RetouchFormer: Semi-supervised High-Quality Face Retouching Transformer with Prior-Based Selective Self-Attention

Wen Xue[1*], **Lianxin Xie**[1*], **Le Jiang**[1], **Tianyi Chen**[1], **Si Wu**[1†], **Cheng Liu**[2†], **Hau-San Wong**[3]
[1]South China University of Technology
[2]Shantou University
[3]City University of Hong Kong
{csxuewen, cslianxin.xie, csjiangle, csttychen}@mail.scut.edu.cn,
cswusi@scut.edu.cn, cliu@stu.edu.cn, cshswong@cityu.edu.hk

Abstract

Face retouching is to beautify a face image, while preserving the image content as much as possible. It is a promising yet challenging task to remove face imperfections and fill with normal skin. Generic image enhancement methods are hampered by the lack of imperfection localization, which typically results in incomplete removal of blemishes at large scales. To address this issue, we propose a transformer-based approach, RetouchFormer, which simultaneously identifies imperfections and synthesizes realistic content in the corresponding regions. Specifically, we learn a latent dictionary to capture clean face priors, and predict the imperfection regions via a reconstruction-oriented localization module. Also based on this, we can realize face retouching by explicitly suppressing imperfections in selective self-attention computation, such that local content will be synthesized from normal skin. On the other hand, multi-scale feature tokens lead to increased flexibility in dealing with the imperfections at various scales. The design elements bring greater effectiveness and efficiency. RetouchFormer outperforms the advanced face retouching methods and synthesizes clean face images with high fidelity in our list of extensive experiments performed.

Introduction

Face retouching aims at beautifying a face image that has suffered from numerous types of imperfections. It is promising to automatically create a flawless skin tone, while professional retouchers may take a few hours to edit the photo. A key challenge of face retouching lies in how to automatically recognize face imperfections, which are diverse in the real world. On the other hand, it is also challenging to fill realistic details, especially for moderate-to-severe acne regions, while at the same time preserving the structures of essential facial features.

The traditional face retouching methods apply local smoothing operators for blemish removal (Arakawa 2004; Batool and Chellappa 2014). Recently, significant progress has been made in generic image-to-image translation, such as style transfer (Isola et al. 2017; Kim et al. 2017; Liu, Breuel, and Kautz 2017; Zhu et al. 2017; Wu et al. 2019; Ling et al. 2021), image enhancement (Zamir et al. 2021; Yang et al. 2021; Wang et al. 2022), and so on. AutoRetouch (Shafaei, Little, and Schmidt 2021) and ABPN (Lei et al. 2022) start to focus on face retouching. However, these methods only consider image-level transformations by global convolution filters, and the imperfection regions cannot be adaptively completed. In addition, these methods typically need pairwise retouching data for model training.

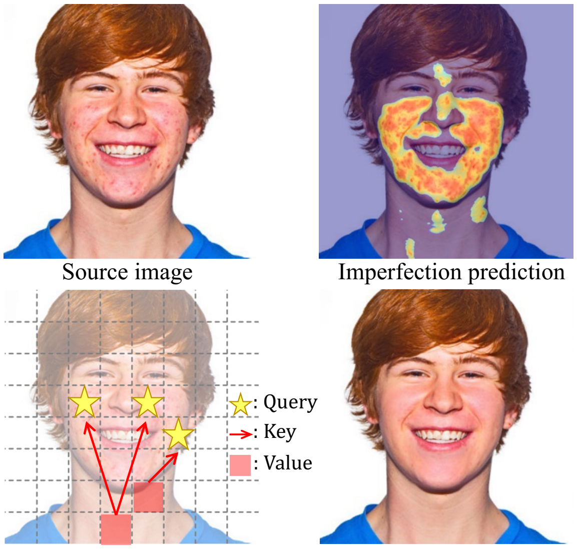

Figure 1: The main idea behind RetouchFormer is to replace the features of the imperfections with those of normal skin via selective self-attention.

In this work, we treat face retouching as a generalized 'soft' version of face image inpainting, and train a transformer with selective self-attention mechanism on partially paired data to synthesize clean face with high realism and fidelity as shown in Figure 1. More specifically, we apply the encoder-transformer-decoder design as shown in Figure 2, and the proposed model is referred to as RetouchFormer, which approximately predicts imperfections to indicate the regions to be filled and simultaneously synthesize realistic details. Toward this end, we learn the clean face prior in

[*]Joint first authors.
[†]Corresponding authors.

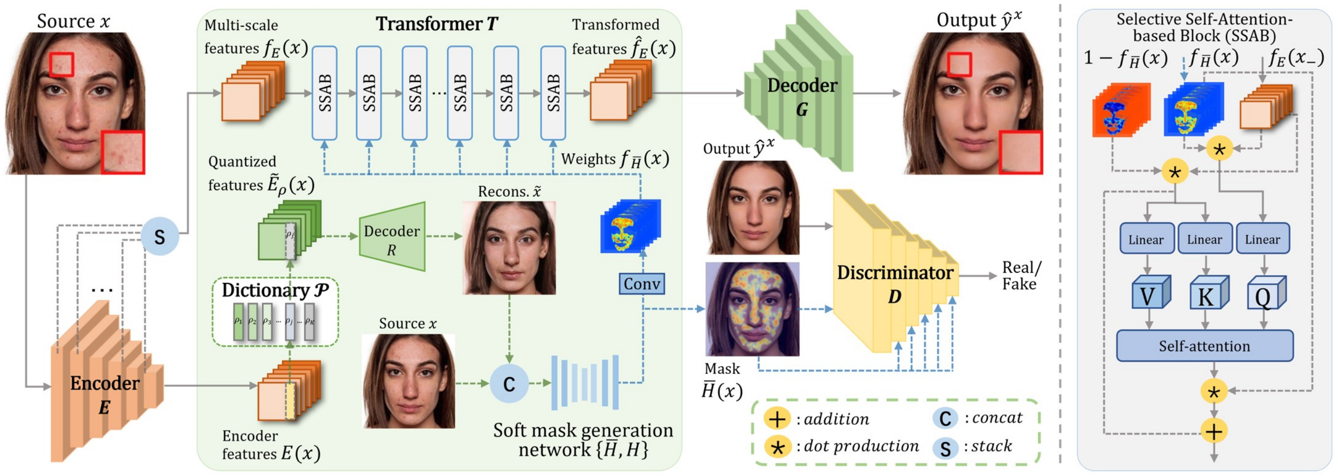

Figure 2: The workflow of RetouchFormer, which is designed to simultaneously predict imperfections and complete the corresponding regions in the encoder feature space. Toward this end, a feature dictionary $\mathcal{P}$ is learnt to capture clean face priors, and is then used for imperfection prediction. Based on the resulting weights $f_{\overline{H}}(x)$, we perform selective self-attention computation by limiting the spatial interactions between imperfections and normal skin, and thus the features of the normal skin regions can be transformed to fill the imperfection regions. By further injecting the masks $\overline{H}(x)$ into a discriminator, more attention is applied to the imperfection regions in real-fake prediction, which in turn improves the synthesized retouching results.

terms of a feature dictionary from retouching images by vector quantization (van den Oord, Vinyals, and Kavukcuoglu 2017). Under the assumption that the skin with imperfections cannot be well reconstructed by the dictionary, the differences before and after feature reconstruction indicate the spatial information related to imperfections to a certain extent. RetouchFormer can conveniently leverage the information for precise correction of face imperfections. Considering that the targets can vary substantially in scale, we adopt a multi-scale encoding scheme in the transformer to hierarchically represent the content in local regions, rather than setting the tokens to a fixed scale. To fill the imperfection regions, the transformer searches contextually coherent contents, and the most relevant patches are identified and transformed by aggregating the attended features. Another key design element of RetouchFormer is its self-attention mechanism with imperfection suppression, based on leveraging the spatial interactions between queries and key-value pairs. The main idea is to softly limit the connections among the tokens associated with imperfection regions, such that the features can be replaced with those associated with normal skin regions. We perform extensive experiments to verify the advantages of RetouchFormer. In summary, the main contribution of this work are as follows: (a) We propose a semi-supervised face retouching transformer to reduce the dependence on paired retouching data for model training. (b) The proposed model, RetouchFormer, is characterized by formulating face retouching as a ‘soft inpainting’ task and realized by joint approximate imperfection prediction and local content synthesis. (c) Based on clean face priors, the spatial information on imperfections is applied to perform selective self-attention computation, through which the spatial interactions between the queries of imperfections and the key-value pairs of normal skin are strengthened.

Related Work

CNN-based Image Translation

Convolutional Neural Networks (CNNs) are the mainstream network architectures in the field of computer vision. The performance of CNN-based image translation has witnessed a rapid progress due to the Generative Adversarial Network (GAN)’s capability of high-fidelity image synthesis (Goodfellow et al. 2014). As a conditional GAN, Pix2Pix (Isola et al. 2017) learnt a mapping to minimize the distribution discrepancy to the target domain data as well as pixel-wise discrepancy to the ground truth. Pix2Pix was trained on the pairwise training data, which impedes its application to the scenarios in which paired data collection is expensive and time-consuming. To overcome this limitation, a number of unpaired image translation methods (Liu, Breuel, and Kautz 2017; Zhu et al. 2017; Kim et al. 2017) performed data distribution alignment between source and target domains or two-way mapping to impose cycle consistency regularization. On the other hand, GPEN (Yang et al. 2021) combined a U-shaped CNN with a GAN to ensure high-quality image generation. MPRNet (Zamir et al. 2021) was designed for multi-stage image translation, which benefits from high-level contextual information and spatial details.

Additionally, the latent space of a pre-trained GAN was observed to possess semantic organization, which allowed semantic editing on images (Goetschalckx et al. 2019; Jahanian, Chai, and Isola 2019; Collins et al. 2020; Shen and Zhou 2021; Wu, Lischinski, and Shechtman 2021). By leveraging high-fidelity GAN inversion methods (Gu, Shen, and Zhou 2020; Richardson et al. 2021; Tov et al. 2021) to project images back to the latent space, semantic manipulation was performed by imposing attribute-associated transformations on the resulting latent vectors. However, these methods perform global latent transformation for semantic image translation, without taking the variation in input images into consideration. In contrast, StyleFlow (Abdal et al. 2021) employed a continuous normalizing flow model to learn a non-linear transformation, conditioned on the starting points in the latent space. By using natural language to express a wide range of visual concepts, StyleGAN-based image editing could benefit from the language-vision pretraining (Patashnik et al. 2021).

Vision Transformer

Transformer (Devlin et al. 2018; Vaswani et al. 2017) is a network architecture designed for modeling long-range dependencies encapsulated in data via attention. Due to its success in the field of natural language processing, transformer has attracted extensive attention and been applied to computer vision. For example, Vision Transformer (ViT) (Dosovitskiy et al. 2021; Wu et al. 2021) showed superior performance in representation learning. To address the high image resolutions and large variations in the scale of visual objects, SwinTransformer (Liu et al. 2021) learnt a hierarchical representation with shifted windows. For low-level computer vision tasks, the Image Processing Transformer (IPT) (Chen et al. 2021) adopted a pre-trained large scale model to perform denoising, deraining, and super-resolution. Different from the self-attention computation in ViT, RestoreFormer (Wang et al. 2022) performed cross-attention with the priors in terms of high-quality key-value pairs to improve the performance of blind face restoration. Apart from the above, the transformer-based methods were also applied to image editing (Hu et al. 2022), object detection (Carion et al. 2020; Zhu et al. 2021), and so on.

Face Retouching

As a specific image translation task, face retouching aims at synthesizing clean and pleasant face images from the ones with imperfections. The traditional methods typically were based on various image processing techniques, and focused more on a single kind of imperfection, such as spots with small scales (Arakawa 2004; Batool and Chellappa 2014; Lipowezky and Cahen 2008; Leyvand et al. 2008). When applying smoothing filters to remove wrinkles and spots, a wavelet-based skin texture restoration method was adopted to restore the lost fine-grained details (Velusamy et al. 2020). AutoRetouch (Shafaei, Little, and Schmidt 2021) was an effective attempt to perform deep image translation for the task, and the retouching performance benefited from the GAN-based framework. To address blemishes at large scale, BPFRe (Xie et al. 2023) adopted a two-stage retouching strategy to progressively restore clean face. ABPN (Lei et al. 2022) propose an adaptive blend pyramid network, which achieved fast local retouching on high-resolution photos.

Different from the above methods, this work explores the adaptation of transformer to face retouching, and is most related to the state-of-the-art methods RestoreFormer (Wang et al. 2022) and BPFRe. The similarity between RetouchFormer and RestoreFormer mainly lies in that the prior learnt from the training data is in terms of a dictionary: RetouchFormer leverages priors to detect imperfections and then perform selective self-attention computation, while RestoreFormer restores degraded images conditioned on priors. Compared with BPFRe, RetouchFormer has two advantages: First, BPFRe performs image-level transformations by global convolution filters, while RetouchFormer adaptively incorporates useful contents via selective self-attention to synthesize the content in imperfection regions. Second, BPFRe is a two-stage retouching method, while RetouchFormer adopts a simpler structure and achieves better retouching performance.

Proposed Approach

Overview

In this section, We elaborate on the proposed approach for synthesizing high-fidelity retouching images in the semi-supervised setting. Let $\mathbb{X} = \mathbb{X}_a \bigcup \mathbb{X}_u$ denote the set of training data, and we have $\|\mathbb{X}_a\| \ll \|\mathbb{X}_u\|$, where $\mathbb{X}_a = \{(x_a, y_a)\}$ represents the subset of paired raw-retouching images, and the remaining images belong to the subset $\mathbb{X}_u = \{x_u, y_u\}$. Note that the raw image x_u and the retouching image y_u may correspond to different identities, and are not paired. This significantly reduces the requirement of training data collection in most practical scenarios.

We formulate face retouching as a ‘soft inpainting’ task to jointly detect and complete imperfection regions. As shown in Figure 2, we adopt an encoder-transformer-decoder architecture. The encoder E extracts multi-scale features from the input, which are then split into non-overlapping patches as tokens. The transformer T aims to learn a transformation from soft-masked features to attended features, such that the imperfections can be transformed into normal skin in the feature space. Toward this end, a feature dictionary is learnt to capture clean face priors, and the imperfections are approximately predicted by measuring the difference before and after the dictionary-based reconstruction. More importantly, we propose a selective self-attention mechanism to substitute the standard one in each transformer block, through which the interactions between tokens associated with imperfection regions are suppressed. A decoder G synthesizes a clean face image from the attended features. The components $\{E, T, G\}$ compete with a discriminator D in an adversarial training scheme to ensure that the synthesized retouching images can match the statistics of the real ones.

RetouchFormer

Clean Face Priors We aim to learn clean face priors in terms of a feature dictionary, which offers rich details for data reconstruction in the encoder feature space. Specifically, we perform Vector Quantization (VQ) (van den Oord, Vinyals, and Kavukcuoglu 2017) by learning the dictionary $\mathcal{P} = \{\rho_1, \rho_2, \ldots, \rho_K\}$, such that the feature vector at the location (i, j) of $E(y_-)$ can be approximated as follows:

$$\rho_*|_{(i,j)} = \arg\min_{\rho_k \in \mathcal{P}} \|E(y_-)|_{(i,j)} - \rho_k\|_2^2. \tag{1}$$

where $y_- \in \{y_a, y_u\}$, and ρ_* denotes the nearest elements that we search for in $\mathcal{P}$. Let $\widetilde{E}_\rho(y_-)$ denote the quantized feature map at each pixel location, and the decoder learns to reconstruct the image y_- from it.

Soft Mask Generation How to effectively generate a mask to indicate the imperfection regions is important for improving face retouching performance. We consider that a raw image cannot be well reconstructed by the clean face prior, and the difference before and after reconstruction provides a clue as to the imperfections. In order to fully use the partially paired training data, we adopt a U-Net with a two-stream structure to generate a mask for our 'soft inpainting' task. The raw image $x_- \in \{x_a, x_u\}$ and the reconstructed one $\widetilde{x}_- = R(\widetilde{E}_\rho(x_-))$ from quantized feature map by using a decoder R are concatenated and passed through the streams denoted by $\overline{H}$ and H, and the output mask has the same spatial resolution of x_-.

For the paired data, the retouching regions can be computed directly, and the output of $\overline{H}$ is evaluated as follows:

$$L_{\overline{H}} = \mathbb{E}_{(x_a,y_a)}\|\overline{H}(x_a, \widetilde{x}_a) - \phi(x_a, y_a)\|_2^2, \tag{2}$$

where the function $\phi(\cdot,\cdot)$ measures the image difference at each pixel location, such as the Euclidean distance. For both paired and unpaired data, the output of H is expected to be confident, and the corresponding loss is defined as follows:

$$\begin{aligned} L_H =& \mathbb{E}_{(x_a,y_a)}[-I_{\{\phi(x_a,y_a)>0\}} \log \sigma(H(x_a, \widetilde{x}_a))] \\ &+ \lambda \mathbb{E}_{x_u}[-\sigma(H(x_u, \widetilde{x}_u)) \log \sigma(H(x_u, \widetilde{x}_u))], \end{aligned} \tag{3}$$

where the function $I_{\{\cdot\}}$ returns 1 if the input is true otherwise 0, $\sigma(\cdot)$ denotes an activation function to map an input into the range $(0, 1)$, e.g., Sigmoid, and λ is a weighting factor for controlling the impact of unpaired training data. H aims to enforce the imperfection prediction to align the view of $\overline{H}$. Therefore, we stop the gradient of $\overline{H}$, and update it by using the Exponential Moving Average (EMA) (Tarvainen and Valpola 2017) as follows:

$$\theta_{\overline{H}} \leftarrow \mu\theta_{\overline{H}} + (1 - \mu)\theta_H, \tag{4}$$

where $\theta_- \in \{\theta_{\overline{H}}, \theta_H\}$ denotes the parameters of the two streams, and μ represents a momentum coefficient to control the effect of moving average. The gradients from the loss functions $L_{\overline{H}}$ in Eq.(2) and L_H in Eq.(3) are propagated back to update H. Compared to H, $\overline{H}$ typically produces more reliable predictions for the raw images x_u from X_u, and the resulting imperfection masks are fed into the selective self-attention-based transformer blocks.

Selective Self-attention The self-attention mechanism used in most of the previous transformer-based method tends to globally attend to contents in terms of the intermediate features of the input image, which will be unsuitable for the face retouching task requiring imperfection removal, since these regions typically have higher similarities with each other. Our transformer T is designed to search suitable contents via a selective self-attention mechanism. We first extract spatial patches from the feature maps of the input image at multiple intermediate layers of the encoder E. The multi-scale patch-based image representation denoted by f_E can effectively address the imperfections of different scales. All the patches are further reshaped into 512-dim vectors. The selective self-attention computation focuses on the imperfection regions indicated by the weighting map $f_{\overline{H}}(x_-)$, and the queries Q, keys K, and values V are formulated as follows:

$$\begin{aligned} Q &= W_q(f_E(x_-) \otimes f_{\overline{H}}(x_-)) + b_q, \\ K &= W_k(f_E(x_-) \otimes (1 - f_{\overline{H}}(x_-))) + b_k, \\ V &= W_v(f_E(x_-) \otimes (1 - f_{\overline{H}}(x_-))) + b_v, \end{aligned} \tag{5}$$

where $W_{q/k/v}$ and $b_{q/k/v}$ denote learnable model parameters, $f_{\overline{H}}(\cdot)$ is learnt from the mask $\overline{H}(\cdot)$ via a learnable convolutional layer, and $\otimes$ is the Hadamard product. The tokens associated with the imperfections serve as queries, while the those playing the parts of keys and values are selected by $1 - f_{\overline{H}}(x_-)$. The attended feature map is obtained as follows:

$$\Delta_{f_E} = \mathtt{softmax}(QK^T/\sqrt{\Lambda})V, \tag{6}$$

where Λ denotes the channel number of the features. The relevant tokens associated with the normal skin regions are selected and transformed. Δ_{f_E} is regarded as residual, and is added with f_E before feeding the result into the subsequent transformer block as follows:

$$\widehat{f}_E(x_-) = f_E(x_-) \otimes (1 - f_{\overline{H}}(x_-)) + \Delta_{f_E} \otimes f_{\overline{H}}(x_-). \tag{7}$$

The above equation indicates that our selective self-attention mechanism aims for suppressing the information from the imperfection regions and replacing it with that from the normal skin regions. Finally, we piece all tokens together, and the resulting feature maps are passed through the decoder G to synthesize high-quality retouching images.

Model Training

We now provide the details of the training loss functions of RetouchFormer, which involve three aspects: the prior-based reconstruction, imperfection prediction, and retouching evaluation.

To capture high-quality clean face priors, the encoder E and the dictionary $\mathcal{P}$ are optimized to reconstruct the features of the retouching images $y_- \in \{y_a, y_u\}$ as follows:

$$L_{prior} = \mathbb{E}_{y_-}\Big[\sum_{(i,j)} \|E(y_-)|_{(i,j)} - \rho_*|_{(i,j)}\|_2^2\Big]. \tag{8}$$

We adopt the VQ algorithm (van den Oord, Vinyals, and Kavukcuoglu 2017) to iteratively optimize the encoder and dictionary. To ensure high-quality reconstruction, we apply the following consistency loss to the synthesized image:

$$L_{con}^{rec} = \mathbb{E}_{y_-}[\eta\|y_- - \widetilde{y}_-\|_1 + \|V(y_-) - V(\widetilde{y}_-)\|_2^2], \tag{9}$$

where $\widetilde{y}_- = R(\widetilde{E}_\rho(y_-))$, η is a weighting factor, and $\mathcal{V}(\cdot)$ denotes the features extracted from a pre-trained VGG-19.

Let $\widehat{y}_-^x = G(\widehat{f}_E(x_-))$ denote the retouching image synthesized by RetouchFormer from the raw image x_-. For the paired training data, the results are evaluated by measuring the degree of consistency with the ground truth as follows:

$$L_{con}^{x2y} = \mathbb{E}_{(x_a,y_a)}[\zeta\|y_a - \widehat{y}_a^x\|_1 + \|V(y_a) - V(\widehat{y}_a^x)\|_2^2], \tag{10}$$

where ζ denotes a weighting factor. High-fidelity content synthesis can also benefit from adversarial training, and we thus adopt the discrimination loss formulated as follows:

$$\begin{aligned} L_{adv}^{syn} &= \mathbb{E}_{x_-}[\log(1 - D(\widehat{y}_-^x, \overline{H}(x_-)))], \\ L_{adv}^{disc} &= \mathbb{E}_{y_-}[\log D(y_-, \overline{H}(y_-))] \\ &\quad + \mathbb{E}_{x_-}[\log(1 - D(\widehat{y}_-^x, \overline{H}(x_-))) \\ &\quad + \log(1 - D(x_-, \overline{H}(x_-)))], \end{aligned} \tag{11}$$

where $D(\cdot,\cdot)$ represents the predicted probability of an input image being real. The imperfection region may be very small, and generic image-level discriminators tend to be deceived in this case, since the pixels on the outside are unchanged. To address this issue, the raw image x_- is also fed into D as fake examples. Further, our discriminator performs pixel-level real-fake identification and can pay more attention by injecting the features of $\overline{H}(\cdot)$ into multiple intermediate layers of the discriminator.

We integrate the above three aspects, and express the optimization formulation of the proposed approach as follows:

$$\begin{aligned} &\min_{\mathcal{P},R} L_{prior} + L_{con}^{rec}, \\ &\min_{H} L_{\overline{H}} + L_{H}, \\ &\min_{E,T,G} L_{prior} + L_{con}^{rec} + L_{con}^{x2y} + L_{adv}^{syn}, \\ &\max_{D} L_{adv}^{disc}. \end{aligned} \tag{12}$$

Please note that the constituent networks are optimized with different loss terms. Only the encoder E is tasked to both clean face prior learning and image synthesis, and the other constituent networks are exclusive for individual tasks.

Experiments

We perform extensive experiments to assess the effectiveness of the proposed approach in face retouching. In the following, we first introduce the datasets and experimental settings. Next, we perform a comparison between RetouchFormer and previous state-of-the-arts quantitatively and qualitatively. More insights are provided via further analysis of the RetouchFormer's design elements.

Datasets

We perform face retouching experiments on the FFHQR dataset (Shafaei, Little, and Schmidt 2021), in which there are 70k pairs of raw-retouching images and the raw data are from the Flickr-Face-HQ (FFHQ) dataset (Karras, Laine, and Aila 2019). According to the setting of (Shafaei, Little, and Schmidt 2021), the training/validation/test data consists of 56k/7k/7k image pairs, respectively. We further evaluate the RetouchFormer and competing methods on FR-wild, which contains 1,000 in-the-wild face images with different types of facial blemishes. Due to the unavailability of retouching ground truth, the FR-wild images are used for qualitative evaluation.

Experimental Settings

Semi-supervised Settings. The existing image translation methods rarely take the levels of supervision into consideration. To demonstrate the stability of our RetouchFormer due to this factor, we define the proportion of paired training data as $\gamma = \frac{\|\mathbb{X}_a\|}{\|\mathbb{X}\|}$, where the paired training images are randomly sampled from the dataset, and the remaining images are used as the unpaired training data. We perform the experiments of face retouching with γ limited in the range $\{0.01, 0.1, 0.2, 1\}$.

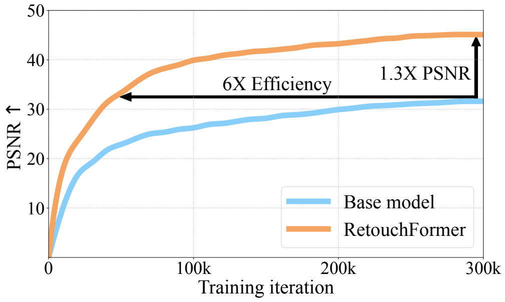

Figure 3: Convergence properties of RetouchFormer and the base model on FFHQR (γ=0.1).

Implementation Details. To perform fair comparison with the competing methods, both training and testing images are resized to 512×512 unless noted otherwise. The feature dictionary $\mathcal{P}$ contains $K = 1024$ elements. The weighting factors: λ in Eq.(3), μ in Eq.(4), η in Eq.(9) and ζ in Eq.(10), are set to 0.5, 0.99, 10 and 10, respectively. During training, the parameters of the proposed model are updated by the Adam optimizer (Kingma and Ba 2015). The learning rate is initially set to $2e^{-4}$ and modified by using a cosine decay schedule. There are a total of 300k training iterations, and each batch contains a single image.

Metrics. To measure the retouching quality, we report the quantitative results by three widely used metrics: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM) and the Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al. 2018), between the synthesized retouching image and the ground truth.

Comparison with State-of-the-arts

We build the base model by disabling soft mask generation, substituting global self-attention for selective self-attention in transformer blocks, and discarding unpaired training data in RetouchFormer. We plot the PSNR scores of the base model and RetouchFormer during training in Figure 3. We find that RetouchFormer is able to converge to a higher PSNR than the base model, and match its best result up to 6 times faster. To demonstrate the advantage of the proposed RetouchFormer on face retouching, we perform a comparison with a wide range of state-of-the-art methods, including CNN-based methods: Pix2PixHD (Isola et al. 2017), MPRNet (Zamir et al. 2021), GPEN (Yang et al. 2021), AutoRetouch (Shafaei, Little, and Schmidt 2021), ABPN (Lei et al. 2022) and BPFRe (Xie et al. 2023), and transformer-based methods: SwinTransformer (Liu et al. 2021) and RestoreFormer (Wang et al. 2022). Pix2PixHD is a generic image translation method serving as baseline. MPRNet, GPEN and RestoreFormer focus on image restoration. AutoRetouch, ABPN and BPFRe is designed for face retouching.

Methods	FFHQR (γ=0.01)			FFHQR (γ=0.1)			FFHQR (γ=0.2)			FFHQR (γ=1)		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Pix2PixHD	25.59	0.7711	0.1585	27.13	0.8008	0.1427	28.88	0.8526	0.1054	29.38	0.9181	0.0766
GPEN	42.70	0.9872	0.0311	42.98	0.9895	0.0169	43.04	0.9901	0.0143	43.12	0.9903	0.0141
SwinTransformer	41.92	0.9840	0.0353	42.29	0.9851	0.0235	42.53	0.9863	0.0199	43.19	0.9878	0.0130
AutoRetouch	38.49	0.9728	0.0161	41.11	0.9791	0.0140	42.22	0.9801	0.0135	44.18	0.9804	0.0133
MPRNet	42.12	0.9874	0.0311	43.29	0.9901	0.0144	43.52	0.9901	0.0137	44.35	0.9907	0.0129
RestoreFormer	39.87	0.9791	0.0178	42.47	0.9879	0.0155	42.86	0.9900	0.0132	42.95	0.9904	0.0129
ABPN	42.09	0.9862	0.0329	43.28	0.9895	0.0234	43.66	0.9903	0.0121	44.41	0.9918	0.0169
BPFRe	43.73	0.9889	0.0127	44.57	0.9901	0.0106	45.06	0.9906	0.0110	45.29	0.9935	0.0092
RetouchFormer	**44.44**	**0.9891**	**0.0116**	**45.13**	**0.9905**	**0.0093**	**45.43**	**0.9913**	**0.0088**	**45.72**	**0.9936**	**0.0078**

Table 1: Quantitative comparison with competing methods on FFHQR.

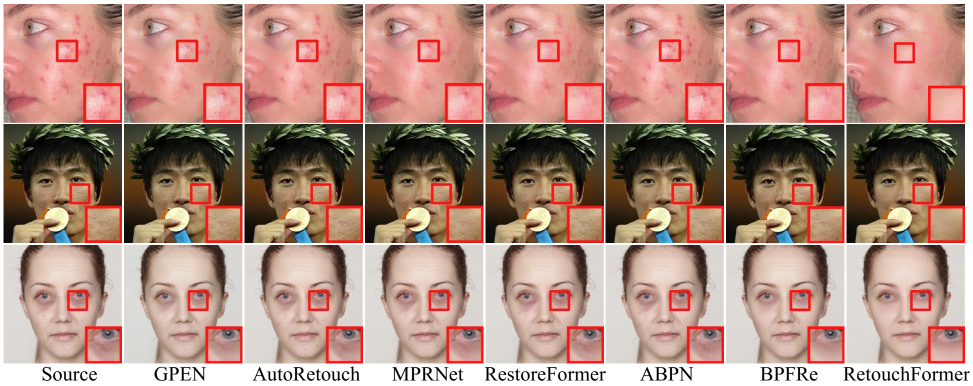

Figure 4: Visual comparison with competing methods on FR-wild images.

Quantitative Results on FFHQR We begin by performing a quantitative comparison between RetouchFormer and the state-of-the-art methods due to the availability of the groundtruth for face retouching on FFHQR. We evaluate the methods by computing the average PSNR, SSIM and LPIPS scores between the synthesized images and the ground truth, and the results are summarized in Table 1. Pix2PixHD provides a lower bound for this task. On FFHQR (γ=0.1), MPRNet, GPEN and ABPN achieve similar performance. BPFRe outperforms all the other comparing motheds while our RetouchFormer achieves a higher PSNR score by 0.56 dB. Related to the transformer-based methods, we observe the superiority of RetouchFormer over RestoreFormer/SwinTransformer with a significant PSNR gain of 2.66/2.84 dB. RetouchFormer can also achieve better results than the competing methods in terms of SSIM and LPIPS. In particular, the proposed approach achieves the lowest LPIPS score of 0.0093, which is lower than the second best method BPFRe by a significant improvement of about 12 percentage points.

Qualitative Results on FR-wild To highlight the superiority of RetouchFormer, we further perform visual comparison with the competing methods. In particular, we evaluate the generalization capability of the methods on the real-world data from FR-wild. Note that both RetouchFormer and other models are trained on FFHQR without seeing any FR-wild images in the training process. All the retouching results are produced by using the trained models. Figure 4 shows a number of retouching images synthesized from the ones with moderate-to-severe acne, pockmarks and dark cicles. One can find that BPFRe and RestoreFormer produce slightly better results than the other existing methods. However, they fail to remove the blemishes at large scales. In accordance with the significant quantitative improvements, the retouching images synthesized by our RetouchFormer are significantly pleasant. Figure 5 shows additional high-quality retouching results of RetouchFormer in removing acne, erasing dark circles and smoothing skin.

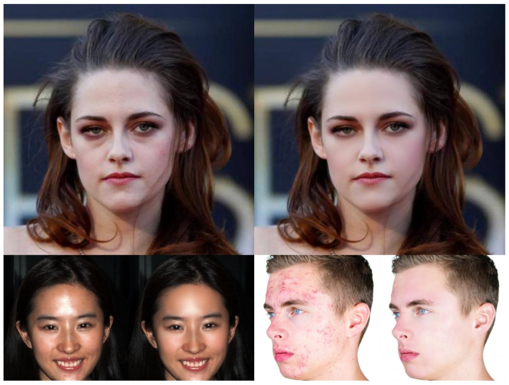

Figure 5: Representative high-quality retouching results of RetouchFormer on FR-wild.

Analysis of Design Elements

According to the quantitative and qualitative comparisons above, we consider that the superior performance of RetouchFormer is due to more accurate imperfection prediction, and adoption of multi-scale tokens, selective self-attention and unpaired data regularization. In this subsection, we ablate the important design elements, and report the results of ablative models in Tables 2-3 and Figures 6-7.

Does the clean face prior make sense? We learn a feature dictionary $\mathcal{P}$ on the retouching images $\{y_a, y_u\}$ for face imperfection prediction. To verify the effectiveness of the prior, we disable $\mathcal{P}$ and obtain a variant 'w/o $\mathcal{P}$', in which the network H generates an imperfection mask from the raw image only. The results shown in Table 2 suggest that $\mathcal{P}$ brings about 17.9% gains in terms of *Soft*-IoU. As to the visual results in Figure 6, we can observe that $\mathcal{P}$ is helpful for discovering more imperfections.

Are the multi-scale tokens important? We adopt the transformer T to learn feature transformations over the spatial patches extracted from the feature maps at multiple intermediate layers of the encoder E. To verify the effectiveness of multi-scale tokens, we build a variant 'w/ *SPS*' by using the encoder output with *Single Patch Size (SPS)*. Table 3 shows that *SPS* leads to a performance drop of 4.93 dB in terms of PSNR. The visual results in Figure 7 confirm that multi-scale tokens allow us to handle the imperfections with different scales.

Are the selective self-attention meaningful? We build the third variant by substituting *Global Self-Attention (GSA)* for selective self-attention, and the resulting model is referred to as 'w/ *GSA*'. Without explicit imperfection prediction and suppression, we can observe that the performance drops significantly. In particular, the PSNR of the variant is 44.07 dB, which is worse than that of RetouchFormer by 1.06 dB. The variant fails to neutralize pockmarks and erase dark circles under the eyes in Figure 7. This indicates that RetouchFormer can utilize the spatial information of imperfections and synthesizes the content from the contextual features of normal skin.

Method	PSNR↑	SSIM↑	*Soft*-IoU↑
w/o $\mathcal{P}$	12.35	0.7970	0.2833
RetouchFormer	**13.21**	**0.8052**	**0.3341**

Table 2: Quantitative results of RetouchFormer with and without the priors $\mathcal{P}$ in soft mask generation.

Method	PSNR↑	SSIM↑	LPIPS↓
w/ *SPS*	40.20	0.9764	0.0423
w/ *GSA*	44.07	0.9882	0.0237
w/o $\mathbb{X}_u$	44.42	0.9891	0.0162
RetouchFormer	**45.13**	**0.9901**	**0.0093**

Table 3: Results of RetouchFormer and ablative models on FFHQR (γ=0.1).

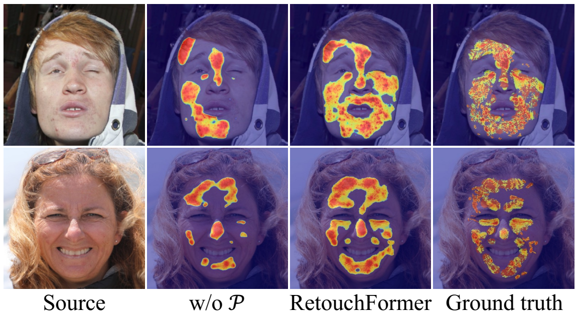

Figure 6: The soft masks generated by RetouchFormer with and without the prior $\mathcal{P}$.

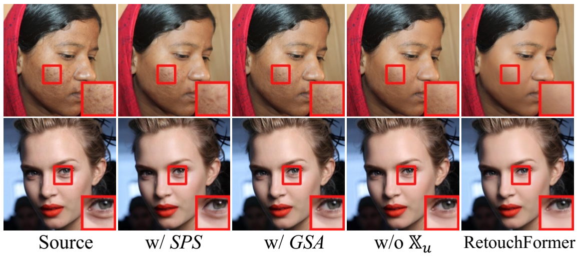

Figure 7: Representative retouching images synthesized by RetouchFormer and ablative models on FR-wild images.

Are the unpaired training data helpful? We also analyze the effect of the unpaired training data $\mathbb{X}_u$. The results in Table 3 suggest that the variant 'w/o $\mathbb{X}_u$' underperforms the full model by about 74% in terms of LPIPS. We consider that both $\mathcal{P}$ and the discriminator D can benefit from $\mathbb{X}_u$, and face retouching is in turn improved. With a decrease in the amount of paired data, the task of face retouching becomes highly challenging, and $\mathbb{X}_u$ is important for stabilizing the performance of RetouchFormer.

Conclusion

This work focuses on face retouching with a transformer-based approach. We formulate face retouching as 'soft inpainting'. RetouchFormer learn clean face priors in terms of a feature dictionary for face imperfection prediction, and further explore prior-based selective self-attention to suppress face imperfections. Furthermore, unpaired training data are utilized to stabilize the performance of RetouchFormer in semi-supervised settings. Extensive comparisons demonstrate the superior capability of RetouchFormer.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Project No. 62072189, 62106136), 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, 2022A1515010434).

References

Abdal, R.; Zhu, P.; Mitra, N. J.; and Wonka, P. 2021. StyleFlow: attribute-conditioned exploration of StyleGAN-Generated images using conditional continuous normalizing flows. *ACM Transactions on Graphics*, 40(3): 1–21.

Arakawa, K. 2004. Nonlinear digital filters for beautifying facial images in multimedia systems. In *Proc. IEEE International Symposium on Curcuits and Systems*.

Batool, N.; and Chellappa, R. 2014. Detection and inpainting of facial wrinkles using texture orientation fields and Markov random field modeling. *IEEE Transactions on Image Processing*, 23(9): 3773–3788.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *Proc. European Conference on Computer Vision*.

Chen, H.; Wang, Y.; Guo, T.; Xu, C.; Deng, Y.; Liu, Z.; Ma, S.; Xu, C.; Xu, C.; and Gao, W. 2021. Pre-trained image processing transformer. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*.

Collins, E.; Bala, R.; Price, B.; and Susstrunk, S. 2020. Editing in style: uncovering the local semantics of GANs. In *Proc. IEEE conference on Computer Vision and Pattern Recognition*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. In *arXiv:1801.04805*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An image is worth 16x16 words: transformers for image recognition at scale. In *International Conference on Learning Representations*.

Goetschalckx, L.; Andonian, A.; Oliva, A.; and Isola, P. 2019. GANalyze: toward visual definitions of congnitive image properties. In *Proc. International Conference on Computer Vision*.

Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *Proc. Neural Information Processing Systems*.

Gu, J.; Shen, Y.; and Zhou, B. 2020. Image processing using multi-code GAN prior. In *Proceedings of IEEE conference on computer vision and pattern recognition*, 3012–3021.

Hu, X.; Huang, Q.; Shi, Z.; Li, S.; Gao, C.; Sun, L.; and Li, Q. 2022. Style transformer for image inversion and editing. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*.

Isola, P.; Zhu, J.-Y.; Zhou, T.; and Efros, A. A. 2017. Image-to-image translation with conditional adversarial networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*.

Jahanian, A.; Chai, L.; and Isola, P. 2019. On the "steerability" of generative adversarial networks. In *arXiv:1907.07171*.

Karras, T.; Laine, S.; and Aila, T. 2019. A style-based generator architecture for generative adversarial networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*.

Kim, T.; Cha, M.; Kim, H.; Lee, J. K.; and Kim, J. 2017. Learning to discover cross-domain relations with generative adversarial networks. In *Proc. International Conference on Machine Learning*.

Kingma, D. P.; and Ba, J. L. 2015. Adam: a method for stochastic optimization. In *Proc. International Conference on Learning Representation*.

Lei, B.; Guo, X.; Yang, H.; Cui, M.; Xie, X.; and Huang, D. 2022. ABPN: Adaptive Blend Pyramid Network for Real-Time Local Retouching of Ultra High-Resolution Photo. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*.

Leyvand, T.; Cohen-Or, D.; Dror, G.; and Lischinski, D. 2008. Data-driven enhancement of facial attractiveness. In *Proc. ACM Conference on Special Interest Group on Computer Graphics and Interactive Techniques*.

Ling, H.; Kreis, K.; Li, D.; Kim, S. W.; Torralba, A.; and Fidler, S. 2021. EditGAN: high-precision semantic image editing. In *Proc. Neural Information Processing Systems*.

Lipowezky, U.; and Cahen, S. 2008. Automatic freckles detection and retouching. In *Proc. IEEE Convention of Electrical and Electronics Engineers in Israel*.

Liu, M.-Y.; Breuel, T.; and Kautz, J. 2017. Unsupervised image-to-image translation networks. In *Proc. Neural Information Processing Systems*.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: hierarchical vision transformer using shifted windows. In *Proc. International Conference on Computer Vision*.

Patashnik, O.; Wu, Z.; Shechtman, E.; Cohen-Or, D.; and Lischinski, D. 2021. StyleCLIP: text-driven manipulation of StyleGAN imagery. In *Proc. International Conference on Computer Vision*.

Richardson, E.; Alaluf, Y.; Patashnik, O.; Nitzan, Y.; Azar, Y.; Shapiro, S.; and Cohen-Or, D. 2021. Encoding in style: a StyleGan encoder for image-to-image translation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2287–2296.

Shafaei, A.; Little, J. J.; and Schmidt, M. 2021. AutoRetouch: automatic professional face retouching. In *Proc. IEEE Winter Conference on Applications of Computer Vision*.

Shen, Y.; and Zhou, B. 2021. Closed-form factorization of latent semantics in GANs. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*.

Tarvainen, A.; and Valpola, H. 2017. Mean teachers are better role models: weight-averaged consistency targets improve semi-supervised deep learning results. In *Proc. Advances in Neural Information Processing Systems*.

Tov, O.; Alaluf, Y.; Nitzan, Y.; Patashnik, O.; and Cohen-Or, D. 2021. Designing an encoder for StyleGan image manipulation. *ACM Transactions on Graphics*, 40(4): 1–14.

van den Oord, A.; Vinyals, O.; and Kavukcuoglu, K. 2017. Neural discrete representation learning. In *Proc. Neural Information Processing Systems*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is all you need. In *Proc. Neural Information Processing Systems*.

Velusamy, S.; Parihar, R.; Kini, R.; and Rege, A. 2020. FabSoften: face beautification via dynamic skin smoothing, guided feathering and texture restoration. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshop*.

Wang, Z.; Zhang, J.; Chen, R.; Wang, W.; and Luo, P. 2022. RestoreFormer: High-quality blind face restoration from undegraded key-value pairs. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*.

Wu, H.; Xiao, B.; Codella, N.; Liu, M.; Dai, X.; Yuan, L.; and Zhang, L. 2021. CvR: Introducing convolutions to vision transformers. *arXiv:2103.15808*.

Wu, P.-W.; Lin, Y.-J.; Chang, C.-H.; Chang, E. Y.; and Liao, S.-W. 2019. RelGAN: multi-domain image-to-image translation via relative attributes. In *Proc. International Conference on Computer Vision*.

Wu, Z.; Lischinski, D.; and Shechtman, E. 2021. StyleSpace analysis: disentangled controls for StyleGAN image generation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*.

Xie, L.; Xue, W.; Xu, Z.; Wu, S.; Yu, Z.; and Wong, H. S. 2023. Blemish-Aware and Progressive Face Retouching With Limited Paired Data. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*.

Yang, T.; Ren, P.; Xie, X.; and Zhang, L. 2021. GAN prior embedded network for blind face restoration in the wild. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*.

Zamir, S. W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F. S.; Yang, M.-H.; and Shao, L. 2021. Multi-stage progressive image restoration. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proc. International Conference on Computer Vision*.

Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; and Dai, J. 2021. Deformable DETR: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations*.