder architecture and trained on paired training data. In addition to the adversarial training loss with a domain discriminator, the consistency regularization between the synthesized image and the ground-truth was imposed on the model. To better balance the high-level contextual information and spatial details, Zamir et al. [47] designed a multi-stage image translation structure to progressively restore the degraded images, and the model was referred to as MPRNet. On the other hand, there are a number of image translation models that focus on unpaired data-based training paradigm. Liu et al. [33] trained the coupled GANs [34] to approximate the joint distribution of images from different domains in a shared latent space, and synthesized domain-specific images with the associated decoder. Zhu et al. [50] extended Pix2Pix by performing two-way transformation, and the resulting model was referred to as CycleGAN, in which the unpaired training images were used to impose the cycle consistency regularization on the translation network. A similar strategy was also adopted to learn cross-domain transformations in DiscoGAN [23]. To efficiently learn the mappings among multiple domains, Choi et al. [7, 8] proposed a StarGAN framework, in which a single generator was trained to translate an input image into different domains. The style transfer was performed via adaptive feature normalization [35], conditioned on the learnable domain label embedding. When dealing with multiple conditions, Bhattarai and Kim [5] applied a graph convolutional network [25] to integrate these conditions, and the resulting vector was injected into a translation network to perform a single step transformation. In AttGAN [15], the domain information was encoded as a part of the latent representation, and an auxiliary classifier was incorporated to ensure the correct modification of target content. Furthermore, a selective translation network [32] was used to edit image content according to the domain discrepancies between the input and reference images.

Another research direction is to leverage the pre-trained GANs due to their capability of high-fidelity image synthesis, and significant progress has been made recently [30,43]. An essential step is to map the input image back to the latent space. Perarnau et al. [36] adopted an encoder to learn the mapping from the data space to the latent space, and the image translation was performed by transforming the resulting latent vector. For severely degraded images, the latent codes inferred by an encoder may be insufficient to synthesize reasonable results, and Yang et al. [45] proposed a GAN Prior Embedded Network (GPEN) to inject the encoder features into the generator blocks. To discover semantically meaningful latent directions without supervision, an effective approach was to perform principal component analysis on the latent vectors of the training images [13, 14]. Shen and Zhou [40] performed factorization on the matrix of the generator weights to determine the latent directions which cause substantial variations. In addition, Ding et al. [11] applied sparse dictionary learning to analyze the intra-class variations and discover the class-irrelevant latent directions. However, the semantics associated with the latent directions may not be well-defined. To semantically control the translation, Shen et al. [39] employed support vector machine [9] to determine a latent direction, which effectively classify the instances with and without the target attribute. In [51], a set of latent directions were learnt to manipulate the content, which were required to be identified by a pretrained regressor. To perform complex manipulation, Abdal et al. [1] proposed a conditional normalizing flow model to infer the latent transformation, which corresponds to a nonlinear path in the latent space.

2.2. Face Retouching

Face image beautification is an interesting application of image processing in media and entertainment industry. Conventional face retouching methods are typically based on nonlinear digital filters. In [2], a variety of smoothing filters were designed to remove roughness and small spots. Layvand et al. [27] improved facial attractiveness by searching for similar face images with higher predicted attractive ratings, and determined a 2D warp field for transformation accordingly. For freckle removal, Lipowezky and Cahen [31] extracted them according to color, shape and texture features and replaced them with the surrounding skin. Batool and Chellappa [4] proposed a bimodal Gaussian mixture model to detect facial blemishes, based on Gabor filter responses and texture orientation. Lin et al. [29] model the densities of melanin and hemoglobin as Gaussian, and modify the skin color by adjusting the means and variance. Velusamy et al. [41] adopted a dynamic smoothing filter to remove blemishes and restore the skin texture via wavelet transform. Recently, as a specific image-to-image translation task, GAN-based methods are applied to face retouching. Shafaei et al. [38] established a large-scale and professionally retouched dataset, and built a base model, which is based on the Relativistic Average GAN [19] as well as perceptual and pixel-level consistency regularization.

3. Proposed Method

3.1. Motivation

The main challenge of face retouching lies in detecting and removing blemishes, while at the same time maintaining close similarity with the original. Generic encoder-decoder-based image translation models are typically optimized by perceptual and pixel-level consistency regularization. These models tend to approximate the mean of local skin and thus fail to remove large blemishes. To address this issue, we partition face retouching into two manageable stages. As shown in Figure 2, there are an encoder and a decoder at the first stage for encoding the global structure, background and local detail together with coarsely retouching. A generator at the second stage aims to achieve more desirable results, conditioned on the decoder features. We consider that precisely suppressing blemishes is crucial to guide the model to fill and replace the contents within the blemish area. Toward this end, we design blemish-aware attention modules to suppress blemishes by weighting the multi-scale intermediate feature maps transferred between the components, rather than simply concatenating the components. As a result, the two stages are seamlessly integrated, and are able to remove the blemishes naturally while making the skin look smooth and clear without affecting other content in the images.

3.2. Notations

We concentrate on the challenging case where a limited amount of paired training images are provided. Let $X^{pair} = \{(x^p_{raw}, x^p_{ret})\}$ denote the set of the raw images x^p_{raw} and paired retouching images x^p_{ret}. In addition, there are a large amount of unpaired data: the raw images $X^{unp}_{raw} = \{x^u_{raw}\}$ and retouched images $X^{unp}_{ret} = \{x^u_{ret}\}$. We typically have $|X^{pair}| \ll |X^{unp}_{raw}| + |X^{unp}_{ret}|$. For simplicity, we use x_{raw} and x_{ret} to represent any raw and retouched images, respectively. The first stage of BPFRe contains a blemish-aware attention module A^{E2D} connecting an encoder E and a decoder D_{deco} for coarse retouching. The second stage consists of a generator G, a discriminator D_{disc}, and an attention module A^{D2G} connecting D_{deco} and G for high-fidelity image synthesis.

3.3. Blemish-aware Attention Module

To better reconstruct the global structure, background and textual detail of a given input image, we transfer the intermediate features of the encoder to the decoder, and further transfer those of the decoder to the generator. The skip connection widely used in the U-Net architectures cannot achieve our purpose of suppressing blemishes and other undesirable skin components. To address this issue, we incorporate a blemish-aware attention module to weight the features before propagating them to the next components.

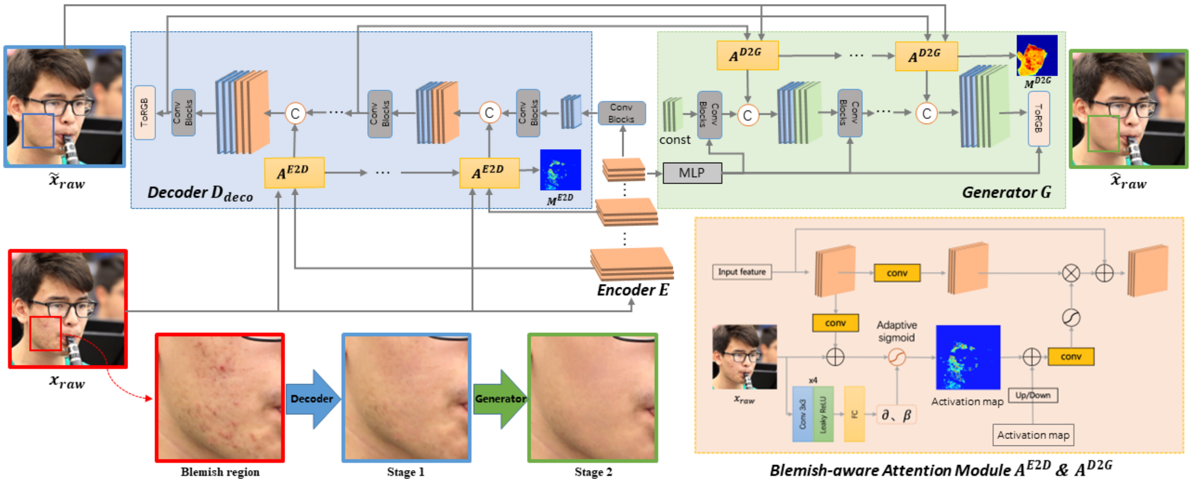

Figure 2. The framework of BPFRe for face retouching. At the first stage, an encoder E and a decoder D_{deco} are trained to coarsely retouch the raw image and retain the global information. To further improve the realism of the content, especially in the blemish regions, a generator G is conditioned on the latent code inferred by E and the intermediate features of D_{deco}. To guide D_{deco} and G to focus more on the blemish regions, two blemish-aware attention modules A^{E2D} and A^{D2G} are incorporated to weight the features transferred from E to D_{deco} and from D_{deco} to G, where the activation parameters, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, are adaptively learnt for local attention.

The module $A^{E2D} = \{h_{param}, h_{map}, h_{weight}\}$ takes the encoder features $f_E(x_{raw})$ and produces a soft mask $\mathcal{M}^{E2D}$, which is expected to highlight blemishes. According to our observation, identifying blemishes heavily depends on the surrounding skin. In view of this, we adopt the convolutional block h_{param} to learn the parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ of the sigmoid activation at local regions as follows:

$$[\boldsymbol{\alpha}, \boldsymbol{\beta}] = h_{param}(x_{raw}), \tag{1}$$

and normalize the attention map produced by h_{map} as:

$$\mathcal{M}^{E2D}|_{u,v} = \frac{1}{1 + \exp^{-\boldsymbol{\alpha} h_{map}(f_E(x_{raw})) - \boldsymbol{\beta}}}\bigg|_{u,v}, \tag{2}$$

where (u, v) denotes a pixel location. To recalibrate the encoder features, we can suppress the blemishes by reducing the response in the attention map, and adopt the block h_{weight} to infer the weighting maps as follows:

$$\mathcal{W}^{E2D} = h_{weight}(1 - \mathcal{M}^{E2D}), \tag{3}$$

where $\mathcal{W}^{E2D}$ have the same dimension as $f_E(x_{raw})$. A^{E2D} is jointly optimized with the other components in the training process. In addition, we construct an explicit supervision in the form of the difference between the raw image and the retouched one as follows:

$$\mathcal{L}^{coarse}_{blem} = \mathbb{E}_{x^p_{raw}} \left[|\mathcal{M}^{E2D} - \Delta^{coarse}|_1 \right], \tag{4}$$

where $\Delta^{coarse} = |x^p_{raw} - x^p_{ret}|_1$. Training with the supervision delivers more reliable attention-guided features for the retouching task. Similarly, we adopt the same architecture to build the other attention module A^{D2G}, and define the corresponding loss as follows:

$$\mathcal{L}^{refine}_{blem} = \mathbb{E}_{x^p_{raw}} \left[|\mathcal{M}^{D2G} - \Delta^{refine}|_1 \right], \tag{5}$$

where $\mathcal{M}^{D2G}$ denotes the attention map yielded by A^{D2G}, and $\Delta^{refine} = |\tilde{x}^p_{raw} - x^p_{ret}|_1$ represents the difference between the output of D_{deco} and retouching ground truth.

3.4. Progressive Retouching

At the first stage, the decoder D_{deco} performs coarse retouching on a raw image, conditioned on the weighted encoder features. For the paired training data X^{pair}, the retouching ground truth is available, and D_{deco} is encouraged to infer the ground truth as accurately as possible. For any retouched image x_{ret}, D_{deco} is required to recover them. By integrating the two aspects, the training loss of the decoder is defined as follows:

$$\begin{aligned} \mathcal{L}^{coarse}_{cons} = \; & \mathbb{E}_{x^p_{raw}} \left[|\tilde{x}^p_{raw} - x^p_{ret}|_1 + |\phi(\tilde{x}^p_{raw}) - \phi(x^p_{ret})|_1 \right] \\ & + \lambda \mathbb{E}_{x_{ret}} \left[|\tilde{x}_{ret} - x^p_{ret}|_1 + |\phi(\tilde{x}_{ret}) - \phi(x_{ret})|_1 \right], \end{aligned} \tag{6}$$

where $\tilde{x}^p_{raw}$ represents the output of D_{deco}:

$$\tilde{x}^p_{raw} = D_{deco}\big(\mathcal{W}^{E2D} \otimes f_E(x^p_{raw})\big), \tag{7}$$

$\otimes$ is the Hadamard product, $\phi(\cdot)$ denotes the features associated with a VGG network pre-trained on ImageNet [10, 18], and the weighting factor λ is used to control the impact of unpaired training data.

When the blemishes are large (e.g., the acne is severe), the first stage may perform less satisfactorily in synthesizing clear face images. Considering the desirable generation capability of StyleGAN [21, 22], we exploit the merits of StyleGAN2 and design the second stage, where the generator receives the latent code inferred by the encoder as well as the weighted intermediate features from the decoder. We represent the synthesized image as follows:

$$\hat{x}_{raw} = G(E(x_{raw}), \mathcal{W}^{D2G} \otimes f_{deco}(x_{raw})), \quad (8)$$

where $E(\cdot)$ denotes the latent representation learnt by E, $\mathcal{W}^{D2G}$ represents the weighting maps, and $f_{deco}(\cdot)$ are the intermediate features of D_{deco}. Similar to StyleGAN2, $E(x_{raw})$ is fed into a MLP to obtain a style code that is broadcasted to each block for feature normalization. On the other hand, $f_{deco}(x_{raw})$ provides rich image information, largely alleviating the difficulty of synthesizing high-fidelity images, and $\mathcal{W}^{D2G}$ prevents the generation process from producing blemishes. We impose the pixel-level and perceptual consistency regularization on G, and the corresponding loss is defined as follows:

$$\mathcal{L}_{cons}^{refine} = \mathbb{E}_{x_{raw}^{p}} \big[|\hat{x}_{raw}^{p} - x_{ret}^{p}|_1 + |\phi(\hat{x}_{raw}^{p}) - \phi(x_{ret}^{p})|_1 \big]. \quad (9)$$

Furthermore, we adopt an adversarial training approach, in which a discriminator D_{disc} is trained to distinguish the retouched images from the raw ones, and G aims to deceive D_{disc}. We define the adversarial training loss as follows:

$$\begin{aligned} \mathcal{L}_{adv}^{disc} = \mathbb{E}_{x_{raw}} \big[& \mathcal{M}^{D2G} \otimes \log(1 - D_{disc}(x_{raw})) \\ & + \mathcal{M}^{D2G} \otimes \log(1 - D_{disc}(\hat{x}_{raw})) \big] \quad (10) \\ & + \mathbb{E}_{x_{ret}} \big[\log D_{disc}(x_{ret}) \big], \end{aligned}$$

$$\mathcal{L}_{adv}^{synt} = \mathbb{E}_{x_{raw}} \big[\mathcal{M}^{D2G} \otimes \log(1 - D_{disc}(\hat{x}_{raw})) \big], \quad (11)$$

where $D_{disc}(\cdot)$ represents the pixel-wise real-fake identification result. Different from generic adversarial loss, we use the produced attention map to weight the real-fake identification result, such that the generator is induced to apply more attention on the regions that correspond to blemishes.

By integrating the above aspects, the optimization formulation of the constituent networks can be expressed as follows:

$$\begin{aligned} & \min_{A^{E2D}, E, D_{deco}} \mathcal{L}_{blem}^{coarse} + \mathcal{L}_{cons}^{coarse}, \\ & \min_{A^{D2G}, G, D_{disc}} \mathcal{L}_{blem}^{refine} + \mathcal{L}_{cons}^{refine} + \mathcal{L}_{adv}^{synt}, \quad (12) \\ & \max_{D_{disc}} \mathcal{L}_{adv}^{disc}. \end{aligned}$$

All the consistent networks in the proposed model are jointly optimized from scratch.

4. Experiments

We evaluate BPFRe on a variety of face retouching tasks. The experiments mainly involve four aspects: (1) We verify the effectiveness of the attention module in specifying the blemishes. (2) We investigate the relative contributions of the main components on face retouching. (3) We further quantitatively and qualitatively compare BPFRe with state-of-the-art image translation models. (4) We finally explore the applicability of BPFRe to image inpainting tasks.

4.1. Datasets and Evaluation Metrics

Datasets. The main experiments are conducted on a large-scale face retouching dataset: FFHQR [38], which is derived from the Flickr-Face-HQ (FFHQ) dataset [21] and covers a wide range of ethnicities and ages. There are 56,000, 7,000 and 7,000 pairs of raw and retouched images for training, validation and testing, respectively. In addition, we collect 1,000 images of faces with large blemishes in the wild to evaluate the performance of the proposed model and competing methods. There are no retouching ground truth available, and we purely use in-the-wild data for testing.

Evaluation metrics. We perform quantitative evaluation based on Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM), which are both widely used metrics in various vision tasks. To further measure the diversity and degree of realism of the synthesized data, we report the Fréchet Inception Distances (FID) [16] and the Learned Perceptual Image Patch Similarity (LPIPS) [49].

4.2. Implementation Details

BPFRe consists of a U-Net $\{E, D_{deco}\}$, a generator G, a discriminator D_{disc} and two attention modules $\{A^{E2D}, A^{D2G}\}$. We adopt the architecture of [37] for the U-Net, and there are 16 and 16 blocks for the encoder and decoder, respectively. In addition, we adopt the StyleGAN2 architecture for the generator and discriminator. A^{E2D} and A^{D2G} have the same light weight architecture that contains 7 convolutional layers. We implement BPFRe using PyTorch on a NVIDIA Tesla V100 GPU. The weighting factor λ in Eq.(6) is set to 0.001. We adopt the Adam optimizer [24] with a learning rate of 0.002. BPFRe is trained for 120k iterations with a batch size of 2.

4.3. Effectiveness of Blemish-aware Attention

We begin by visually verifying the effectiveness of the attention modules A^{E2D} and A^{D2G} in identifying the regions that could be blemishes. In the proposed model, the two modules are expected to suppress the blemishes by weighting the intermediate features transferred from the encoder to the decoder, and from the decoder to the generator, such that the content in the blemish regions can be synthesized from contextual information. In Figure 3, we visualize the attention maps $\mathcal{M}^{E2D}$ and $\mathcal{M}^{D2G}$ and the corresponding synthesized images $\tilde{x}_{raw}$ and $\hat{x}_{raw}$ at the two stages. We observe that both attention maps cover most of the blemish regions indicated in the difference map

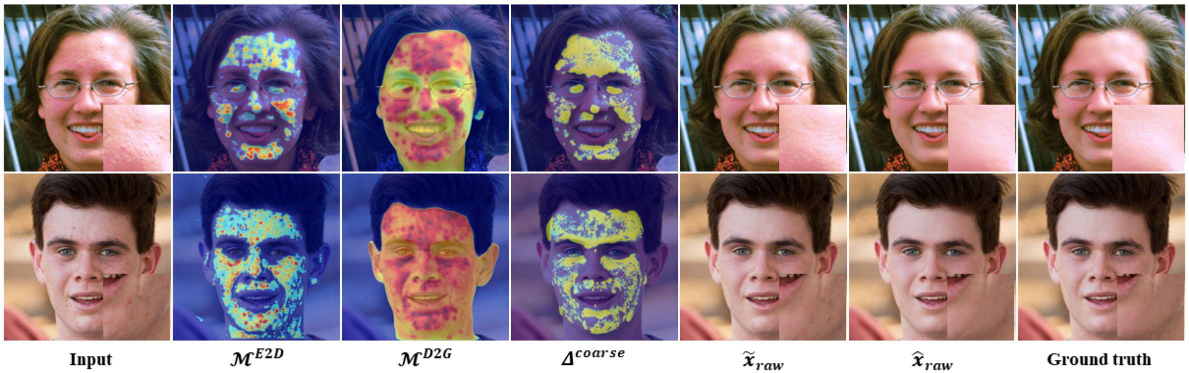

Figure 3. Visualization of attention maps and corresponding retouching results at the two stages of BPFRe.

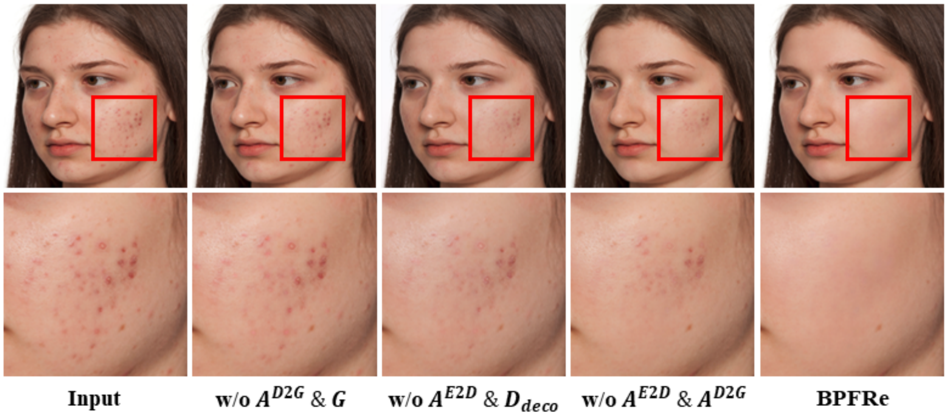

Figure 4. Face retouching results of BPFRe and the ablative models.

$\Delta^{coarse} = |x^p_{raw} - x^p_{ret}|_1$ between the raw image and retouching ground truth. Compared to the first stage, the second stage focuses more on synthesizing the details, and the high response area in $\mathcal{M}^{D2G}$ is greater than that in $\mathcal{M}^{E2D}$. As a result, the second stage is able to produce clear face images with realistic details.

4.4. Ablation Study

To analyze the roles of the main components of BPFRe, we perform ablative experiments by constructing three variants. The first variant is built by disabling the decoder and denoted by 'BPFRe w/o A^{E2D} & D_{deco}'. The encoder features are weighted by A^{D2G} and transferred to the generator. We build the second variant by removing the generator, and the resulting model is referred to as 'BPFRe w/o A^{D2G} & G'. In addition to minimizing $\mathcal{L}^{coarse}_{cons}$ in Eq.(6), the U-Net also competes with the discriminator. 'BPFRe w/o A^{E2D} & A^{D2G}' is the third variant, which refers to our model without the blemish-aware attention modules.

We summarize the PSNR, SSIM and LPIPS results of the proposed BPFRe and its three variants in Table 1. One can observe that the full model is able to achieve better quantitative results than its variants in terms of all the metrics. Both 'BPFRe w/o A^{D2G} & G' and 'BPFRe w/o A^{E2D} & D_{deco}' are single-stage models. When compared to 'BPFRe w/o A^{D2G} & G', the second stage of BPFRe leads to a PSNR gain of 1.49. When disabling the first stage of BPFRe, the performance drop reaches 4.59 PSNR points, although 'BPFRe w/o A^{E2D} & D_{deco}' also includes the generator. This implies that the two-stage architecture plays an important role in the generation process. In addition, we consider that without the blemish-aware attention modules, the result of 'BPFRe w/o A^{E2D} & A^{D2G}' is not as good as BPFRe. We also show representative results of the methods in Figure 4, and find that 'BPFRe w/o A^{E2D} & A^{D2G}' produces a better retouching image than the other variants but cannot completely remove the blemishes. This demonstrates the effectiveness of the combination of our two-stage architecture and blemished-aware attention mechanism.

Table 1. Results of BPFRe and the ablative models on FFHQR.

Method	PSNR ↑	SSIM ↑	LPIPS ↓
BPFRe w/o A^{E2D} & D_{deco}	40.70	0.9883	0.0140
BPFRe w/o A^{D2G} & G	43.80	0.9915	0.0135
BPFRe w/o A^{E2D} & A^{D2G}	44.38	0.9923	0.0101
BPFRe	**45.29**	**0.9935**	**0.0092**

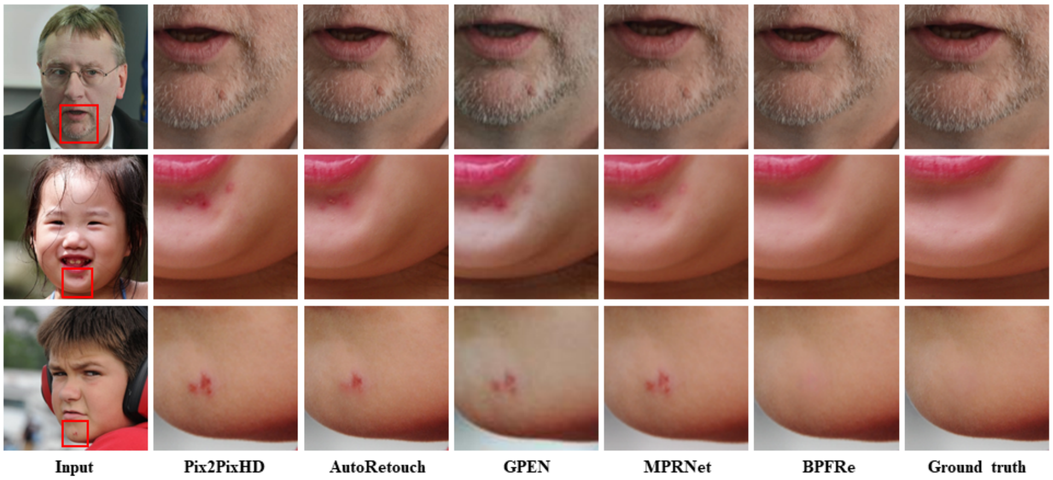

Figure 5. Visual comparison of BPFRe and the competing methods on the FFHQR images.

Table 2. Results of BPFRe and competing methods on FFHQR.

	PSNR ↑		SSIM ↑		LPIPS ↓	
Method	*All*	*Hard*	*All*	*Hard*	*All*	*Hard*
Raw images	43.89	37.69	0.9910	0.982	0.9906	0.0310
Pix2PixHD [17]	29.38	30.10	0.9181	0.9035	0.0766	0.0844
GPEN [45]	43.12	37.88	0.9911	0.9792	0.0141	0.0697
AutoRetouch [38]	44.18	38.01	0.9910	0.9812	0.0133	0.0292
MPRNet [47]	44.35	38.67	0.9931	0.9854	0.0129	0.0301
BPFRe	**45.29**	**38.98**	**0.9935**	**0.9856**	**0.0092**	**0.0205**

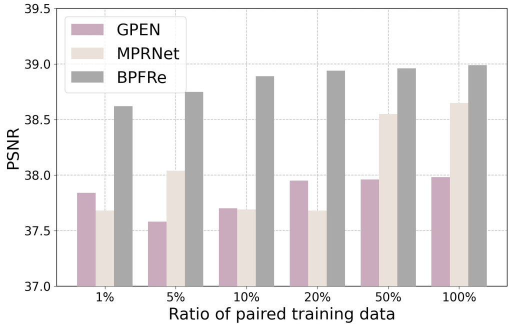

Figure 6. The impact of the amount of paired training data on the performance of BPFRe and competing methods on FFHQR-*Hard*.

4.5. Comparison

To demonstrate the superiority of BPFRe, we compare the proposed model with a number of representative competing methods, including Pix2PixHD [17], MPRNet [47], GPEN [45] and AutoRetouch [38]. Pix2PixHD is a typical image-to-image translation method, MPRNet and GPEN serve as state-of-the-art image restoration methods, and AutoRetouch focuses on the face retouching task.

4.5.1 Results on FFHQR

To demonstrate the capability of our framework to remove blemishes, we manually select about 1.4k comparatively difficult images to build a subset: FFHQR-*Hard*. We summarize the results of BPFRe and the competing methods in Table 2. One can find that BPFRe outperforms the competing methods in terms of PSNR, SSIM and LPIPS. Compared to AutoRetouch that serves as a customized method for this task, BPFRe still achieves a competitive advantage in terms of all the metrics. We visually compare the methods on challenging images in Figure 5. BPFRe is effective in removing blemishes of different types and scales, and the produced images are visually pleasant and consistent with the ground-truth data. On the other hand, the competing methods do not completely remove the blemishes. We further compare with MPRNet and GPEN when the amount of paired training images decreases. The proportion of paired data is limited in the range of {1%, 5%, 10%, 20%, 50% 100%}. Figure 6 demonstrates that BPFRe consistently obtains better PSNR scores than the competing methods on FFHQR-*Hard*. Although GPEN adopts the StyleGAN generator, our improvement over GPEN is as large as 0.78 to 1.19 PSNR point(s).

4.5.2 Results on Images in the Wild

We further evaluate BPFRe and the competing methods on in-the-wild face images. It is worth noting that all the methods are trained only on FFHQR. Figure 7 presents the representative synthesized images. These results lead to similar conclusions as the experiment on FFHQR. GPEN fails to perform retouching on the images. AutoRetouch oversmoothes the content and has limited generalization capability to synthesize realistic content in the blemish regions. MPRNet is able to restore the textual detail but fails to remove severe acne. In contrast, BPFRe removes blemishes

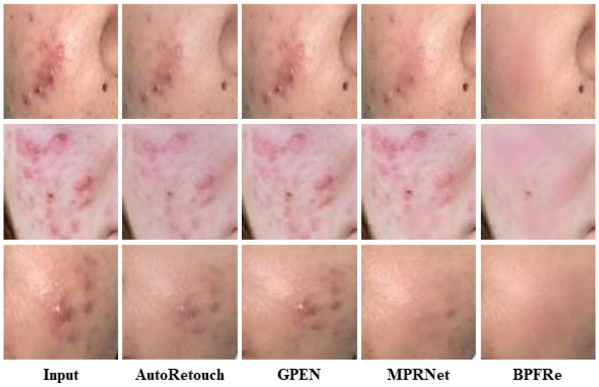

Figure 7. Visual comparison of BPFRe and the competing methods on the in-the-wild face images.

Table 3. The voting result (%) of user study on in-the-wild data.

Method	Rank-1	Rank-2	Rank-3	Rank-4	Rank-5
Pix2PixHD [17]	0.001	0.005	5.303	5.650	**88.258**
GPEN [45]	0.387	1.053	15.909	**74.388**	9.589
AutoRetouch [38]	0.003	9.298	**71.022**	16.384	1.566
MPRNet [47]	14.401	**74.211**	7.576	3.577	0.587
BPFRe	**85.208**	15.433	0.190	0.001	0.000

Table 4. The image inpainting results of BPFRe and the competing methods on CelebA-HQ.

	1%		5%		10%		20%	
Method	FID ↓	PSNR ↑	FID ↓	PSNR ↑	FID ↓	PSNR ↑	FID ↓	PSNR ↑
MPRNet [47]	41.73	25.79	27.27	28.02	28.45	**28.76**	28.41	28.21
GPEN [45]	14.21	26.02	10.56	26.37	10.88	26.26	10.47	26.30
BPFRe	**7.57**	**28.17**	**7.20**	**28.19**	**7.16**	28.21	**7.29**	**28.32**

naturally, and make the skin look clear and smooth, which demonstrates the strong generalization capability.

4.6. User Study

We perform a subjective evaluation on in-the-wild data, and there are 50 questions constructed. Given a raw image, the workers are required to rank the retouching results of BPFRe and the competing methods, and high-ranking results should represent delightful content with realistic details. For a fair assessment, the results of the methods are presented in a random order. We employ 80 validated workers to answer each question, and Table 3 presents the average ranking result. BPFRe achieves the best performance on in-the-wild data, which demonstrates that our results are consistent with human visual perception.

4.7. Applied to Image Inpainting

Although BPFRe is originally designed for face retouching, the attention-guided two-stage architecture is capable of performing image inpainting with limited paired data, and the training loss and optimization scheme can be used directly in this task. We compare BPFRe with the representative state-of-the-art image restoration methods: MPRNet and GPEN. These models are trained on FFHQ for the cases where 1%, 5%, 10% and 20% of the training images are paired. We adopt the image degradation method [45], and evaluate the trained models on CelebA-HQ [20]. The results are summarized in Table 4. As the amount of paired training data decreases, the superiority of BPFRe over the competing methods becomes significant (up to 6.64 FID and 2.15 PSNR points). Figure 8 shows that BPFRe is still able to produce reasonable results for severely degraded images.

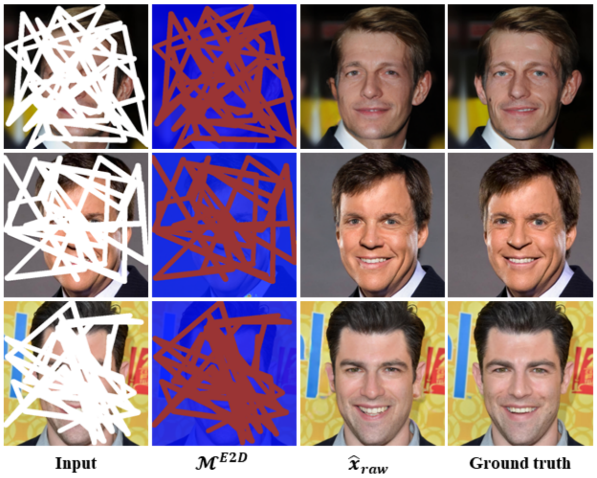

Figure 8. Visualization of the attention maps and corresponding inpainting results on the CelebA-HQ images.

5. Conclusion

In this paper, we propose an attention-guided progressive face retouching framework to remove blemishes naturally and synthesize high-fidelity content. We design a two-stage structure to exploit the merit of the U-Net architecture in restoring the image details and that of the GAN generator architecture in generating realistic images. The core idea is to explicitly suppress blemishes when transferring the intermediate features from the encoder to the decoder, and from the decoder to the generator. Toward this end, we adopt a blemish-aware attention module to learn the weighting maps. Our model can be effectively trained on partially paired data, and the experimental results demonstrate the effectiveness qualitatively and qualitatively.

Acknowledgments

This work was supported in part by the China Scholarship Council, in part by the National Natural Science Foundation of China (Project No. 62072189), in part by the Research Grants Council of the Hong Kong Special Administration Region (Project No. CityU 11206622), and in part by the Natural Science Foundation of Guangdong Province (Project No. 2022A1515011160).

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*, 40(3):1–21, 2021. 2, 3

[2] Kaoru Arakawa. Nonlinear digital filters for beautifying facial images in multimedia systems. In *Proc. IEEE International Symposium on Curcuits and Systems*, 2004. 1, 3

[3] Yuval Bahat and Tomer Michaeli. Explorable super resolution. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 2

[4] Nazre Batool and Rama Chellappa. Detection and inpainting of facial wrinkles using texture orientation fields and Markov random field modeling. *IEEE Transactions on Image Processing*, 23(9):3773–3788, 2014. 3

[5] Binod Bhattarai and Tae-Kyun Kim. Inducing optimal attribute representations for conditional GANs. In *Proc. European Conference on Computer Vision*, 2020. 2

[6] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natual image synthesis. In *Proc. International Conference on Learning Representation*, 2019. 2

[7] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. StarGAN: unified generative adversarial networks for multi-domain image-to-image translation. In *Proc. IEEE conference on Computer Vision and Pattern Recognition*, pages 8789–8797, 2018. 2

[8] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. StarGAN v2: diverse image synthesis for multiple domains. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, June 2020. 2

[9] Corinna Cortes and Vladimir Vapnik. Support vector networks. *Machine learning*, 20(3):273–297, 1995. 3

[10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: a large-scale hierarchical image database. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2009. 4

[11] Guanqi Ding, Xinzhe Han, Shuhui Wang, Shuzhe Wu, Xin Jin, Dandan Tu, and Qingming Huang. Attribute group editing for reliable few-shot image generation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 2, 3

[12] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Proc. Neural Information Processing Systems*, 2014. 2

[13] Erik Harkonen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. GANSpace: discovering interpretable GAN controls. In *Proc. Neural Information Processing Systems*, 2020. 3

[14] Zhenliang He, Meina Kan, and Shiguang Shan. EigenGAN: layer-wise eigen-learning for GANs. In *Proc. International Conference on Computer Vision*, 2021. 3

[15] Zhenliang He, Wangmeng Zuo, Meina Kan, Shiguang Shan, and Xilin Chen. AttGAN: facial attribute editing by only changing what you want. *IEEE Transactions on Image Processing*, 28(11):5464–5478, 2019. 2

[16] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, and Bernhard Nessler. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Proc. Neural Information Processing Systems*, 2017. 5

[17] Philip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2, 7, 8

[18] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Proc. European Conference on Computer Vision*, 2016. 4

[19] Alexia Jolicoeur-Martineau. The relativistic discriminator: a key element missing from standard GAN. In *Proc. International Conference on Learning Representation*, 2019. 3

[20] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *Proc. International Conference on Learning Representation*, 2018. 8

[21] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 5

[22] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 2, 5

[23] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. In *Proc. International Conference on Machine Learning*, 2017. 2

[24] Diederik P. Kingma and Jimmy Lei Ba. Adam: a method for stochastic optimization. In *Proc. International Conference on Learning Representation*, 2015. 5

[25] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *Proc. International Conference on Learning Representation*, 2017. 2

[26] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2

[27] Tommer Leyvand, Daniel Cohen-Or, Gideon Dror, and Dani Lischinski. Data-driven enhancement of facial attractiveness. In *Proc. ACM Conference on Special Interest Group on Computer Graphics and Interactive Techniques*, 2008. 3

[28] Tianwei Lin, Zhuoqi Ma, Fu Li, Dongliang He, Xin Li, Errui Ding, Nannan Wang, Jie Li, and Xinbo Gao. Drafting and revision: Laplacian pyramid network for fast high-quality artistic style transfer. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[29] Tsung-Ying Lin, Yu-Ting Tsai, Tsung-Shian Huang, Wen-Chieh Lin, and Jung-Hong Chuang. Exemplar-based freckle retouching and skin tone adjustment. *Computers & Graphics*, 78:54–63, 2019. 3

[30] Huan Ling, Karsten Kreis, Daiqing Li, Seung Wook Kim, Antonio Torralba, and Sanja Fidler. EditGAN: high-precision semantic image editing. In *Proc. Neural Information Processing Systems*, 2021. 2

[31] Uri Lipowezky and Sarah Cahen. Automatic freckles detection and retouching. In *Proc. IEEE Convention of Electrical and Electronics Engineers in Israel*, 2008. 3

[32] Ming Liu, Yukang Ding, Min Xia, Xiao Liu, Errui Ding, Wangmeng Zuo, and Shilei Wen. STGAN: a unified selective transfer network for arbitrary image attribute editing. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2

[33] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *Proc. Neural Information Processing Systems*, 2017. 2

[34] Ming-Yu Liu and Oncel Tuzel. Coupled generative adversarial networks. In *Proc. Neural Information Processing Systems*, 2016. 2

[35] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2

[36] Guim Perarnau, Joost van de Weijer, Bogdan Raducanu, and Jose M. Alvarez. Invertible conditional GANs for image editing. In *Proc. NIPS workshop on Adversarial Training*, 2016. 2

[37] 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 *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 2287–2296, 2021. 5

[38] Alireza Shafaei, James J. Little, and Mark Schmidt. AutoRetouch: automatic professional face retouching. In *Proc. IEEE Winter Conference on Applications of Computer Vision*, 2021. 1, 3, 5, 7, 8

[39] Yujun Shen, Ceyuan Yang, Xiaoou Tang, and Bolei Zhou. InterFaceGAN: interpreting the disentangled face representation learned by GANs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(4):2004–2018, 2020. 3

[40] Yujun Shen and Bolei Zhou. Closed-form factorization of latent semantics in GANs. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 3

[41] Sudha Velusamy, Rishubh Parihar, Raviprasad Kini, and Aniket Rege. FabSoften: face beautification via dynamic skin smoothing, guided feathering and texture restoration. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshop*, 2020. 1, 3

[42] Tengfei Wang, Hao Ouyang, and Qifeng Chen. Image inpainting with external-internal learning and monochromic bottleneck. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 2

[43] Po-Wei Wu, Yu-Jing Lin, Che-Han Chang, Edward Y. Chang, and Shih-Wei Liao. RelGAN: multi-domain image-to-image translation via relative attributes. In *Proc. International Conference on Computer Vision*, 2019. 2

[44] Yanze Wu, Xintao Wang, Yu Li, Honglun Zhang, Xun Zhao, and Ying Shan. Towards vivid and diverse image colorization with generative color prior. In *Proc. IEEE International Conference on Computer Vision*, 2021. 2

[45] Tao Yang, Peiran Ren, Xuansong Xie, and Lei Zhang. GAN prior embedded network for blind face restoration in the wild. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 2, 3, 7, 8

[46] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S. Huang. Generative image inpainting with contextual attention. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2

[47] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Huang Yang, and Ling Shao. Multi-stage progressive image restoration. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 1, 2, 7, 8

[48] Cheng Zhang, Shaolin Su, Yu Zhu, Qingsen Yan, Jinqiu Sun, and Yanning Zhang. Exploring and evaluating image restoration potential in dynamic scenes. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 2

[49] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 5

[50] Jun-Yan Zhu, Taesung Park, Philip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proc. International Conference on Computer Vision*, 2017. 2

[51] Peiye Zhuang, Oluwasanmi Koyejo, and Alexander G. Schwing. Enjoy your editing: controllable GANs for image editing via latent space navigation. In *Proc. International Conference on Learning Representation*, 2021. 3