Old Photo Restoration via Deep Latent Space Translation

Ziyu Wan, Bo Zhang, Dongdong Chen, Pan Zhang, Dong Chen, Jing Liao, Fang Wen

Abstract—We propose to restore old photos that suffer from severe degradation through a deep learning approach. Unlike conventional restoration tasks that can be solved through supervised learning, the degradation in real photos is complex and the domain gap between synthetic images and real old photos makes the network fail to generalize. Therefore, we propose a novel triplet domain translation network by leveraging real photos along with massive synthetic image pairs. Specifically, we train two variational autoencoders (VAEs) to respectively transform old photos and clean photos into two latent spaces. And the translation between these two latent spaces is learned with synthetic paired data. This translation generalizes well to real photos because the domain gap is closed in the compact latent space. Besides, to address multiple degradations mixed in one old photo, we design a global branch with a partial nonlocal block targeting to the structured defects, such as scratches and dust spots, and a local branch targeting to the unstructured defects, such as noises and blurriness. Two branches are fused in the latent space, leading to improved capability to restore old photos from multiple defects. Furthermore, we apply another face refinement network to recover fine details of faces in the old photos, thus ultimately generating photos with enhanced perceptual quality. With comprehensive experiments, the proposed pipeline demonstrates superior performance over state-of-the-art methods as well as existing commercial tools in terms of visual quality for old photos restoration.

Index Terms—Image Restoration, Image Generation, Latent Space Translation, Mixed degradation

1 Introduction

PHOTOS are taken to freeze the happy moments that otherwise gone. Even though time goes by, one can still evoke memories of the past by viewing them. Nonetheless, old photo prints deteriorate when kept in poor environmental condition, which causes the valuable photo content permanently damaged. Fortunately, as mobile cameras and scanners become more accessible, people can now digitalize the photos and invite a skilled specialist for restoration. However, manual retouching is usually laborious and time-consuming, which leaves piles of old photos impossible to get restored. Hence, it is appealing to design automatic algorithms that can instantly repair old photos for those who wish to bring old photos back to life.

Prior to the deep learning era, there are some attempts [1], [2], [3], [4] that restore photos by automatically detecting the localized defects such as scratches and blemishes, and filling in the damaged areas with inpainting techniques. Yet these methods focus on completing the missing content and none of them can repair the spatially-uniform defects such as film grain, sepia effect, color fading, etc., so the photos after restoration still appear outdated compared to modern photographic images. With the emergence of deep learning, one can address a variety of low-level image restoration problems [5], [6], [7], [8], [9], [10], [11] by exploiting the powerful representation capability of convolutional neural networks, *i.e.*, learning the mapping for a specific task from a large amount of synthetic images.

The same framework, however, does not apply to old photo restoration and the reason is three-fold. First, the degradation process of old photos is rather complex, and there exists no degradation model that can realistically render the old photo artifact. Therefore, the model learned from those synthetic data generalizes poorly on real photos. Second, old photos are plagued with a compound of degradation and inherently require different strategies for repair: unstructured defects that are spatially homogeneous, *e.g.*, film grain and color fading, should be restored by utilizing the pixels in the neighborhood, whereas the structured defects, *e.g.*, scratches, dust spots, etc., should be repaired with a global image context. Furthermore, people are fastidious to tiny artifacts around faces yet a network trained on general natural images cannot capture facial intrinsic characteristics. Thus, a network targeting for face retouching is needed especially considering portraits account for large proportion of old photos.

To circumvent these issues, we formulate the old photo restoration as a triplet domain translation problem. Different from previous image translation methods [12], we leverage data from three domains (*i.e.*, real old photos, synthetic images and the corresponding ground truth), and the translation is performed in latent space. Synthetic images and the real photos are first transformed to the same latent space

- *Z. Wan is with the Department of Computer Science, City University of Hong Kong, Hong Kong SAR, China. E-mail: ziyuwan2-c@my.cityu.edu.hk*
- *B. Zhang, D. Chen and F. Wen are with the Visual Computing Group, Microsoft Research, Beijing, China. E-mail: {Tony.Zhang, doch, fangwen}@microsoft.com*
- *D. Chen is with Microsoft Cloud+AI, Redmond, Washington, USA. E-mail: cddlyf@gmail.com*
- *P. Zhang is with the Department of Automation, University of Science and Technology of China, Hefei, Anhui, China. E-mail: zhangpan@mail.ustc.edu.cn*
- *J. Liao is with the Department of Computer Science, City University of Hong Kong, Hong Kong SAR, China. E-mail: jingliao@cityu.edu.hk*
- *J. Liao is the corresponding author.*

Manuscript received April 19, 2005; revised August 26, 2015.

Fig. 1: **Old photo restoration results produced by our method.** Our method can handle the complex degradation mixed by both unstructured and structured defects in real old photos. In particular, we recover high-frequency details for face regions, further improving the perceptual quality for portraits. For each image pair, left is the input while the retouched output is shown on the right.

with a shared variational autoencoder [13] (VAE). Meanwhile, another VAE is trained to project ground truth clean images into the corresponding latent space. The mapping between the two latent spaces is then learned with the synthetic image pairs, which restores the corrupted images to clean ones. The advantage of the latent restoration is that the learned latent restoration can generalize well to real photos because of the domain alignment within the first VAE. Besides, we differentiate the mixed degradation and propose a partial nonlocal block that considers the long-range dependencies of latent features to specifically address the structured defects during the latent translation. Finally, considering that faces are the most important visual stimuli, we propose a post-processing step with a coarse-to-fine generator to reconstruct high-resolution faces with hierarchical spatial adaptive conditions. Some results are shown in Figure 1. In comparison with several leading restoration methods, we prove the effectiveness of our approach in restoring multiple degradations of real photos.

2 Related Work

Single degradation image restoration. Existing image degradation can be roughly categorized into two groups: unstructured degradation such as noise, blurriness, color fading, and low resolution, and structured degradation such as holes, scratches, and spots. For the former unstructured ones, traditional works often impose different image priors, including nonlocal self-similarity [14], [15], [16], sparsity [17], [18], [19], [20] and local smoothness [21], [22], [23]. Recently, a lot of deep learning based methods have also been proposed for different image degradation, like image denoising [5], [6], [24], [25], [26], [27], [28], super-resolution [7], [29], [30], [31], [32], and deblurring [8], [33], [34], [35].

Compared to unstructured degradation, structured degradation is more challenging and often modeled as the "image painting" problem. Thanks to powerful semantic modeling ability, most existing best-performed inpainting methods are learning based. For example, Liu *et al*. [36] masked out the hole regions within the convolution operator and enforces the network focus on non-hole features only. To get better inpainting results, many other methods consider both local patch statistics and global structures. Specifically, Yu *et al*. [37] and Liu *et al*. [38] proposed to employ an attention layer to utilize the remote context. And the appearance flow is explicitly estimated by Ren *et al*. [39] so that textures in the hole regions can be directly synthesized based on the corresponding patches.

No matter for unstructured or structured degradation, though the above learning-based methods can achieve remarkable results, they are all trained on the synthetic data. Therefore, their performance on the real dataset highly relies on synthetic data quality. For real old images, since they are often seriously degraded by a mixture of unknown degradation, the underlying degradation process is much more difficult to be accurately characterized. In other words, the network trained on synthetic data only, will suffer from the domain gap problem and perform badly on real old photos. In this paper, we model real old photo restoration as a new triplet domain translation problem and some new techniques are adopted to minimize the domain gap.

Mixed degradation image restoration. In the real world, a corrupted image may suffer from complicated defects mixed with scratches, loss of resolution, color fading, and film noises. However, research solving mixed degradation is much less explored. The pioneer work RL-Restore [40] proposed a toolbox that comprises multiple light-weight networks, and each of them responsible for a specific degradation. Then they learn a controller that dynamically selects the operator from the toolbox. Inspired by RL-Restore [40], Suganum *et al*. [41] performs different convolutional operations in parallel and uses the attention mechanism to select the most suitable combination of operations. However, these methods still rely on supervised learning from synthetic data and hence cannot generalize to real photos. Besides, they only focus on unstructured defects and do not support structured defects like image inpainting. On the other hand, DIP [42] found that the deep neural network inherently

resonates with low-level image statistics and thereby can be utilized as an image prior for blind image restoration without external training data. This method has the potential, though not claimed in DIP [42], to restore in-the-wild images corrupted by mixed factors. In comparison, our approach excels in both restoration performance and efficiency.

Face restoration. A variety of methods specifically designed for face restoration have been proposed. Early works [43], [44] attempt to deblur faces by the guidance of an external reference, but an exemplar image with suitable texture for transfer is inconvenient to retrieve and the requirement of an external face database makes it cumbersome for practical usage. On the other hand, most contemporary works [45] rely on generative adversarial network (GAN) to resolve the blurriness and produce realistic result. It is noteworthy that the restoration quality could be boosted by explicitly considering intrinsic facial priors such as face parsing [46], facial landmarks [47], identity prior [48] or 3D morphable models [49]. Nonetheless, these methods require extra networks to perform those auxiliary tasks, which brings robustness issue when processing the face images that suffer from large pose and severe degradations. A recent work [50] utilizes a pre-trained generative model and searches the latent code that conforms to the input. Albeit impressive, the generated faces suffer from fidelity issue. In this work, we aim to restore in-the-wild faces with well-preserved identity while caring for robustness. To this end, we do not rely on face prior and learn the restoration by synthesis: instead of letting the network digest the degraded faces as input, the output is synthesized from a latent noise with the latent features modulated by the degraded faces through spatially-variant de-normalization. We will show that this approach achieves preferable quality in restoring vintage portraits.

Old photo restoration. Old photo restoration is a classical mixed degradation problem, but most existing methods [1], [2], [3], [4] focus on inpainting only. They follow a similar paradigm *i.e.*, defects like scratches and blotches are first identified according to low-level features and then inpainted by borrowing the textures from the vicinity. However, the hand-crafted models and low-level features they used are difficult to detect and fix such defects well. Moreover, none of these methods consider restoring some unstructured defects such as color fading or low resolution together with inpainting. Thus photos still appear old fashioned after restoration. In this work, we reinvestigate this problem by virtue of a data-driven approach, which can restore images from multiple defects simultaneously and turn heavily-damaged old photos to modern style.

3 METHOD

In contrast to conventional image restoration tasks, old photo restoration is more challenging. First, old photos contain far more complex degradation that is hard to be modeled realistically and there always exists a domain gap between synthetic and real photos. As such, the network usually cannot generalize well to real photos by purely learning from synthetic data. Second, the defects of old photos is a compound of multiple degradations, thus essentially requiring different strategies for restoration. Unstructured defects such as film noise, blurriness and color fading, etc. can be restored with spatially homogeneous filters by making use of surrounding pixels within the local patch; structured defects such as scratches and blotches, on the other hand, should be inpainted by considering the global context to ensure the structural consistency. In the following, we first describe our main framework to address the aforementioned *generalization issue* and *mixed degradation issue* respectively. After that, we introduce auxiliary network for face enhancement, so as to further improve the restoration quality.

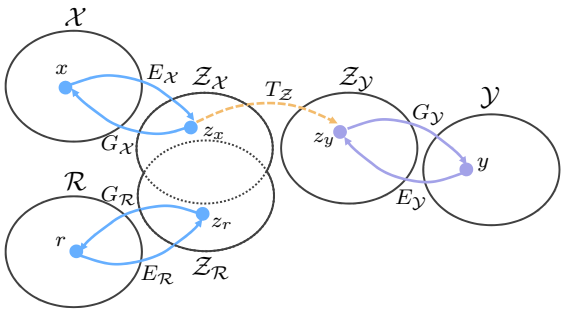

Fig. 2: **Illustration of our translation method with three domains.** The domain gap between $\mathcal{Z}_{\mathcal{X}}$ and $\mathcal{Z}_{\mathcal{R}}$ will be reduced in the shared latent space.

3.1 Restoration via latent space translation

In order to mitigate the domain gap, we formulate the old photo restoration as an image translation problem, where we treat clean images and old photos as images from distinct domains and we wish to learn the mapping in between. However, as opposed to general image translation methods that bridge two different domains [12], [51], we translate images across three domains: the real photo domain $\mathcal{R}$, the synthetic domain $\mathcal{X}$ where images suffer from artificial degradation, and the corresponding ground truth domain $\mathcal{Y}$ that comprises images without degradation. Such triplet domain translation is crucial in our task as it leverages the unlabeled real photos as well as a large amount of synthetic data associated with ground truth.

We denote images from three domains respectively with $r \in \mathcal{R}$, $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, where x and y are paired by data synthesis, *i.e.*, x is degraded from y. Directly learning the mapping from real photos $\{r\}_{i=1}^{N}$ to clean images $\{y\}_{i=1}^{N}$ is hard since they are not paired and thus unsuitable for supervised learning. We thereby propose to decompose the translation with two stages, which are illustrated in Figure 2. First, we propose to map $\mathcal{R}, \mathcal{X}, \mathcal{Y}$ to corresponding latent spaces via $E_{\mathcal{R}} : \mathcal{R} \mapsto \mathcal{Z}_{\mathcal{R}}$, $E_{\mathcal{X}} : \mathcal{X} \mapsto \mathcal{Z}_{\mathcal{X}}$, and $E_{\mathcal{Y}} : \mathcal{Y} \mapsto \mathcal{Z}_{\mathcal{Y}}$, respectively. In particular, because synthetic images and real old photos are both corrupted, sharing similar appearances, we align their latent space into the shared domain by enforcing some constraints. Therefore we have $\mathcal{Z}_{\mathcal{R}} \approx \mathcal{Z}_{\mathcal{X}}$. This aligned latent space encodes features for all the corrupted images, either synthetic or real ones. Then we propose to learn the image restoration in the latent space. Specifically, by utilizing the synthetic data pairs $\{x, y\}_{i=1}^{N}$, we learn the translation from the latent space of corrupted images, $\mathcal{Z}_{\mathcal{X}}$, to the latent space of ground truth, $\mathcal{Z}_{\mathcal{Y}}$, through the mapping $T_{\mathcal{Z}} : \mathcal{Z}_{\mathcal{X}} \mapsto \mathcal{Z}_{\mathcal{Y}}$, where $\mathcal{Z}_{\mathcal{Y}}$ can be further reversed to $\mathcal{Y}$ through generator $G_{\mathcal{Y}} : \mathcal{Z}_{\mathcal{Y}} \mapsto \mathcal{Y}$. By learning

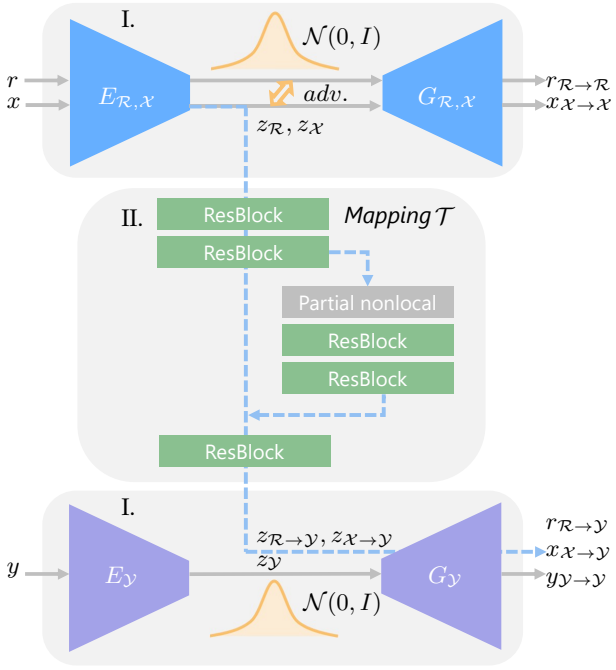

Fig. 3: **Architecture of our restoration network.** (I.) We first train two VAEs: VAE_1 for images in real photos $r \in \mathcal{R}$ and synthetic images $x \in \mathcal{X}$, with their domain gap closed by jointly training an adversarial discriminator; VAE_2 is trained for clean images $y \in \mathcal{Y}$. With VAEs, images are transformed to compact latent space. (II.) Then, we learn the mapping that restores the corrupted images to clean ones in the latent space.

the latent space translation, real old photos r can be restored by sequentially performing the mappings,

$$r_{\mathcal{R}\to\mathcal{Y}} = G_{\mathcal{Y}} \circ T_{\mathcal{Z}} \circ E_{\mathcal{R}}(r). \tag{1}$$

Domain alignment in the VAE latent space One key of our method is to meet the assumption that $\mathcal{R}$ and $\mathcal{X}$ are encoded into the same latent space. To this end, we propose to utilize variational autoencoder [13] (VAE) to encode images with compact representation, whose domain gap is further examined by an adversarial discriminator [52]. We use the network architecture shown in Figure 3 to realize this concept.

In the first stage, two VAEs are learned for the latent representation. Old photos $\{r\}$ and synthetic images $\{x\}$ share the first one termed VAE_1, with the encoder $E_{\mathcal{R},\mathcal{X}}$ and generator $G_{\mathcal{R},\mathcal{X}}$, while the ground true images $\{y\}$ are fed into the second one, VAE_2 with the encoder-generator pair $\{E_{\mathcal{Y}}, G_{\mathcal{Y}}\}$. VAE_1 is shared for both r and x in the aim that images from both corrupted domains can be mapped to a shared latent space. The VAEs assumes Gaussian prior for the distribution of latent codes, so that images can be reconstructed by sampling from the latent space. We use the re-parameterization trick to enable differentiable stochastic sampling [53] and optimize VAE_1 with data $\{r\}$ and $\{x\}$ respectively. The objective with $\{r\}$ is defined as:

$$\begin{aligned}\mathcal{L}_{\text{VAE}_1}(r) = \ & \text{KL}(E_{\mathcal{R},\mathcal{X}}(z_r|r)||\mathcal{N}(0, I)) \\ & + \alpha\mathbb{E}_{z_r\sim E_{\mathcal{R},\mathcal{X}}(z_r|r)}\left[\|G_{\mathcal{R},\mathcal{X}}(r_{\mathcal{R}\to\mathcal{R}}|z_r) - r\|_1\right] \\ & + \mathcal{L}_{\text{VAE}_1,\text{GAN}}(r)\end{aligned} \tag{2}$$

where, $z_r \in \mathcal{Z}_{\mathcal{R}}$ is the latent codes for r, and $r_{\mathcal{R}\to\mathcal{R}}$ is the generation outputs. The first term in equations is the KL-divergence that penalizes deviation of the latent distribution from the Gaussian prior. The second ℓ_1 term lets the VAE reconstruct the inputs, implicitly enforcing latent codes to capture the major information of images. Besides, we introduce the least-square loss (LSGAN) [54], denoted as $\mathcal{L}_{\text{VAE}_1,\text{GAN}}$ in the formula, to address the well-known over-smooth issue in VAEs, further encouraging VAE to reconstruct images with high realism. The objective with $\{x\}$, denoted as $\mathcal{L}_{\text{VAE}_1}(x)$, is defined similarly. And VAE_2 for domain $\mathcal{Y}$ is trained with a similar loss so that the corresponding latent representation $z_y \in \mathcal{Y}$ can be derived.

We use VAE rather than vanilla autoencoder because VAE features denser latent representation due to the KL regularization (which will be proved in ablation study), and this helps produce closer latent space for $\{r\}$ and $\{x\}$ with VAE_1 thus leading to smaller domain gap. To further narrow the domain gap in this reduced space, we propose to use an adversarial network to examine the residual latent gap. Concretely, we train another discriminator $D_{\mathcal{R},\mathcal{X}}$ that differentiates $\mathcal{Z}_{\mathcal{R}}$ and $\mathcal{Z}_{\mathcal{X}}$, whose loss is defined as,

$$\begin{aligned}\mathcal{L}^{\text{latent}}_{\text{VAE}_1,\text{GAN}}(r, x) = \ & \mathbb{E}_{x\sim\mathcal{X}}[D_{\mathcal{R},\mathcal{X}}(E_{\mathcal{R},\mathcal{X}}(x))^2] \\ & + \mathbb{E}_{r\sim\mathcal{R}}[(1 - D_{\mathcal{R},\mathcal{X}}(E_{\mathcal{R},\mathcal{X}}(r)))^2].\end{aligned} \tag{3}$$

Meanwhile, the encoder $E_{\mathcal{R},\mathcal{X}}$ of VAE_1 tries to fool the discriminator with a contradictory loss to ensure that $\mathcal{R}$ and $\mathcal{X}$ are mapped to the same space. Combined with the latent adversarial loss, the total objective function for VAE_1 becomes,

$$\min_{E_{\mathcal{R},\mathcal{X}}, G_{\mathcal{R},\mathcal{X}}} \max_{D_{\mathcal{R},\mathcal{X}}} \mathcal{L}_{\text{VAE}_1}(r) + \mathcal{L}_{\text{VAE}_1}(x) + \mathcal{L}^{\text{latent}}_{\text{VAE}_1,\text{GAN}}(r, x). \tag{4}$$

Restoration through latent mapping With the latent code captured by VAEs, in the second stage, we leverage the synthetic image pairs $\{x, y\}$ and propose to learn the image restoration by mapping their latent space (the mapping network M in Figure 3). The benefit of latent restoration is threefold. First, as $\mathcal{R}$ and $\mathcal{X}$ are aligned into the same latent space, the mapping from $\mathcal{Z}_{\mathcal{X}}$ to $\mathcal{Z}_{\mathcal{Y}}$ will also generalize well to restoring the images in $\mathcal{R}$. Second, the mapping in a compact low-dimensional latent space is in principle much easier to learn than in the high-dimensional image space. In addition, since the two VAEs are trained independently and the reconstruction of the two streams would not be interfered with each other. The generator $G_{\mathcal{Y}}$ can always get an absolutely clean image without degradation given the latent code $z_{\mathcal{Y}}$ mapped from $\mathcal{Z}_{\mathcal{X}}$, whereas degradations will likely remain if we learn the translation in pixel level.

Let $r_{\mathcal{R}\to\mathcal{Y}}$, $x_{\mathcal{X}\to\mathcal{Y}}$ and $y_{\mathcal{Y}\to\mathcal{Y}}$ be the final translation outputs for r, x and y, respectively. At this stage, we solely train the parameters of the latent mapping network $\mathcal{T}$ and fix the two VAEs. The loss function $\mathcal{L}_{\mathcal{T}}$, which is imposed at both the latent space and the end of generator $G_{\mathcal{Y}}$, consists of three terms,

$$\mathcal{L}_{\mathcal{T}}(x, y) = \lambda_1\mathcal{L}_{\mathcal{T},\ell_1} + \mathcal{L}_{\mathcal{T},\text{GAN}} + \lambda_2\mathcal{L}_{\text{FM}} \tag{5}$$

where the latent space loss, $\mathcal{L}_{\mathcal{T},\ell_1} = \mathbb{E}\|\mathcal{T}(z_x) - z_y)\|_1$, penalizes the ℓ_1 distance of the corresponding latent codes. We introduce the adversarial loss $\mathcal{L}_{\mathcal{T},\text{GAN}}$, still in the form

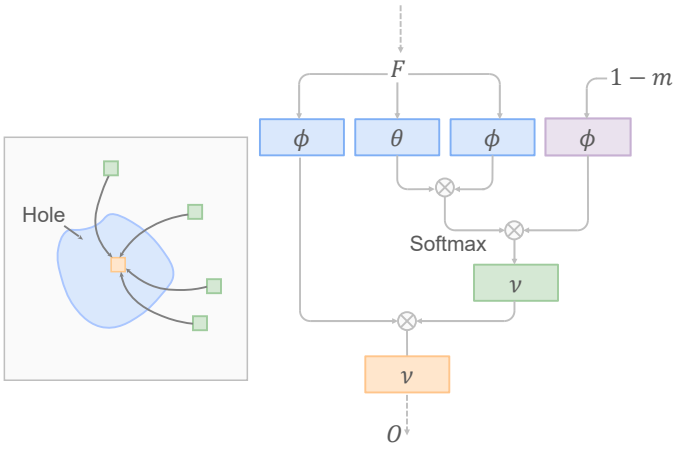

Fig. 4: **Partial nonlocal block**. Left shows the principle. The pixels within the hole areas are inpainted by the context pixels outside the corrupted region. Right shows the detailed implementation.

of LSGAN [54], to encourage the ultimate translated synthetic image $x_{\mathcal{X}\to\mathcal{Y}}$ to look real. Besides, we introduce feature matching loss L_{FM} to stabilize the GAN training. Specifically, L_{FM} matches the multi-level activations of the adversarial network D_M, and that of the pretrained VGG network (also known as perceptual loss in [12], [55]), *i.e.*,

$$\mathcal{L}_{\text{FM}} = \mathbb{E}\Big[\sum_i \frac{1}{n^i_{D_\mathcal{T}}}\|\phi^i_{D_\mathcal{T}}(x_{\mathcal{X}\to\mathcal{Y}}) - \phi^i_{D_\mathcal{T}}(y_{\mathcal{Y}\to\mathcal{Y}})\|_1 + \sum_i \frac{1}{n^i_{\text{VGG}}}\|\phi^i_{\text{VGG}}(x_{\mathcal{X}\to\mathcal{Y}}) - \phi^i_{\text{VGG}}(y_{\mathcal{Y}\to\mathcal{Y}})\|_1\Big], \quad (6)$$

where $\phi^i_{D_\mathcal{T}}$ (ϕ^i_{VGG}) denotes the i^{th} layer feature map of the discriminator (VGG network), and $n^i_{D_\mathcal{T}}$ (n^i_{VGG}) indicates the number of activations in that layer.

3.2 Multiple degradation restoration

The latent restoration using the residual blocks, as described earlier, only concentrates on local features due to the limited receptive field of each layer. Nonetheless, the restoration of structured defects requires plausible inpainting, which has to consider long-range dependencies so as to ensure global structural consistency. Since legacy photos often contain mixed degradations, we have to design a restoration network that simultaneously supports the two mechanisms. Towards this goal, we propose to enhance the latent restoration network by incorporating a global branch as shown in Figure 3, which composes of a nonlocal block [56] that considers global context and several residual blocks in the following. While the original block proposed in [56] is unaware of the corruption area, our nonlocal block explicitly utilizes the mask input so that the pixels in the corrupted region will not be adopted for completing those area. Since the context considered is a part of the feature map, we refer to the module specifically designed for the latent inpainting as a *partial nonlocal block*, which is shown in Figure 4.

Formally, let $F \in \mathbb{R}^{C\times HW}$ be the intermediate feature map in M (C, H and W are number of channels, height and width respectively), and $m \in \{0,1\}^{HW}$ represents the binary mask downscaled to the same size, where 1 represents the defect regions to be inpainted and 0 represents the intact regions. The affinity between i^{th} location and j^{th} location in F, denoted by $s_{i,j} \in \mathbb{R}^{HW\times HW}$, is calculated by the correlation of F_i and F_j modulated by the mask $(1-m_j)$, *i.e.*,

$$s_{i,j} = (1-m_j)f_{i,j} / \sum_{\forall k}(1-m_k)f_{i,k}, \quad (7)$$

where,

$$f_{i,j} = \exp(\theta(F_i)^T \cdot \phi(F_j)) \quad (8)$$

gives the pairwise affinity with embedded Gaussian. Here, θ and ϕ project F to Gaussian space for affinity calculation. According to the affinity $s_{i,j}$ that considers the holes in the mask, the partial nonlocal finally outputs

$$O_i = \nu\left(\sum_{\forall j} s_{i,j}\mu(F_j)\right), \quad (9)$$

which is a weighted average of correlated features for each position. We implement the embedding functions θ, ϕ, μ and ν with 1×1 convolutions.

We design the global branch specifically for inpainting and hope the non-hole regions are left untouched, so we fuse the global branch with the local branch under the guidance of the mask, *i.e.*,

$$F_{fuse} = (1-m)\odot\rho_{\text{local}}(F) + m\odot\rho_{\text{global}}(O), \quad (10)$$

where operator $\odot$ denotes Hadamard product, and ρ_{local} and ρ_{global} denote the nonlinear transformation of residual blocks in two branches. In this way, the two branches constitute the latent restoration network, which is capable to deal with multiple degradation in old photos. We will detail the derivation of the defect mask in Section 4.1.

Table 1 shows the detailed network structure.

3.3 Defect Region Detection

Since the global branch of our restoration network requires a mask m as the guidance, in order to get the mask automatically, we train a scratch detection network in a supervised way by using a mixture of real scratched dataset and synthetic dataset. Specifically, let $\{s_i, y_i | s_i \in \mathcal{S}, y_i \in \mathcal{Y}\}$ denote the whole training pairs, where s_i and y_i are the scratched image and the corresponding binary scratch mask respectively, we use the cross-entropy loss to minimize the difference between the predicted mask $\hat{y}_i$ and y_i,

$$\mathcal{L}_{CE} = \mathbb{E}_{(s_i,y_i)\sim(\mathcal{S},\mathcal{Y})}\Big\{\alpha\sum_{h=1}^{H}\sum_{w=1}^{W} -y_i^{(h,w)}\log\hat{y}_i^{(h,w)} - (1-\alpha)\sum_{h=1}^{H}\sum_{w=1}^{W}(1-y_i^{(h,w)})\log(1-\hat{y}_i^{(h,w)})\Big\}. \quad (11)$$

Since the scratch regions are often a small portion of the whole image, here we use a weight α_i to remedy the imbalance of positive and negative pixel samples. To determine the detailed value of α_i, we compute the positive/negative proportion of y_i on the fly,

$$\alpha_i = \frac{[y_i = 1]}{[y_i = 1] + [y_i = 0]]}. \quad (12)$$

Module	Layer	Kernel size / stride	Output size
Encoder E	Conv	$7 \times 7/1$	$256 \times 256 \times 64$
	Conv	$4 \times 4/2$	$128 \times 128 \times 64$
	Conv	$4 \times 4/2$	$64 \times 64 \times 64$
	ResBlock×4	$3 \times 3/1$	$64 \times 64 \times 64$
Generator G	ResBlock×4	$3 \times 3/1$	$64 \times 64 \times 64$
	Deconv	$4 \times 4/2$	$128 \times 128 \times 64$
	Deconv	$4 \times 4/2$	$256 \times 256 \times 64$
	Conv	$7 \times 7/1$	$256 \times 256 \times 3$
Mapping $\mathcal{T}$	Conv	$3 \times 3/1$	$64 \times 64 \times 128$
	Conv	$3 \times 3/1$	$64 \times 64 \times 256$
	Conv	$3 \times 3/1$	$64 \times 64 \times 512$
	Partial nonlocal	$1 \times 1/1$	$64 \times 64 \times 512$
	Resblock×2	$3 \times 3/1$	$64 \times 64 \times 512$
	ResBlock×6	$3 \times 3/1$	$64 \times 64 \times 512$
	Conv	$3 \times 3/1$	$64 \times 64 \times 256$
	Conv	$3 \times 3/1$	$64 \times 64 \times 128$
	Conv	$3 \times 3/1$	$64 \times 64 \times 64$

TABLE 1: **Detailed network structure.** The modules in the global branch of the mapping network are highlighted in gray.

Besides, we also introduce the focal loss to focus on the hard samples,

$$\mathcal{L}_{FL} = \mathbb{E}_{(s_i,y_i)\sim(\mathcal{S},\mathcal{Y})}\left\{\sum_{h=1}^{H}\sum_{w=1}^{W} -(1-p_i^{(h,w)})^{\gamma}\log p_i^{(h,w)}\right\}, \tag{13}$$

where,

$$p_i^{(h,w)} = \begin{cases} \hat{y_i}^{(h,w)} & \text{if } y_i^{(h,w)} = 1 \\ 1-\hat{y_i}^{(h,w)} & \text{otherwise} \end{cases} \tag{14}$$

Therefore, the whole detection objective is

$$\mathcal{L}_{Seg} = \mathcal{L}_{CE} + \beta\mathcal{L}_{FL}. \tag{15}$$

By default, we set the parameters in Equations (13) and (15) with $\gamma = 0.2$ and $\beta = 10$. And the detection network adopts U-Net architecture which reuses low-level features through extensive skip connection.

3.4 Face Enhancement

The restoration network proposed above is general to all kinds of old photos. However, considering restoration quality on faces is most sensitive to human perception, we further propose a network for face enhancement. Given one real degraded photo r, we hope to reconstruct degraded faces r_f in r into a much detailed and clean version with proposed face enhancement network G_f. The classical pixel-wise translation method could not solve such a blind restoration problem well because the degradation prior is totally unknown. Here, we solve this problem from the perspective of generative models.

As shown in Figure 5, we employ a coarse-to-fine generator to translate a low-dimensional code z into corresponding high-resolution and clean faces, where z is a down-sampled patch of r_f (8×8 in our implementation). At the same time of progressive generation, r_f will be injected into the generator in each scale with a spatial adaptive manner [57], which could capture the style and structure information of degraded faces as much as possible. Specifically, let $h \in \mathbb{R}^{H\times W\times C}$ be the activation map of previous layer and r_f^i be the condition of current scale i. h will be modulated as follows,

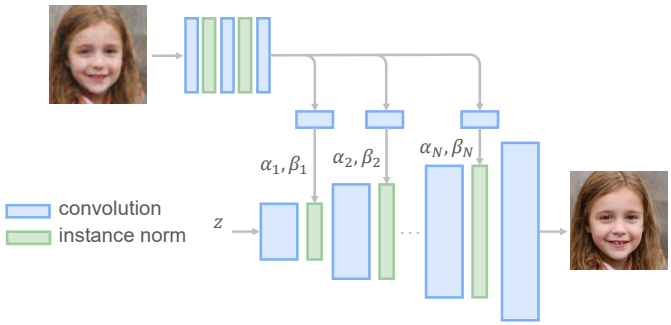

Fig. 5: **The progressive generator network of face enhancement.** Starting from a latent vector z, the network up-samples the feature map by deconvolution progressively. The degraded face will be injected into different resolutions in a spatial condition manner.

$$\gamma_{x,y,c}(r_f^i)\frac{h_{x,y,c}-\mu_c}{\sqrt{\sigma_c^2+\epsilon}} + \beta_{x,y,c}(r_f^i), \tag{16}$$

where $h_{x,y,c}$ denotes each element of h, $x \in H$ and $y \in W$ span spatial dimensions and $c \in C$ is the feature channel. μ_c and σ_c are the mean and standard deviation of the activation h in channel c. ϵ is a constant factor to avoid outlier values, $\gamma_{x,y,c}(r_f^i)$ and $\beta_{x,y,c}(r_f^i)$ are two learnable scalars locally controlling the influence from r_f^i. In practice, we use two convolutional layers to generate these two coefficients at each element location.

To train the proposed face enhancement network, we penalize the perceptual distance between the generated face $G_f(z, r_f)$ and high-resolution r_c as follows,

$$\mathcal{L}_{\text{perc}}^{\text{face}} = \mathbb{E}\left[\sum_i \frac{1}{n_{\text{VGG}}^i}\|\phi_{\text{VGG}}^i(G_f(z,r_f)) - \phi_{\text{VGG}}^i(r_c)\|_1\right], \tag{17}$$

where r_f is the degraded face of r_c and z is the latent code of r_f. Besides, another adversarial loss is involved in the training procedure to ensure the synthesis of high-frequency details,

$$\begin{aligned}\mathcal{L}_{\text{GAN}}^{\text{face}}(z, r_f, r_c) = \mathbb{E}_{z\sim\mathcal{Z}, r_f\sim\mathcal{R}_f}[D_f(G_f(z,r_f))^2] \\ + \mathbb{E}_{r_c\sim\mathcal{R}_c}[(1-D_f(r_c))^2].\end{aligned} \tag{18}$$

The face enhance network is jointly trained with previous restoration network to ensure better generalization ability, *i.e.*, r_f is the output of triplet domain translation network. We found such training scheme could effectively suppress the generated artifacts. More detailed analysis about joint training could be found in Section 4.4.3. During inference, we firstly search the face parts of arbitrary photos, and then refine this region with proposed enhancement network. As a result of generative model, there sometimes exists color shifting between reconstructed faces and input

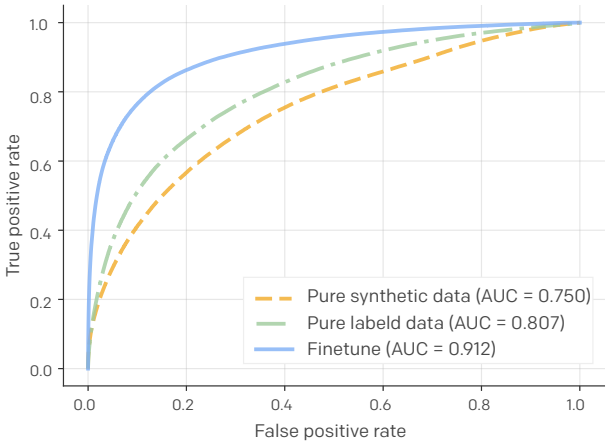

Fig. 6: **ROC curve for scratch detection of different data settings.** Combining both synthetic structured degradations and a small amount of labeled data, the scratch detection network could achieve great results.

degraded faces. We solve this issue by histogram matching. Finally the reconstructed face will be combined with original input photo using linear blending to produce the final results.

4 Experiment

4.1 Implementation

Training Dataset We synthesize old photos using images from the Pascal VOC dataset [58]. In Section 4.2, we introduce how to render realistic defects. Besides, we collect 5,718 old photos to form the images old photo dataset. To train the face enhancement network, we use 50,000 aligned high-resolution face images from FFHQ [59].

Training details We adopt Adam solver [62] with $\beta_1 = 0.5$ and $\beta_2 = 0.999$. The learning rate is set to 0.0002 for the first 100 epochs, with linear decay to zero thereafter. During training, we randomly crop images to 256×256. In all the experiments, we empirically set the parameters in Equations (2) and (5) with $\alpha = 10$, $\lambda_1 = 60$ and $\lambda_2 = 10$ respectively.

4.2 Data Generation

Next, we brief the old photo synthesis procedure. Though we cannot fully emulate the old photo style, a careful synthesis is vital to high-quality restoration as support overlap between two domain distributions eases domain adaptation [63].

Unstructured Degradation We use the following operations to simulate the unstructured degradation. Specifically,

1) Gaussian white noise with $\sigma \in (5, 50)$.
2) Gaussian blur with kernel size $k \subset \{3, 5, 7\}$ and standard deviation $\sigma \in (1.0, 5.0)$;
3) JPEG compression whose quality level in the range of $(40, 100)$;
4) Color jitter which randomly shifts the RGB color channels by $(-20, 20)$;
5) Box blur to mimic the lens defocus.

We apply above types of augmentations with varying parameters in random order. To achieve more variations, we stochastically drop out each type of operation with 30% probability. Still, the synthesis cannot exactly match the appearance of real photo defects, thus requiring the proposed network to further reduce the domain gap.

Structured Degradation As described in Section 3.3, to train the defect region detection network, a mixture of synthetic and real scratch datasets are used (pretrain on synthetic and finetune on real). For the synthetic part, we collect 62 scratch texture images and 55 paper texture images, which are further augmented with elastic distortions. Then we use layer addition, lighten-only and screen modes with random level of opacity to blend the scratch textures over the natural images from the Pascal VOC dataset [58]. Besides, in order to simulate large-area photo damage, we generate holes with feathering and random shape where the underneath paper texture is unveiled. Note that we also introduce film grain noise and blur with random kernel to simulate the global defects at this stage so that the synthetic data has a similar global style as the real old photos. These injected noises are beneficial in that they make the distribution of synthetic and real data become more overlapped. Examples of synthesized scratched old photos are shown in Figure 8.

To improve the detection performance on real old photos, we collect 783 real old photos and manually annotate the local defects, among which 400 images are used for training and remaining for testing.As shown in Figure 6, adding the real data into training can significantly boost the scratch detection performance on real old photos and achieve AUC as 0.912. Some sampled scratch detection masks and restoration results of test dataset are shown in Figure 9.

4.3 Comparisons

Baselines We compare our method against state-of-the-art approaches. For fair comparison, we train all the methods with the same training dataset (Pascal VOC) and test them on the corrupted images synthesized from DIV2K dataset [64] and the test set of our old photo dataset. The following methods are included for comparison.

- Operation-wise attention [41] performs multiple operations in parallel and uses an attention mechanism to select the proper branch for mixed degradation restoration. It learns from synthetic image pairs with supervised learning.
- Deep image prior [42] learns the image restoration given a single degraded image, and has been proven powerful in denoising, super-resolution and blind inpainting.
- Pix2Pix [65] is a supervised image translation method, which leverages synthetic image pairs to learn the translation in image level.
- CycleGAN [51] is a well-known unsupervised image translation method that learns the translation using unpaired images from distinct domains.
- The last baseline is to sequentially perform BM3D [66], a classical denoising method, and EdgeConnect [67], a state-of-the-art inpainting method, to restore the unstructured and structured defects respectively.

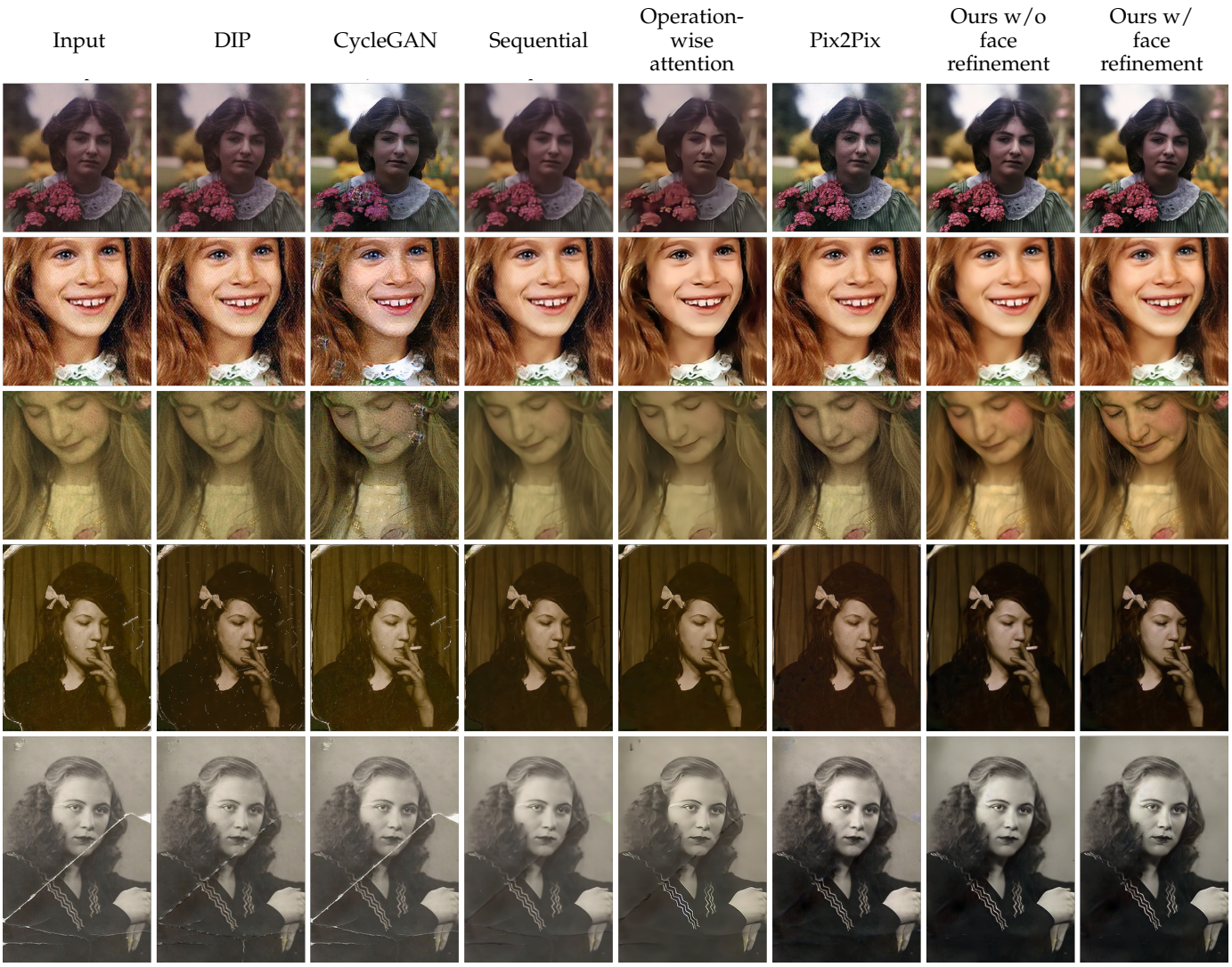

Fig. 7: **Qualitative comparison against state-of-the-art methods.** It shows that our method can restore both unstructured and structured degradation and our recovered results are significantly better than other methods.

Fig. 8: **Some examples of synthetic photos with scratches.**

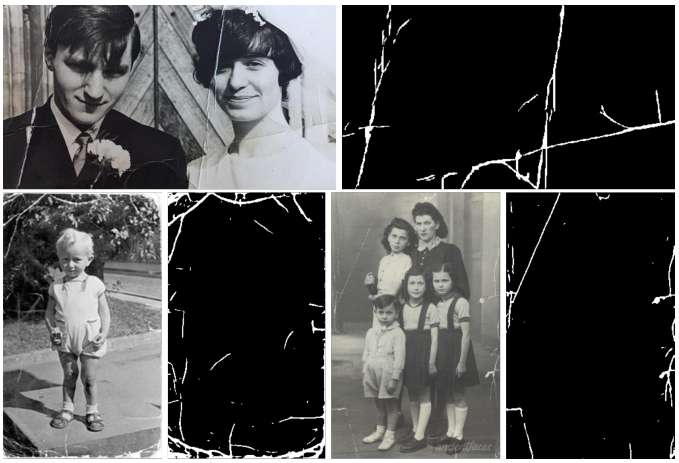

Fig. 9: **Some defect region detection results on real photos.**

Quantitative comparison We test different models on the synthetic images from DIV2K dataset and adopt four metrics for comparison. Table 2 gives the quantitative results. The peak signal-to-noise ratio (PSNR) and the structural similarity index (SSIM) are used to compare the low-level differences between the restored output and the ground truth. The operational-wise attention method unsurprisingly achieves the best PSNR/SSIM score since this method directly optimizes the pixel-level ℓ_1 loss. Our method ranks second-best in terms of PSNR/SSIM. However, these

Method	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓
Input	12.92	0.49	0.59	306.80
Attention [41]	**24.12**	**0.70**	0.33	208.11
DIP [42]	22.59	0.57	0.54	194.55
Pix2pix [65]	22.18	0.62	**0.23**	**135.14**
Sequential [66], [67]	22.71	0.60	0.49	191.98
Ours w/o partial nonlocal	23.14	0.68	0.26	143.62
Ours w/ partial nonlocal	**23.33**	**0.69**	**0.25**	**134.35**

TABLE 2: **Quantitative results on the DIV2K dataset.** Upward arrow (↑) indicate that a higher score denotes a good image quality. We highlight the best two scores for each measure.

two metrics characterizing low-level discrepancy, usually do not correlate well with human judgment, especially for complex unknown distortions [68]. Therefore, we also adopt the recently learned perceptual image patch similarity (LPIPS) [68] metric which calculates the distance of multi-level activations of a pretrained network and is deemed to better correlate with human perception. This time, Pix2pix and our method give the best scores with a negligible difference. The operation-wise attention method, however, shows inferior performance under this metric, demonstrating it does not yield good perceptual quality. Besides, we adopt Fréchet Inception Distance (FID) [69] which is a widely used metric for assessing the quality of generative models. Specifically, the FID score calculates the distance between the feature distributions of the final outputs and the real images. Still, our method and Pix2pix rank the best, while our method shows a slight quantitative advantage. In all, our method is comparable to the leading methods on synthetic data.

Qualitative comparison To prove the generalization to real old photos, we conduct experiments on the real photo dataset. For a fair comparison, we retrain the CycleGAN to translate real photos to clean images. Since we lack the restoration ground truth for real photos, we cannot apply reference-based metrics for evaluation. Therefore, we qualitatively compare the results, which are shown in Figure 7. The DIP method can restore mixed degradations to some extent. However, there is a trade off between the defect restoration and the structural preservation: more defects reveal after a long training time while fewer iterations induce the loss of fine structures. CycleGAN, learned from unpaired images, tends to focus on restoring unstructured defects and neglect to restore all the scratch regions. Both the operation-wise attention method and the sequential operations give comparable visual quality. However, they cannot amend the defects that are not covered in the synthetic data, such as sepia issue and color fading. Besides, the structured defects still remain problematic, possibly because they cannot handle the old photo textures that are subtly different from the synthetic dataset. Pix2pix, which is comparable to our approach on synthetic images, however, is visually inferior to our method. Some film noises and structured defects still remain in the final output. This is due to the domain gap between synthetic images and real photos, which makes the method fail to generalize. In comparison, our method gives clean, sharp images with the scratches plausibly filled with unnoticeable artifacts. Besides successfully addressing the artifacts considered in data synthesis, our method can also enhance the photo color appropriately. In general, our method gives the most visually pleasant results and the photos after restoration appear like modern photographic images.

Method	Top 1	Top 2	Top 3	Top 4	Top 5
DIP [42]	2.54	8.49	19.26	39.09	74.22
CycleGAN [51]	4.24	8.21	19.54	28.32	50.42
Sequential [66], [67]	4.81	18.13	47.87	79.60	94.61
Attention [41]	6.79	21.24	49.85	73.08	88.38
Pix2Pix [65]	16.14	60.90	73.65	86.68	94.90
Ours	**65.43**	**83.00**	**89.80**	**93.20**	**97.45**

TABLE 3: **User study results.** The percentage (%) of each method being selected as the top K ($K = 1 - 5$) by users.

User study To better illustrate the subjective quality, we conduct a user study to compare with other methods. We randomly select 21 old photos from the test set and let users sort the results according to the restoration quality. We collect subjective opinions from 24 users and count the percentage of each method being selected as the top K ($K = 1 - 5$). The results are shown in Table 3, which clearly demonstrates the advantage of our approach, with 65.43% chances to be selected as the top 1.

Comparison with Commercial Software Some commercial software and applications like Meitu [70] and Remini Photo Enhancer [71] start to provide the service of automatic old photos restoration. To demonstrate the effectiveness of our pipeline, we also compare the restoration performance with them in Figure 10. Based on the observation of their outputs, it could be found that their methods ignore the existing structured degradations and color fading. By contrast, our method alleviates these problems and generates more visual-pleasant results like the first row and third row of Figure 10. Meanwhile, the proposed latent domain translation network better deals with real-world local defects such as noise because of a smaller domain gap in the second row of Figure 10. Finally, with a dedicated face enhancement network, the refined human face also contains more details. Overall, our method could achieve more clean, sharp and vibrant results compared with commercial counterparts.

4.4 Analysis

In order to prove the effectiveness of individual technical contributions, we perform the following ablation study.

4.4.1 Latent translation with VAEs

Let us consider the following variants, with proposed components added one-by-one:

- Pix2Pix which learns the translation in image-level. The model is trained using only synthetic pairs.
- Two VAEs with an additional KL loss to penalize the latent space. The VAEs and latent mapping are all trained simultaneously.

Fig. 10: **Qualitative comparisons against commercial tools.** Remini Photo Enhancer [71], Meitu [70] and our full pipeline results are included.

- VAEs with two-stage training (VAEs-TS): the two VAEs are first trained separately and the latent mapping is learned thereafter with the two fixed VAEs, which ensure the translation is performed in two fix domains.
- Full model, which also adopts latent adversarial loss.

We first calculate the Wasserstein distance [72] between the latent space of old photos and synthetic images. Table 4 shows that distribution distance gradually reduces after adding each component. The main reason is that the KL loss of VAEs could lead to a more compact latent space. Training with the two-stage manner isolates the two VAEs, and the latent adversarial loss further closes the domain gap. A smaller domain gap will improve the model generalization to real photo restoration. To demonstrate this point, several visual comparison results (without any face post-processing) are provided in Figure 11. We observe that Pix2Pix could not handle these blind distortions well. The restoration is gradually improved with a more compact latent space. Besides, we also adopt a blind image quality assessment metric, BRISQUE [73], to measure photo quality after restoration. The BRISQUE score in Table 4 progressively improves by applying these mentioned techniques, which is consistent with corresponding visual results.

4.4.2 *Partial nonlocal block*

We propose the partial nonlocal block to make the triplet domain translation network support the restoration of structured degradations. As shown in Figure 12, both Pix2Pix [12]

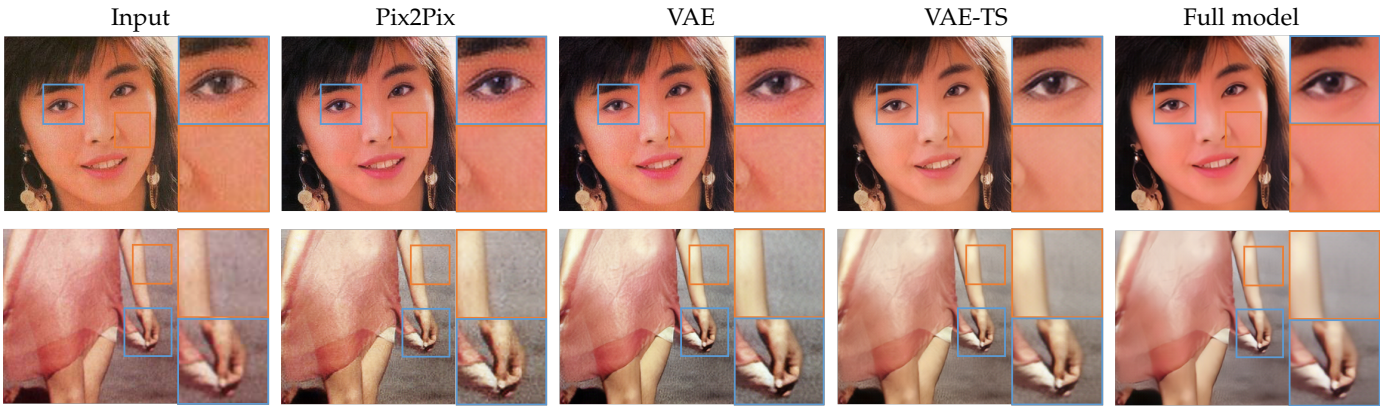

Fig. 11: **Ablation study for latent translation with VAEs.** By involving feature translation and feature-level adversarial loss, the domain gap between synthetic degradations and real-world defects could be narrowed gradually, leading to better restoration results step by step.

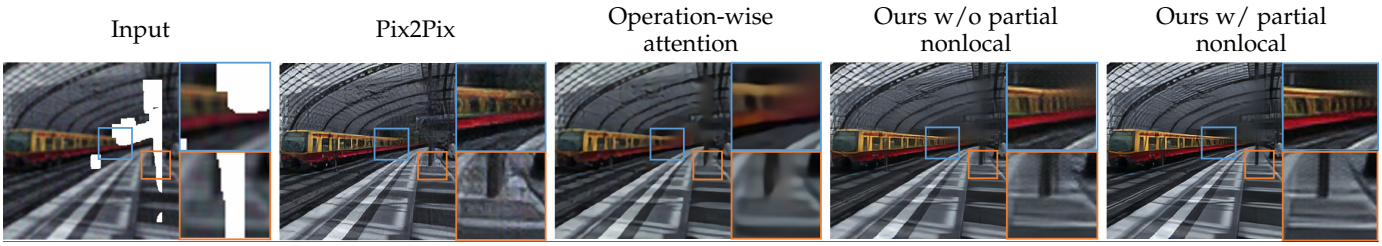

Fig. 12: **Ablation study of partial nonlocal block.** Partial nonlocal better inpaints the structured defects.

Method	Pix2Pix	VAEs	VAEs-TS	full model
Wasserstein ↓	1.837	1.048	0.765	**0.581**
BRISQUE ↓	25.549	23.949	23.396	**23.016**

TABLE 4: **Ablation study of latent translation with VAEs.** We provide some quantitative comparisons here to demonstrate the superior performance of the full model. Our method achieve best results on both distribution distance and BRISQUE metric.

Fig. 13: **Ablation study of partial nonlocal block.** Partial nonlocal does not touch the non-hole regions as this operation is aware of the corruption area.

and mixed-distortion restoration method [41] could not simultaneously handle local and global defects well. Because of the utilization of large image context (partial nonlocal), the scratches can be inpainted with fewer visual artifacts and better globally consistent restoration can be achieved in our method. In addition, we find that the partial nonlocal block could also ensure that the inpainting is only applied in the localized defect areas. In Figure 13, the background of origin photos will be modified if we remove this block. Besides, the quantitative result in Table 2 also shows that the partial nonlocal block consistently improves the restoration performance on all the metrics.

4.4.3 Ablation Study of Face Enhancement Network

Effectiveness of Joint Training The face enhancement network is jointly trained with the triplet domain translation network, i.e., input corrupted faces will first pass through this translation network, and then be reconstructed into a high-resolution version with the proposed enhancement network. To demonstrate the effectiveness of this training scheme, we provide some qualitative results of real old photos in Figure 14. We could observe that without joint training, unnatural redundant textures and artifacts are visible in the generated faces. One reason may be there exist some distribution bias between generated faces of the first stage and real degraded faces. By introducing the joint training, this kind of gap could be alleviated, leading more pleasant and stable results.

Comparison with Other Generative Model Generally, a straightforward question could arise here: if a simple pixel-wise translation model like [45] could obtain desired results? To verify this point, we train another Pix2Pix [12] model using the same loss function and discriminator. There are mainly two differences between Pix2Pix [12] and our enhancement network: 1) We adopt the progressive generator with the spatial condition rather than the image-level concatenation of Pix2Pix [12]. 2) The model parameter amount of Pix2Pix [12] is about 188.9M, which is almost twice as

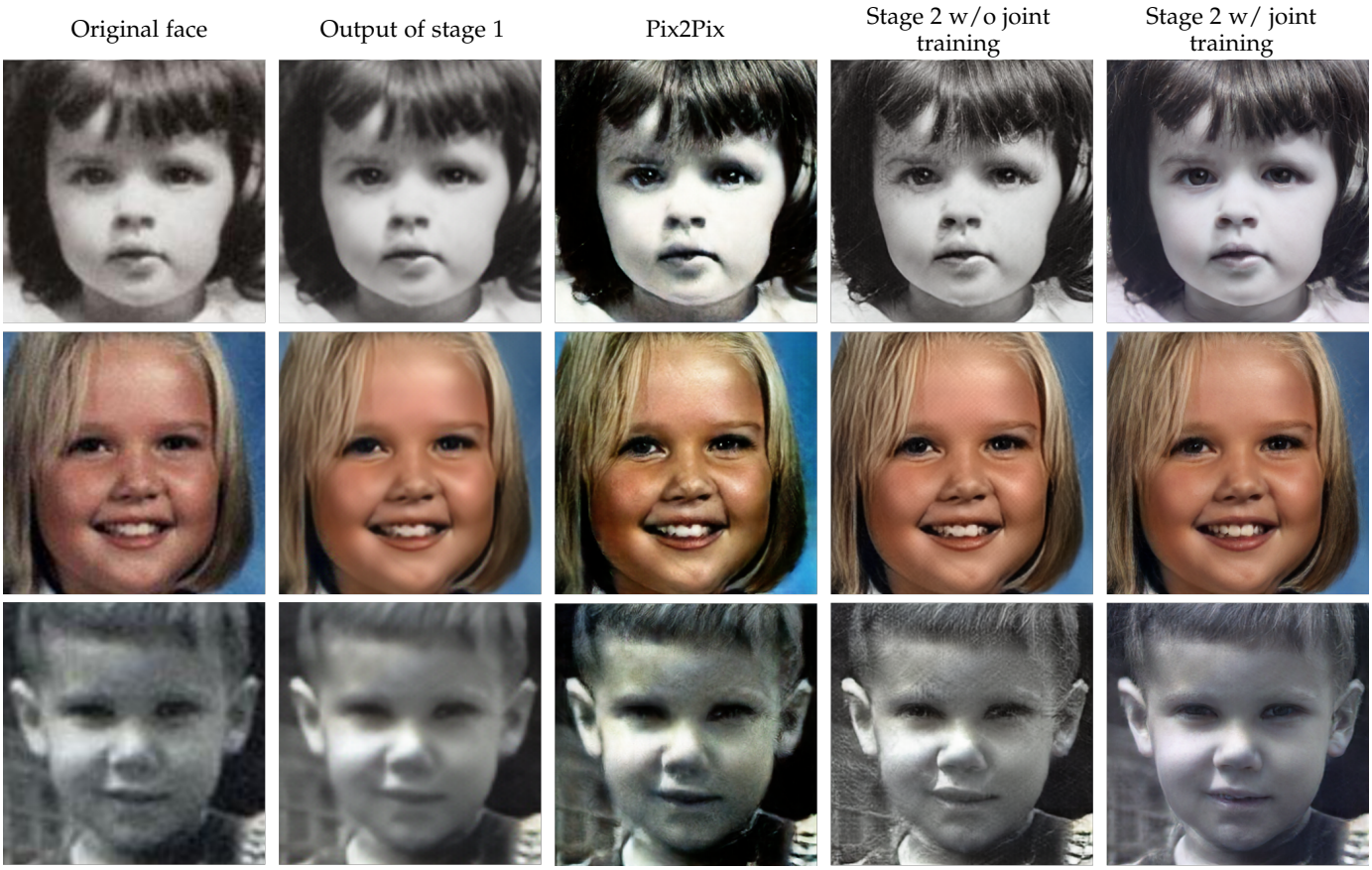

Fig. 14: **Comparisons with Pix2Pix [12] and w/o joint training.** To ensure fair comparison, the Pix2Pix model is also trained jointly with the domain translation network. Stage 1: Triplet domain translation. Stage 2: Face enhancement.

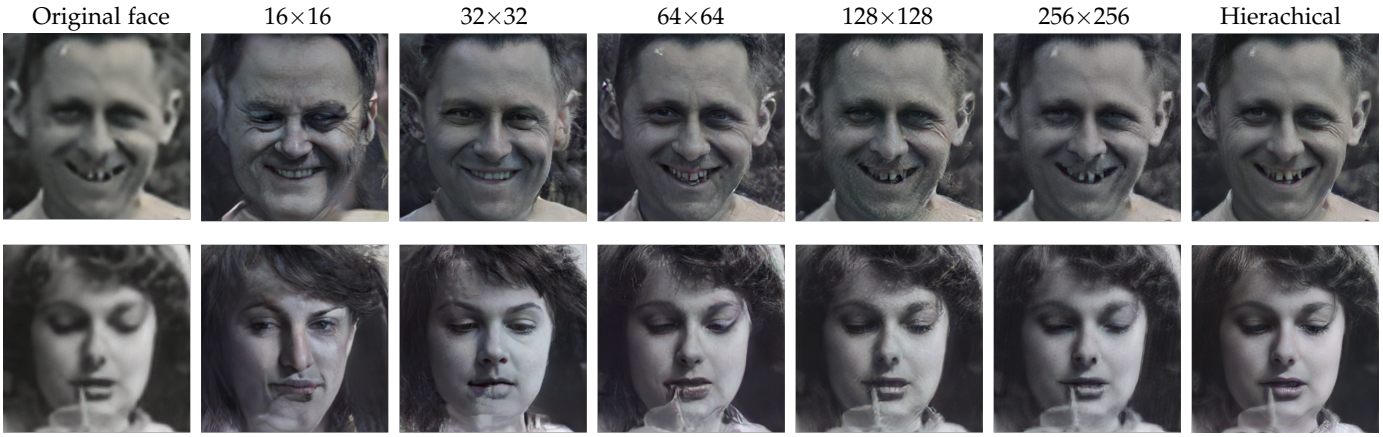

Fig. 15: **Face reconstruction results with different injection methods.** The hierarchical manner leads to the best results.

	Input	16 x 16	32 x 32	64 x 64	128 x 128	256 x 256	Hierarchical
PSNR ↑	22.918	17.677	20.931	23.088	24.622	24.938	**25.282**
SSIM ↑	0.655	0.545	0.618	0.677	0.724	0.740	**0.743**
FID ↓	42.421	24.177	17.993	15.768	14.236	15.653	**13.175**
LPIPS ↓	0.376	0.271	0.193	0.150	0.129	0.133	**0.120**

TABLE 5: **Quantitative comparisons for different injection positions.** We test the results on synthetic degraded face images.

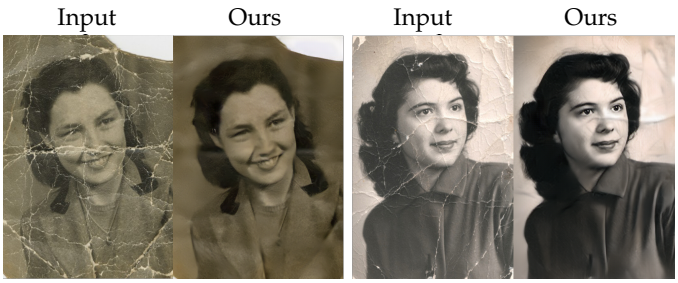

Fig. 16: **Limitation.** Our method cannot handle complex shading artifacts due to the uneven lighting.

much as ours (92.1M). The visual comparisons could be found in Figure 14. Obviously, the results of [12] are far below our expectations, which contains many artifacts and color degradations. By contrast, the reconstructed faces of our method are more vivid. The potential reason for this phenomenon is that the face post-processing of [45] is to replenish slight details of faces, but our target here is to reconstruct the face based on the corrupted observation, which is a more challenging task. By the feature-level spatial modulation, the generator learns to reconstruct a clean face while capturing the original structure and style information.

Effectiveness of Hierarchical Spatial Injection To reconstruct a high-resolution face from real photos meanwhile maintaining underlying structure and style information, we propose to modulate the features of the coarse-to-fine generator in a hierarchical spatial condition manner. To demonstrate the importance of this point, we compare this method with the single spatial injection of different layers, i.e., from the lowest scale (16×16) to the highest (256×256) one. Qualitatively, as shown in Figure 15, although we could generate a more vivid face at the lowest scale, the identity is not preserved since a low-dimensional condition could not constrain the generator well. With the increase of injection resolution, the reconstructed face becomes more accurate gradually. However, we find that the generated faces contain lots of noise and artifacts when the injection is performed at the highest scale only. The reason may be that the position of highest scale injection is too close to the generator output and less relevant with the semantic feature in previous layers, thus resulting in the incomplete modulation. By contrast, our hierarchical spatial injection achieves natural restoration results with the right structures and styles, as shown in the last column of Figure 15. To further prove this point, we also calculate the quantitative statistics of each scale on a synthetic dataset. We randomly select 2,000 test images and add varying degradations to construct paired data. As shown in Table 5, although scale 16×16 and 32×32 achieve better performance on FID and LPIPS compared with input which demonstrates the distribution of generated face become close to real HR faces, the PSNR and SSIM are even lower than the input because of the loss on original information. By introducing the method of hierarchical injection, our enhancement network obtains the best scores on all four metrics.

5 Discussion and Conclusion

We propose a novel triplet domain translation network that opens new avenue to restore the mixed degradation for in-the-wild old photos. The domain gap is reduced between old photos and synthetic images, and the translation to clean images is learned in latent space. Our method suffers less from generalization issue compared with prior methods. Besides, we propose a partial nonlocal block which restores the latent features by leveraging the global context, so the scratches can be inpainted with better structural consistency. Furthermore, we propose a coarse-to-fine generator with spatial adaptive condition to reconstruct the face regions of old photos. Our method demonstrates good performance in restoring severely degraded old photos. However, our method cannot handle complex shading as shown in Figure 16. This is because our dataset contains few old photos with such defects. One could possibly address this limitation using our framework by explicitly considering the shading effects during synthesis or adding more such photos as training data.

References

[1] F. Stanco, G. Ramponi, and A. De Polo, "Towards the automated restoration of old photographic prints: a survey," in *The IEEE Region 8 EUROCON 2003. Computer as a Tool.*, vol. 2. IEEE, 2003, pp. 370–374.

[2] V. Bruni and D. Vitulano, "A generalized model for scratch detection," *IEEE transactions on image processing*, vol. 13, no. 1, pp. 44–50, 2004.

[3] R.-C. Chang, Y.-L. Sie, S.-M. Chou, and T. K. Shih, "Photo defect detection for image inpainting," in *Seventh IEEE International Symposium on Multimedia (ISM'05)*. IEEE, 2005, pp. 5–pp.

[4] I. Giakoumis, N. Nikolaidis, and I. Pitas, "Digital image processing techniques for the detection and removal of cracks in digitized paintings," *IEEE Transactions on Image Processing*, vol. 15, no. 1, pp. 178–188, 2005.

[5] K. Zhang, W. Zuo, S. Gu, and L. Zhang, "Learning deep cnn denoiser prior for image restoration," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3929–3938.

[6] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising," *IEEE Transactions on Image Processing*, vol. 26, no. 7, pp. 3142–3155, 2017.

[7] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *European conference on computer vision*. Springer, 2014, pp. 184–199.

[8] L. Xu, J. S. Ren, C. Liu, and J. Jia, "Deep convolutional neural network for image deconvolution," in *Advances in Neural Information Processing Systems*, 2014, pp. 1790–1798.

[9] W. Ren, S. Liu, H. Zhang, J. Pan, X. Cao, and M.-H. Yang, "Single image dehazing via multi-scale convolutional neural networks," in *European conference on computer vision*. Springer, 2016, pp. 154–169.

[10] B. Zhang, M. He, J. Liao, P. V. Sander, L. Yuan, A. Bermak, and D. Chen, "Deep exemplar-based video colorization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8052–8061.

[11] Q. Gao, X. Shu, and X. Wu, "Deep restoration of vintage photographs from scanned halftone prints," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 4120–4129.

[12] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, 2017.

[13] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[14] A. Buades, B. Coll, and J.-M. Morel, "A non-local algorithm for image denoising," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 2. IEEE, 2005, pp. 60–65.

[15] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Non-local sparse models for image restoration," in *2009 IEEE 12th international conference on computer vision*. IEEE, pp. 2272–2279.

[16] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Image denoising by sparse 3-d transform-domain collaborative filtering," *IEEE Transactions on image processing*, vol. 16, no. 8, pp. 2080–2095, 2007.

[17] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Transactions on Image processing*, vol. 15, no. 12, pp. 3736–3745, 2006.

[18] J. Mairal, M. Elad, and G. Sapiro, "Sparse representation for color image restoration," *IEEE Transactions on image processing*, vol. 17, no. 1, pp. 53–69, 2007.

[19] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE transactions on image processing*, vol. 19, no. 11, pp. 2861–2873, 2010.

[20] J. Xie, L. Xu, and E. Chen, "Image denoising and inpainting with deep neural networks," in *Advances in neural information processing systems*, 2012, pp. 341–349.

[21] Y. Weiss and W. T. Freeman, "What makes a good model of natural images?" in *2007 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2007, pp. 1–8.

[22] S. D. Babacan, R. Molina, and A. K. Katsaggelos, "Total variation super resolution using a variational approach," in *2008 15th IEEE International Conference on Image Processing*. IEEE, 2008, pp. 641–644.

[23] S. Z. Li, *Markov random field modeling in image analysis*. Springer Science & Business Media, 2009.

[24] K. Zhang, W. Zuo, and L. Zhang, "Ffdnet: Toward a fast and flexible solution for cnn-based image denoising," *IEEE Transactions on Image Processing*, vol. 27, no. 9, pp. 4608–4622, 2018.

[25] X. Mao, C. Shen, and Y.-B. Yang, "Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections," in *Advances in neural information processing systems*, 2016, pp. 2802–2810.

[26] S. Lefkimmiatis, "Universal denoising networks: a novel cnn architecture for image denoising," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3204–3213.

[27] D. Liu, B. Wen, Y. Fan, C. C. Loy, and T. S. Huang, "Non-local recurrent network for image restoration," in *Advances in Neural Information Processing Systems*, 2018, pp. 1673–1682.

[28] Y. Zhang, K. Li, K. Li, B. Zhong, and Y. Fu, "Residual non-local attention networks for image restoration," *arXiv preprint arXiv:1903.10082*, 2019.

[29] J. Kim, J. Kwon Lee, and K. Mu Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1646–1654.

[30] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," *arXiv preprint*, 2017.

[31] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. Change Loy, "Esrgan: Enhanced super-resolution generative adversarial networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 0–0.

[32] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2472–2481.

[33] J. Sun, W. Cao, Z. Xu, and J. Ponce, "Learning a convolutional neural network for non-uniform motion blur removal," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 769–777.

[34] S. Nah, T. Hyun Kim, and K. Mu Lee, "Deep multi-scale convolutional neural network for dynamic scene deblurring," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3883–3891.

[35] O. Kupyn, V. Budzan, M. Mykhailych, D. Mishkin, and J. Matas, "Deblurgan: Blind motion deblurring using conditional adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8183–8192.

[36] G. Liu, F. A. Reda, K. J. Shih, T.-C. Wang, A. Tao, and B. Catanzaro, "Image inpainting for irregular holes using partial convolutions," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 85–100.

[37] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Generative image inpainting with contextual attention," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5505–5514.

[38] H. Liu, B. Jiang, Y. Xiao, and C. Yang, "Coherent semantic attention for image inpainting," *arXiv preprint arXiv:1905.12384*, 2019.

[39] Y. Ren, X. Yu, R. Zhang, T. H. Li, S. Liu, and G. Li, "Structureflow: Image inpainting via structure-aware appearance flow," *arXiv preprint arXiv:1908.03852*, 2019.

[40] K. Yu, C. Dong, L. Lin, and C. Change Loy, "Crafting a toolchain for image restoration by deep reinforcement learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2443–2452.

[41] M. Suganuma, X. Liu, and T. Okatani, "Attention-based adaptive selection of operations for image restoration in the presence of unknown combined distortions," *arXiv preprint arXiv:1812.00733*, 2018.

[42] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Deep image prior," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 9446–9454.

[43] Y. Hacohen, E. Shechtman, and D. Lischinski, "Deblurring by example using dense correspondence," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 2384–2391.

[44] J. Pan, Z. Hu, Z. Su, and M.-H. Yang, "Deblurring face images with exemplars," in *European conference on computer vision*. Springer, 2014, pp. 47–62.

[45] C. Chan, S. Ginosar, T. Zhou, and A. A. Efros, "Everybody dance now," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 5933–5942.

[46] Z. Shen, W.-S. Lai, T. Xu, J. Kautz, and M.-H. Yang, "Deep semantic face deblurring," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8260–8269.

[47] A. Bulat and G. 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*, 2018, pp. 109–117.

[48] K. Grm, W. J. Scheirer, and V. Štruc, "Face hallucination using cascaded super-resolution and identity priors," *IEEE Transactions on Image Processing*, vol. 29, no. 1, pp. 2150–2165, 2019.

[49] W. Ren, J. Yang, S. Deng, D. Wipf, X. Cao, and X. Tong, "Face video deblurring using 3d facial priors," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 9388–9397.

[50] S. Menon, A. Damian, S. Hu, N. Ravi, and C. 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*, 2020, pp. 2437–2445.

[51] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2223–2232.

[52] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[53] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *CoRR*, vol. abs/1312.6114, 2013.

[54] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley, "Least squares generative adversarial networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2794–2802.

[55] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European conference on computer vision*. Springer, 2016, pp. 694–711.

[56] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7794–7803.

[57] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, "Semantic image synthesis with spatially-adaptive normalization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2337–2346.

[58] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *International journal of computer vision*, vol. 111, no. 1, pp. 98–136, 2015.

[59] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2019, pp. 4401–4410.

[60] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.
[61] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.
[62] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
[63] A. Kumar, T. Ma, and P. Liang, "Understanding self-training for gradual domain adaptation," *arXiv preprint arXiv:2002.11361*, 2020.
[64] E. Agustsson and R. Timofte, "Ntire 2017 challenge on single image super-resolution: Dataset and study," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017.
[65] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional gans," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8798–8807.
[66] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Bm3d image denoising with shape-adaptive principal component analysis," 2009.
[67] K. Nazeri, E. Ng, T. Joseph, F. Qureshi, and M. Ebrahimi, "Edgeconnect: Generative image inpainting with adversarial edge learning," 2019.
[68] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 586–595.
[69] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *Advances in Neural Information Processing Systems*, 2017, pp. 6626–6637.
[70] "Meitu," https://www.meitu.com/en.
[71] "Remini photo enhancer," https://www.bigwinepot.com/index_en.html.
[72] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein gan," *arXiv preprint arXiv:1701.07875*, 2017.
[73] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on image processing*, vol. 21, no. 12, pp. 4695–4708, 2012.