From Face to Natural Image: Learning Real Degradation for Blind Image Super-Resolution

Xiaoming Li[1,5], Chaofeng Chen[2], Xianhui Lin[3], Wangmeng Zuo[1,4(✉)], and Lei Zhang[5]

[1] Faculty of Computing, Harbin Institute of Technology, China
[2] S-Lab, Nanyang Technological University, Singapore
[3] DAMO Academy, Alibaba Group, Shenzhen, China
[4] Peng Cheng Lab, Shenzhen, China
[5] Department of Computing, The Hong Kong Polytechnic University
{csxmli, chaofenghust, xhlin129}@gmail.com, wmzuo@hit.edu.cn, cslzhang@comp.polyu.edu.hk

Abstract How to design proper training pairs is critical for super-resolving real-world low-quality (LQ) images, which suffers from the difficulties in either acquiring paired ground-truth high-quality (HQ) images or synthesizing photo-realistic degraded LQ observations. Recent works mainly focus on modeling the degradation with handcrafted or estimated degradation parameters, which are however incapable to model complicated real-world degradation types, resulting in limited quality improvement. Notably, LQ face images, which may have the same degradation process as natural images, can be robustly restored with photo-realistic textures by exploiting their strong structural priors. This motivates us to use the real-world LQ face images and their restored HQ counterparts to model the complex real-world degradation (namely ReDegNet), and then transfer it to HQ natural images to synthesize their realistic LQ counterparts. By taking these paired HQ-LQ face images as inputs to explicitly predict the degradation-aware and content-independent representations, we could control the degraded image generation, and subsequently transfer these degradation representations from face to natural images to synthesize the degraded LQ natural images. Experiments show that our ReDegNet can well learn the real degradation process from face images. The restoration network trained with our synthetic pairs performs favorably against SOTAs. More importantly, our method provides a new way to handle the real-world complex scenarios by learning their degradation representations from the facial portions, which can be used to significantly improve the quality of non-facial areas. The source code is available at https://github.com/csxmli2016/ReDegNet.

Keywords: real world degradation, blind image super-resolution

1 Introduction

It is widely known that Convolutional Neural Networks (CNNs) are proficient in handling the data they have seen, but perform inferior on these deviating

Figure 1: (a): A real-world LQ image. (b)∼(g): Restoration comparisons with inverse halftone method [43], RealSR [19], Real-ESRGAN [48], BSRGAN [56], BSRGAN* fine-tuned with halftone degradation [13], and Ours^s that is specifically trained with the synthetic pairs in (i). (h): Face restoration result by GPEN [53]. (i): Our synthetic LQ sample with the degradation representation from (h).

from the training sets. This property makes the blind image super-resolution networks difficult to handle the real-world LQ images which are usually corrupted with complex and unsynthesizable degradation. However, building these pairs of real-world LQ and HQ datasets is neither feasible nor practical, because the real-world degradation types are too diverse and some of them are not brought by the imaging system. Figure 1 (a) shows a real-world LQ image that is degraded with halftone related artifacts. One can see that the synthetic LQ image (on the top-left of (b)) by the inverse halftoning method [13] is hardly consistent with the complex real-world degradation, which makes these types of restoration methods (*e.g.*, [43]) fail to generate photo-realistic result (see (b)).

To alleviate the difficulties in restoring the real-world LQ images, some works attempt to predict the degradation parameters [16, 17, 19, 20, 34] and then handle the LQ input with the non-blind restoration works. However, the real degradation usually combines with various corruption types, each of which has lost its intrinsic characteristics. This inevitably makes these methods sensitive to the prediction errors of the degradation parameters, and consequently makes them fail to handle the real-world LQ image (see (c) in Figure 1).

Recently, data-driven methods are suggested to design a practical degradation model by handcrafting the complex combinations of blur, downsampling, noise and JPEG compression with random [56] or high orders [48]. Albeit these methods have more diverse degradation types [10, 31, 57] and show great generalization in handling the real-world LQ images in most cases, they still fail to cover some

complex real degradation which cannot be well synthesized (see (d) and (e) in Figure 1). By incorporating the synthetic halftone degradation [13], BSRGAN* has slight improvement (see (f)), but still contains obvious linearity artifacts.

In contrast, face image has specific and strong structure prior, and can be better restored while exhibiting great generalization ability on real-world LQ images in most cases [27, 47, 53]. Although the image is corrupted by intractable degradation, the face restoration result is very plausible and photo-realistic (see Figure 1 (h)). Since the face and non-face (natural) regions in an image share the same degradation, once we have known the degradation process on face regions, transferring it to natural HQ images would bring considerable benefits, *e.g.*, we can apply this degradation process on the HQ natural image to synthesize these types of natural image pairs (see (i)) for training restoration network (see (g)).

In this paper, we make the first attempt to explore the **re**al **deg**radation with ReDegNet, which contains (i) learning the real degradation from the pairs of real-world LQ and pseudo HQ face images with DegNet, and (ii) transferring it to HQ natural images to synthesizing their realistic LQ ones with SynNet. As for (i), instead of taking a single LQ image to predict its degradation parameters [19], our DegNet takes the real-world LQ and its pseudo HQ face images as input to generate the degradation representation, which models the degradation process of how the HQ image is degraded to the LQ one. To disentangle the image content and degradation type, we adopt two manners, *i.e.*, a) carefully designed framework by predicting the degradation representation through several fully connected layers to generate the convolution weights which can be regarded as the styles in StyleGANs [22, 23], and b) contrastive loss [46] by minimizing the representation distance between the pairs with different content but degraded with the same degradation parameters, and meanwhile maximizing these with the same content but different degradation. This process is fully supervised by the paired LQ/HQ face images. As for (ii), our SynNet synthesizes the realistic LQ natural images with these degradation representations extracted from face images, which can help us to learn the real-world restoration mapping. Note that our method may perform limited on scenarios without faces. By extending the degradation space with face images share the similar degradation, our model would be further improved. The main contributions are summarized as follows:

- We propose the ReDegNet to explore the real degradation from face images by explicitly learning the degradation-aware and content-independent representations which control the degraded image generation.
- We transfer these real-world degradation representations to HQ natural images to generate their realistic LQ ones for supervised real restoration.
- We provide a new manner for handling intractable degraded images by learning their degradation from face regions within them, which can be used for synthesizing these types of LQ natural images for specifically fine-tuning.
- Experimental results demonstrate that our ReDegNet can well learn the degradation representations from face images and can effectively transfer to natural ones, contributing to the comparable performance on general restoration and superior performance in specific scenarios against the SOTAs.

2 Related Work

2.1 Blind Face Restoration

Different from the complex textures in natural images, the specific structure in face images make it feasible to well handle the real-world LQ face images [5, 7, 8, 18, 24, 60]. To alleviate the sensibility for the unknown degradation, reference images or component features are suggested for guiding the blind restoration process [27–29]. Most recently, generative face prior [22, 23] based methods [4, 47, 53] are proposed to improve and stabilize the restoration quality, which can robustly restore the real-world LQ face images in most scenarios. Their great generalization on face images inspires us to explore the possibility of extending the restoration performance from the local region (*i.e.*, face) to the whole image.

2.2 Degradation Estimation Based Blind Image Super-Resolution

The real-world LQ images are mainly corrupted with unknown degradation parameters, so some works focus on estimating these degradation parameters and then apply non-blind restoration methods to recover it. Bell-Kligler *et al.* [2] firstly propose the image-specific KernelGAN to predict the blur kernels and feed them to ZSSR [41] for non-blind restoration. Gu *et al.* [16] introduce iterative kernel correction method to estimate the blur kernel which further benefits the restoration results. Luo *et al.* [34] alternate the optimization of restoring HQ images with the predicted kernel and estimating the blur kernel with the restored results, both of which can compensate each other. Wang *et al.* [46] suggest a degradation-aware super-resolution network that learn the degradation related parameters to guide the restoration process. However, real-world LQ images usually have high frequency noises or compression artifacts, and these methods are sensitive with them, which brings adverse effect for parameter prediction.

2.3 Data-driven Based Blind Image Super-Resolution

The main challenge of blind image super-resolution task can be ascribed to the lack of suitable training pairs. So a straightforward way is to collect the real-world LQ and HQ pairs. Cai *et al.* [3] adjust the focal length of the digital cameras to capture the paired LQ/HQ images on the same scene. Wei *et al.* [50] build a larger dataset with a large-scale diverse benchmark by zooming the digital cameras. Except for the cumbersome capturing process, the spatial and brightness misalignment easily leads to uncontrollable errors. Moreover, although these images are realistic, they are more suitable for the specific super-resolution task that has the similar capturing scenarios. These types of collecting data occupies very few of these complex real-world degraded images, resulting in the failure cases when handling other real degradation, *e.g.*, noise or compression.

To alleviate the difficulties in synthesizing real-world LQ images, recent works tend to learn the restoration mapping with unpaired LQ and HQ images. Yuan *et al.* [54] suggest a Cycle-in-Cycle network by firstly mapping the LQ input to

noise-free space and then super-resolving it through a pre-trained super-resolution model. Similarly, Lugmayr *et al.* [32] adopt the cycle consistent loss to learn a domain distribution network to generate new LQ/HQ pairs for supervised restoration. Fritsche *et al.* [12] also propose the unsupervised DSGAN model to generate the degraded LQ images with the same characteristics as the original ones. To constitute more realistic LQ images, Ji *et al.* [19] extract the blur kernels via KernelGAN [2] and noise injection through [6, 59], which perform on HQ images to simulate the real degradation process. Although these methods achieve great performance in most cases, they still show limited generalization ability in super-resolving real-world LQ images, because 1) the estimated degradation parameters from only a single image is highly ill-posed and they are not enough to infer how the HQ images degraded (Figure 1 (a)), and 2) the real-world LQ images usually suffer from complex degradation, which is challenging to model due to the lack of paired data. In contrast, our ReDegNet adopts the pairs of real-world LQ and pseudo HQ face images to explore the real degradation process.

Another way is to extend the degradation space. Instead of the traditional degradation process that degrades the HQ image with Gaussian blurring, followed by the bicubic downsampling operation, and the injection of Gaussian noise and JPEG compression, Zhang *et al.* [56] propose a practical degradation model with randomly shuffled orders of these operations which tremendously cover the diverse degradation space. Similarly, Wang *et al.* [48] suggest a high order degradation model with several repeated degradation process. Although these two methods show great generalization in handling real-world images, they are still incapable for those images corrupted with complex degradation like the halftone image in Figure 1 (a). Traditional methods remove these continuous noisy dots mainly through filters [26, 33, 36], look-up-tables [9], dictionary learning [11], or maximum a posteriori estimation [44]. Recent CNN-based inverse halftoning methods [13, 43, 51, 52] and even these estimation or data-driven based methods still fail to generate photo-realistic results on these types of real-world LQ images, which can be ascribed to the difficulties in synthesizing proper LQ images.

3 Methodology

Our ReDegNet aims to learn the real degradation from the pairs of real-world LQ and pseudo HQ face images, and transfer it to natural ones. So it mainly contains two sub-networks, *i.e.*, DegNet for learning the degradation representation Ω, and SynNet for synthesizing the LQ images with the given HQ input and Ω. With the collected real-world LQ face images I_f^{ReaL} and their pseudo HQ ones I_f^{PseH}, the learning process of DegNet ($\mathcal{F}_{Deg}$) and SynNet ($\mathcal{F}_{Syn}$) can be formulated as:

$$\Omega_f^{Rea} = \mathcal{F}_{Deg}\left(I_f^{ReaL}, I_f^{PseH}; \Theta_{Deg}\right), \tag{1}$$

$$\hat{I}_f^{L} = \mathcal{F}_{Syn}\left(I_f^{PseH}, \Omega_f^{Rea}; \Theta_{Syn}\right), \tag{2}$$

where Θ_{Deg} and Θ_{Syn} are the learnable parameters for DegNet and SynNet.

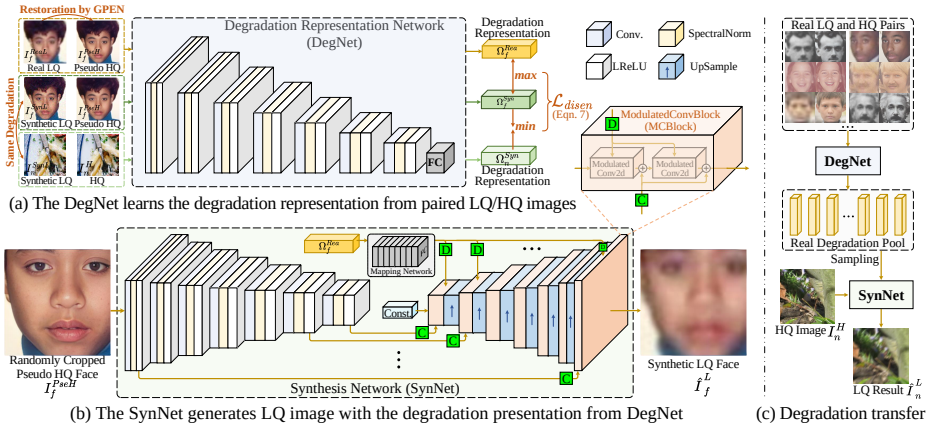

Figure 2: Overview of our ReDegNet. (a) The DegNet learns the degradation representation. (b) The SynNet synthesizes the LQ image with the degradation presentation Ω from DegNet. D denotes a learned affine transform from Ω that produces a degradation style. C represents the content features that will be degraded by D through modulated convolution. (c) The HQ natural image together with the degradation representation sampled from the face pairs are taken into SynNet to generate their synthetic LQ one.

After jointly end-to-end learning through degradation disentanglement, the synthetic realistic LQ natural images can be obtained in the inference through:

$$\hat{I}_n^L = \mathcal{F}_{Syn}\left(I_n^H, \Omega_f^{Rea}; \Theta_{Syn}\right), \tag{3}$$

where I_n^H and $\hat{I}_n^L$ are the HQ and the synthetic LQ natural images, respectively. Ω_f^{Rea} can be sampled from these real degradation representations which are extracted from the collected real-world face pairs. The whole framework and each sub-network are illustrated in Figure 2 and will be introduced in the following.

3.1 Learning Real Degradation from Face Image

Instead of predicting the degradation related representations from only a single LQ image [19, 46], we take the LQ and HQ pairs as input to explore the degradation process about how the HQ image is degraded to the LQ one. The degradation representation network (DegNet) shown in Figure 2 (a) is stacked with several convolutional layers, each of which followed by spectral normalization [39] and LeakyReLU activation. A fully convolutional (FC) layer is incorporated in the last to predict the degradation representation vector Ω, which has the size of 1×512. This sub-network is optimized through two terms, *i.e.*, the disentanglement loss in Eqn. 7 and the gradient back propagated from the following SynNet.

3.2 Synthesizing the LQ Image

After obtaining the degradation representation vector Ω, the remaining problem is about how to utilize it to control the degradation process. Inspired by the Style-

GANs [22,23] that control the style of the generated image with one vector within $\mathcal{W}$ space, we adopt the similar structure to map the degradation representation Ω to $\mathcal{W}$ space through several fully convolutional (FC) layers. Then, instead of feeding the broadcast noise in StyleGAN, the image content of our SynNet is provided by the features of the input HQ images. Finally, with the degradation styles D and image content C, the degraded image is reconstructed with the modulated convolution operation (MCBlock) in which the degradation styles serve as the convolutional weights to control the degradation process of the given image content [23]. With several cascaded MCBlocks, the final LQ result which is expected to have the similar degradation types with the given degradation representation can be synthesized. Since the degradation vector Ω should be a global representation without any spatial information, here we randomly crop the HQ image as the input of SynNet to alleviate the spatial dependency.

To introduce different scales of textures in the training phase, we adopt the random rotation, resampling, and cropping on face images in DegNet and SynNet, simultaneously. The proposed SynNet combined with DegNet constitutes our ReDegNet that can be jointly optimized in a supervised end-to-end manner.

3.3 Transferring Degradation to Natural Image

After training on face images, our ReDegNet can not only extract the real degradation representation from pairs of face images, but also generate the corresponding LQ image with the expected degradation styles. So as for general restoration, we store large amounts of degradation representations that are extracted from real-world LQ and their pseudo HQ face images, which will be sampled to imitate the real degradation process on natural HQ images (Figure 2 (c)). Here we also resample and rotate the LQ/HQ face pairs to augment the degradation space. Notably, our ReDegNet can be utilized in some specific restoration, in which the degradation types are not easy to synthesize with current degradation model. For the intractable old photos (*e.g.*, Figure 1) or old films, we can obtain their degradation representations with DegNet through the pairs of LQ face region within them and its pseudo HQ result. Then the HQ natural images can be utilized to generate the corresponding LQ image by SynNet to synthesize these degradation types of natural training pairs, which can be used to fine-tune the specific restoration on the whole image.

3.4 Learning Objective

Two types of loss functions are collaborated together to constrain the learning of our ReDegNet, *i.e.*, (i) disentanglement loss that is introduced to extract the degradation-related representations, and (ii) reconstruction loss that is suggested to constrain the synthetic results close to the ground-truth.

Disentanglement Loss. The degradation representations Ω_f^{Rea} learned from face images are expected to perform on natural ones to control the degradation styles, so it should be degradation-aware and content-independent. To achieve this goal, we adopt contrastive learning [40, 46] to minimize the distances of Ωs that

are obtained from images with different content but have the same degradation parameters, and meanwhile maximize these negative pairs. To synthesize the degraded face and natural images with the same degradation parameters, we adopt the handcrafted degradation model from BSRGAN [56] to control the degradation process. To clarify the notations, we give a unified definition $I^{\blacktriangle}_{\blacktriangledown}$ in which $\blacktriangle \in \{SynL, ReaL, PseH, H\}$ denotes the handcrafted synthetic LQ image with BSRGAN [56], real-world LQ image, the restored pseudo HQ image, and real-world HQ image, respectively, $\blacktriangledown \in \{f, n\}$ represents the face and natural image, respectively. Denote these synthetic face and natural pairs with BSRGAN by $\{I_f^{PseH}, I_f^{SynL}\}$ and $\{I_n^{H}, I_n^{SynL}\}$. It should be noted that I_f^{SynL} and I_n^{SynL} are obtained from I_f^{PseH} and I_n^{H} with the same degradation sequence and parameters. As for these three types of pairs, *i.e.*, real-world LQ and HQ face pairs, synthetic LQ and HQ face pairs, as well as synthetic LQ and HQ natural pairs, their degradation representations can be formulated as:

$$\Omega_f^{Rea} = \mathcal{F}_{Deg}\left(I_f^{ReaL}, I_f^{PseH}; \Theta_{Deg}\right), \tag{4}$$

$$\Omega_f^{Syn} = \mathcal{F}_{Deg}\left(I_f^{SynL}, I_f^{PseH}; \Theta_{Deg}\right), \tag{5}$$

$$\Omega_n^{Syn} = \mathcal{F}_{Deg}\left(I_n^{SynL}, I_n^{H}; \Theta_{Deg}\right). \tag{6}$$

Then the disentanglement loss $\mathcal{L}_{disen}$ can be further formulated as:

$$\mathcal{L}_{disen} = \left\|\Omega_f^{Syn} - \Omega_n^{Syn}\right\|_2^2 + \frac{\lambda}{\|\Omega_f^{Syn} - \Omega_f^{Rea}\|_2^2 + \epsilon} + \frac{1}{2}\left\|\Theta_{Deg}\right\|_2^2, \tag{7}$$

where λ is the trade-off parameter. By minimizing the distance between Ω_f^{Syn} and Ω_n^{Syn} which share the same degradation process but have the different contents (*i.e.*, face and nature), we can constrain the extraction of degradation-aware and content-independent representations. On the contrary, by maximizing the distance between Ω_f^{Syn} and Ω_f^{Rea} which have the same contents (*i.e.*, I_f^{PseH}) but are corrupted with different degradation process, the degradation representation can be further constrained to the degradation-aware learning.

Reconstruction Loss. It mainly contains three terms, *i.e.*, i) mean square error loss $\mathcal{L}_{mse}$, ii) realistic loss $\mathcal{L}_{real}$, and iii) degradation-consistent loss $\mathcal{L}_{cons}$.

i) The MSE loss $\mathcal{L}_{mse}$ contains two terms and is formulated as:

$$\mathcal{L}_{mse} = \ell_{mse}(\hat{I}_f^L, I_f^{ReaL}) = \frac{1}{\mathcal{C}\mathcal{H}\mathcal{W}}\left\|\hat{I}_f^L - I_f^{ReaL}\right\|^2 + \sum_{i=1}^{4}\frac{0.1}{\mathcal{C}_i\mathcal{H}_i\mathcal{W}_i}\left\|\Phi_i(\hat{I}_f^L) - \Phi_i(I_f^{ReaL})\right\|^2 \tag{8}$$

where $\hat{I}_f^L$ is the generated LQ face image in Eqn. 2 and I_f^{ReaL} is the collected real-world LQ image. $\mathcal{C}_*$, $\mathcal{H}_*$, $\mathcal{W}_*$ are the dimensions and Φ_i is the i-th convolution layer of the pre-trained VGG-19 model [42]. This objective constrains the synthetic LQ images close to the real-world LQ images in both pixel and feature space [21].

ii) The realistic loss $\mathcal{L}_{real}$ mainly considers two types of constraints, *i.e.*, style loss [14] and adversarial loss [15]. The first one is computed with the Gram matrix

on the feature spaces of VGG-19 model and can be formulated as:

$$\mathcal{L}_{style} = \sum_{i=1}^{4} \frac{1}{\mathcal{C}_i \mathcal{H}_i \mathcal{W}_i} \left\| \Phi_i(\hat{I}_f^L)^T \Phi_i(\hat{I}_f^L) - \Phi_i(I_f^{ReaL})^T \Phi_i(I_f^{ReaL}) \right\|^2 , \tag{9}$$

in which the variants have the same definitions as these in Eqn. 8. The second one is the widely used adversarial loss which is effective in constraining the results within the natural manifold. In this paper, we adopt the discriminator from SNGAN [39] by incorporating the spectral normalization behind each convolutional layer. It is worth noting that the result $\hat{I}_f^L$ is expected to be a LQ image and visually blur in most cases, which is difficult for discriminator to distinguish whether it is a real LQ or fake LQ image due to the wider space of LQ types. So instead of only taking the synthetic result into the discriminator, we take the HQ image and their degradation representation as additional conditions [37]. The hinge version of adversarial loss [39, 55] is given by:

$$\mathcal{L}_D = -\mathbb{E}[\min(0, -1 + D(I_f^{ReaL}, I_f^{PseH}, \Omega_f^{Rea}))] - \mathbb{E}[\min(0, -1 - D(\hat{I}_f^L, I_f^{PseH}, \Omega_f^{Rea}))] \tag{10}$$

$$\mathcal{L}_G = -\mathbb{E}[D(\mathcal{F}_{Syn}(I_f^{PseH}, \mathcal{F}_{Deg}(I_f^{ReaL}, I_f^{PseH}; \Theta_{Deg}); \Theta_{Syn}), I_f^{PseH}, \Omega_f^{Rea})] . \tag{11}$$

Combining the two terms together, the final realistic loss is formulated as:

$$\mathcal{L}_{real} = 0.1 \cdot \mathcal{L}_{style} + \mathcal{L}_G . \tag{12}$$

iii) The third one is the degradation-consistent loss. As analyzed before, the degradation representation Ω_f^{Syn} and Ω_n^{Syn} in Eqns. 5 and 6 are obtained from the face and natural pairs that are corrupted by the same degradation process. Therefore, switching Ω_f^{Syn} and Ω_n^{Syn} should have the same LQ results. Thus the degradation-consistent loss is suggested as:

$$\mathcal{L}_{cons} = \ell_{mse}(\mathcal{F}_{Syn}(I_n^H, \Omega_f^{Syn}; \Theta_{Syn}), I_n^{SynL}) + \ell_{mse}(\mathcal{F}_{Syn}(I_f^{PseH}, \Omega_n^{Syn}; \Theta_{Syn}), I_f^{SynL}) , \tag{13}$$

where ℓ_{mse} is the MSE loss defined in Eqn. 8. With the constraints on the degradation representation Ω_f^{Syn} (Ω_n^{Syn}) that is extracted from face (natural) images and performed on natural (face) ones, we can further optimize the disentanglement learning, and benefit the training process of the SynNet.

To sum up, the final learning objective is formulated as:

$$\mathcal{L} = \lambda_{disen}\mathcal{L}_{disen} + \lambda_{mse}\mathcal{L}_{mse} + \lambda_{real}\mathcal{L}_{real} + \lambda_{cons}\mathcal{L}_{cons} , \tag{14}$$

where λ_{disen}, λ_{mse}, λ_{real} and λ_{cons} are set to 5, 1, 0.1, and 2, respectively.

4 Experiments

Since our ReDegNet is proposed to design a degradation model for synthesizing LQ images, in this work, we mainly compare with three related works, *i.e.*, RealSR [19], BSRGAN [56] and Real-ESRGAN [48], in which RealSR synthesizes the LQ image with the estimated kernel and noise from the single real-world photograph, BSRGAN and Real-ESRGAN focus on handcrafted design of diverse degradation.

These three methods and Ours adopt the same network (*i.e.*, ESRGAN [49]) , so we can fairly compare with their released models. To evaluate the effectiveness of blind image super-resolution methods on handling the real-world LQ images, here we analyze the performance on two types of real-world images, *i.e.*, real-world pairs collected by digital camera, and real-world single LQ images. As for the quantitative evaluation, we use PSNR, SSIM, and LPIPS [58] to measure the distance between the result and ground-truth. Since real-world single LQ images do not have the ground-truth, we follow the competing methods [19, 48, 56] and adopt NIQE [38] to evaluate the non-reference image quality.

4.1 Dataset and Implementation Details

We collect real-world LQ face images from Internet, and then adopt GPEN [53] to obtain their pseudo HQ counterparts. These images cover diverse degradation types, from slightness to severeness, oldness to present, *etc*. Among them, 10,000 images are used for training, 1,000 images for validating, and the remaining 5,000 images for testing. Except these collected images, we also introduce the synthetic LQ face images from FFHQ [22] with common degradation, *e.g.*, blur, noise, JPEG compression, and downsampling operation, *etc*, to improve the generalization ability. During the inference, we conduct the degradation representation pool $\{\Omega_f^{ReaL}\}^N$ from these face pairs, which will be sampled to constitute the natural pairs for training our general restoration network (*i.e.*, F2N-ESRGAN).

As for the natural image, we follow BSRGAN [56], and adopt DIV2K [1], Flick2K [30, 45] and FFHQ [22] for training our ReDegNet and F2N-ESRGAN. Adam optimizer [25] with $\beta_1 = 0.5$ and $\beta_2 = 0.999$ is adopted to train ReDegNet and F2N-ESRGAN. The initial learning rate is set to 2×10^{-4} and will decrease by 0.5 when the MSE loss $\mathcal{L}_{mse}$ on the validation set tends to be stable. All the experiments are implemented on a PC server with 4 Tesla V100 GPUs.

4.2 Quantitative Comparison

Table 1 lists the quantitative results. One can see that (i) as for these real-world pairs (RealSR Canon and Nikon [3], and DRealSR [50]), although the PSNR and SSIM of Ours is comparable against others, the LPIPS of Ours obtains the best performance, which indicates that our results are more consistent with human perception [58]. The best LPIPS of Ours in turn validates the effectiveness of our ReDegNet in synthesizing the realistic training pairs. (ii) As for the non-reference image quality metric, we collect two groups of real-world images, *i.e.*, RealSRSet proposed in BSRGAN [56], and RealLQSet that contains LQ images collected from Internet and LQ frames extracted from 480P videos. We can see that results of Ours are better than others in most cases, but inferior to RealSR [19] in RealSRSet [56]. We analyze that the RealLQSet (1,000 images) covers more types of common real-world LQ images than RealSRSet (only 20 images), which indicates RealLQSet is more suitable in evaluating the performance of super-resolving the real-world LQ images. The better NIQE of Ours may be attributed to the usage of degradation that are learned from real-world LQ face images.

Table 1: Quantitative comparison on two types of real-world LQ images.

Methods	Real-world Pairs									Real-world LQ	
	RealSR-Canon			RealSR-Nikon			DRealSR			RealSRSet	RealLQSet
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	NIQE↓	NIQE↓
RealSR	25.58	.723	.458	**25.49**	.693	.459	**27.69**	**.759**	.438	**4.82**	5.62
BSRGAN	**25.61**	**.768**	.363	24.51	.711	.391	26.64	.744	.380	5.60	5.36
Real-ESRGAN	24.95	**.768**	.366	24.50	**.716**	.388	26.57	.753	.374	5.75	5.24
Ours	25.57	.765	**.362**	25.43	**.716**	**.385**	26.91	.758	**.373**	4.85	**4.93**
Ours (-D)	24.63	.749	.463	24.35	.684	.460	26.32	.740	.425	6.45	6.27
Ours (U)	25.05	.752	.428	24.72	.708	.421	26.35	.741	.404	5.81	5.93

Figure 3: Visual comparison of these competing methods on real-world LQ images.

4.3 Visual Comparison on Real-world LQ Images

Except the quantitative metrics, visual comparison appears to be critically important in evaluating the restoration performance, especially for these real-

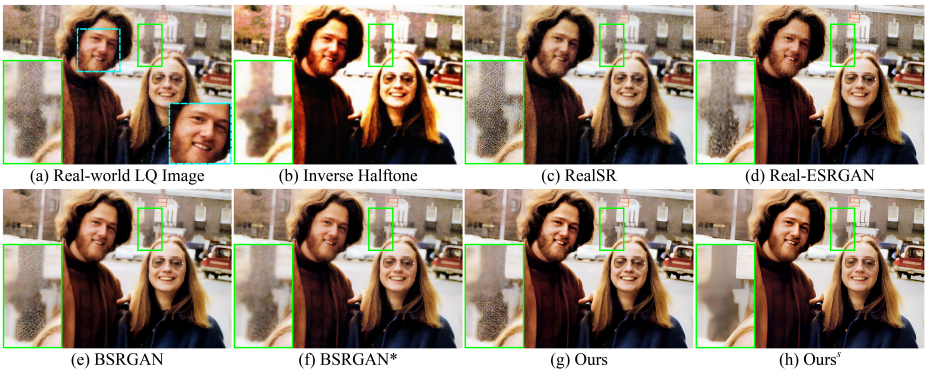

Figure 4: Restoration results on real-world LQ image. Close-up on the right bottom of (a) is the face restoration result by GPEN [53]. BSRGAN* denotes the official BSRGAN model fine-tuned with the incorporation of halftone degradation [13]. Ours^s represents our general model (Ours) that is specifically fine-tuned with the synthetic natural pairs with the degradation representation from the face region. Best view it by zooming in.

world LQ images. In this paper, we select these real-world LQ images from three types of datasets, *i.e.*, RealSRSet from BSRGAN [56], RealSR dataset [3], and the collected real-world LQ images from our RealLQSet. Visual results of the competing methods are shown in Figure 3. One can see that results of Ours are much clearer than others, not only in the smooth regions ($1st$ row), but also in these with rich and complex textures ($2\sim4th$ rows). Due to the limited ability in predicting the kernel and noise from the real-world LQ images, RealSR [19] fails to generate plausible and photo-realistic textures when handling the input with complex degradation. Although BSRGAN [56] and Real-ESRGAN [48] show great generalization due to the wider handcrafted degradation spaces, our F2N-ESRGAN performs comparable against them with these degradation representations that are extracted from the real-world LQ face images, which indicates the effectiveness of our method in synthesizing the photo-realistic LQ images, and in turn contributes to the better restoration performance.

4.4 Fine-tuning for Specific Restoration

Except the general super-resolution task mentioned above, our method can also fine-tune the restoration model on specific scenarios which have face images in them. Figures 1 and 4 show the specific cases. We can observe that (1) although they are similar to the halftone degradation, the restoration result by the inverse halftone method [43] can not well handle it (see (b)) due to the complex degradation that these real-world LQ images usually suffer from. (2) The general restoration methods, *i.e.*, RealSR [19], BSRGAN [56], and Real-ESRGAN [48] also fail to generate plausible results on these unsynthesizable degradation (see (c∼e)), while Ours perform favorable but still contain obvious artifacts (see (g)). (3) By fine-tuning BSRGAN with the synthetic halftone degradation [13],

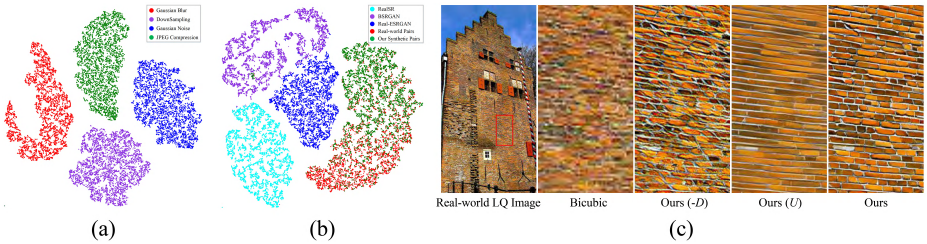

Figure 5: (a) The t-SNE results of four groups degradation with only blur, downsampling, noise and JPEG compression. (b) The t-SNE results of the synthetic degradation by the competing methods. (c) Restoration comparison of different variants.

BSRGAN* has slight improvement in reducing the artifacts, but still can not generate photo-realistic structures (see (f)). (4) By restoring the face region with GPEN [53] and synthesizing the similar degradation types on natural images (see our suppl.), results of Ours^s are much better than others, which indicates the effectiveness of our method in learning the degradation from face images and transferring to natural ones. Compared with BSRGAN and Real-ESRGAN, our method can not only handle the general restoration with limited real degradation, but also fine-tune the model for some specific scenarios that have face images, which are common in the consumer photography and old photos or films.

4.5 Ablation Study

Firstly, to illustrate the degradation extraction ability of our DegNet, we introduce t-SNE [35] to visualize the degradation representation Ω for different degradation types. To this end, we generate four groups of LQ face pairs by separately degrading 5,000 HQ test images with Gaussian blurring, downsampling, Gaussian noise and JPEG compression. Then DegNet is utilized to extract their degradation representations. The visualization of each group mapping to 2D space by t-SNE is shown in Figure 5 (a). We can observe that these four groups of degradation representations are embedded into four clusters completely, which indicates that our DegNet can well capture and distinguish the different degradation types.

Secondly, we explore the degradation space of these competing methods. With the 5,000 real-world test LQ and HQ face pairs, we synthesize the LQ images by utilizing the degradation models of RealSR [19], BSRGAN [56], Real-ESRGAN [48] and our ReDegNet on the pseudo HQ images. Among them, the kernel and noise of RealSR are extracted from the real-world images. BSRGAN and Real-ESRGAN are used with their default settings from their released models. As for ours, we randomly sample from the degradation representation pool $\{\Omega_f^{ReaL}\}^N$ via SynNet to generate the LQ images. Note that $\{\Omega_f^{ReaL}\}^N$ have no overlap with the 5,000 test pairs. The visualization of the degradation representations of these five groups, *i.e.*, RealSR, BSRGAN, Real-ESRGAN, Ours and the real-world LQ/HQ pairs, is shown in Figure 5 (b). One can see that our synthetic LQ images are more consistent with the real-world LQ ones than the

competing methods, indicating the effectiveness of our method in extracting the degradation from real pairs of face images. Albeit BSRGAN and Real-ESRGAN have more diversities due to the random/high orders and handcrafted degradation, only few LQ images are similar to the real-world LQ ones within 5,000 pairs.

Finally, to evaluate the necessities of the disentanglement loss and our pairs of LQ and HQ face images, we design two variants, *i.e.*, Ours (U) by using unpaired data which feeds only the LQ images into DegNet and random HQ images into SynNet, respectively, and adopts the discriminator to distinguish whether the result has the similar degradation with the LQ input or not, and Ours (-D) by removing the disentanglement loss. The comparisons on real-world LQ images are shown in Table 1 and Figure 5 (c). We can see that compared with Ours (U), results of Ours are clearer and more photo-realistic, indicating the effectiveness of our supervised manner in predicting the real degradation from the pairs of face images. Besides, by removing the disentanglement loss, results of Ours (-D) easily have distorted structures and obvious artifacts, which may be caused by the inaccurate degradation representation that may contain the face related content.

4.6 Limitations

This work is intuitively motivated by the observation that the face region usually shares the similar degradation with the non-face region. However, the background is sometimes out of the depth of field, which easily has the inconsistent degradation with face region, thereby bring limited benefits for the specific restoration. Besides, our general restoration model performs not obviously superior to the competing methods, especially on these camera captured test sets in Table 1. It is better to collect face images under the similar scenarios to augment the degradation space.

5 Conclusion

In this work, we made the first attempt to model the real degradation from the real-world LQ face images and their pseudo HQ counterparts, and transfer these real degradation processes to HQ natural images by disentangling the degradation-aware and content-independent representations. With the synthetic natural image pairs generated by our ReDegNet, the trained blind image super-resolution models (*i.e.*, F2N-ESRGAN) demonstrated competitive performance against SOTA methods, especially on real-world LQ images. Our method provided a new solution to synthesize more realistic LQ natural images with the degradation representation that are extracted from the facial regions within them, which are beneficial for restoring the details of non-facial regions. Experiments showed that our ReDegNet can well learn the real degradation from face images, and can effectively generate the photo-realistic LQ natural ones, thereby leading to promising performance in general and specific restoration.

Acknowledgment This work is partially supported by the National Natural Science Foundation of China under grant No. U19A2073, the Major Key Project of PCL under grant No. PCL2021A12, and Hong Kong RGC RIF grant (R5001-18).

References

1. Agustsson, E., Timofte, R.: Ntire 2017 challenge on single image super-resolution: Dataset and study. In: CVPRW (2017)
2. Bell-Kligler, S., Shocher, A., Irani, M.: Blind super-resolution kernel estimation using an internal-gan. In: NeurIPS (2019)
3. Cai, J., Zeng, H., Yong, H., Cao, Z., Zhang, L.: Toward real-world single image super-resolution: A new benchmark and a new model. In: ICCV (2019)
4. Chan, K.C., Wang, X., Xu, X., Gu, J., Loy, C.C.: Glean: Generative latent bank for large-factor image super-resolution. In: CVPR (2021)
5. Chen, C., Li, X., Yang, L., Lin, X., Zhang, L., Wong, K.Y.K.: Progressive semantic-aware style transformation for blind face restoration. In: CVPR (2021)
6. Chen, J., Chen, J., Chao, H., Yang, M.: Image blind denoising with generative adversarial network based noise modeling. In: CVPR (2018)
7. Chen, Y., Tai, Y., Liu, X., Shen, C., Yang, J.: Fsrnet: End-to-end learning face super-resolution with facial priors. In: CVPR (2018)
8. Chrysos, G.G., Zafeiriou, S.: Deep face deblurring. In: CVPRW (2017)
9. Chung, K.L., Wu, S.T.: Inverse halftoning algorithm using edge-based lookup table approach. IEEE TIP (2005)
10. Elad, M., Feuer, A.: Restoration of a single superresolution image from several blurred, noisy, and undersampled measured images. IEEE TIP (1997)
11. Freitas, P.G., Farias, M.C., Araújo, A.P.: Enhancing inverse halftoning via coupled dictionary training. Signal Processing: Image Communication (2016)
12. Fritsche, M., Gu, S., Timofte, R.: Frequency separation for real-world super-resolution. In: ICCVW (2019)
13. Gao, Q., Shu, X., Wu, X.: Deep restoration of vintage photographs from scanned halftone prints. In: ICCV (2019)
14. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: CVPR (2016)
15. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: NeurIPS (2014)
16. Gu, J., Lu, Zuo, W., Dong, C.: Blind super-resolution with iterative kernel correction. In: CVPR (2019)
17. Guo, S., Yan, Z., Zhang, K., Zuo, W., Zhang, L.: Toward convolutional blind denoising of real photographs. In: CVPR (2019)
18. Huang, H., He, R., Sun, Z., Tan, T.: Wavelet-srnet: A wavelet-based cnn for multi-scale face super resolution. In: ICCV (2017)
19. Ji, X., Cao, Y., Tai, Y., Wang, C., Li, J., Huang, F.: Real-world super-resolution via kernel estimation and noise injection. In: CVPRW (2020)
20. Jiang, J., Zhang, K., Timofte, R.: Towards flexible blind JPEG artifacts removal. In: ICCV (2021)
21. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: ECCV (2016)
22. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: CVPR (2019)
23. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: CVPR (2020)
24. Kim, D., Kim, M., Kwon, G., Kim, D.S.: Progressive face super-resolution via attention to facial landmark. In: BMVC (2019)
25. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: ICLR (2015)

26. Kite, T.D., Damera-Venkata, N., Evans, B.L., Bovik, A.C.: A fast, high-quality inverse halftoning algorithm for error diffused halftones. IEEE TIP (2000)
27. Li, X., Chen, C., Zhou, S., Lin, X., Zuo, W., Zhang, L.: Blind face restoration via deep multi-scale component dictionaries. In: ECCV (2020)
28. Li, X., Li, W., Ren, D., Zhang, H., Wang, M., Zuo, W.: Enhanced blind face restoration with multi-exemplar images and adaptive spatial feature fusion. In: CVPR (2020)
29. Li, X., Liu, M., Ye, Y., Zuo, W., Lin, L., Yang, R.: Learning warped guidance for blind face restoration. In: ECCV (2018)
30. Lim, B., Son, S., Kim, H., Nah, S., Mu Lee, K.: Enhanced deep residual networks for single image super-resolution. In: CVPRW (2017)
31. Liu, C., Sun, D.: On bayesian adaptive video super resolution. IEEE TPAMI (2013)
32. Lugmayr, A., Danelljan, M., Timofte, R.: Unsupervised learning for real-world super-resolution. In: ICCVW (2019)
33. Luo, J., De Queiroz, R., Fan, Z.: A robust technique for image descreening based on the wavelet transform. IEEE Transactions on Signal Processing (1998)
34. Luo, Z., Huang, Y., Li, S., Wang, L., Tan, T.: Unfolding the alternating optimization for blind super resolution. In: NeurIPS (2020)
35. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research (2008)
36. Miceli, C.M., Parker, K.J.: Inverse halftoning. Journal of Electronic Imaging (1992)
37. Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784 (2014)
38. Mittal, A., Soundararajan, R., Bovik, A.C.: Making a "completely blind" image quality analyzer. IEEE SPL (2012)
39. Miyato, T., Kataoka, T., Koyama, M., Yoshida, Y.: Spectral normalization for generative adversarial networks. In: ICLR (2018)
40. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: CVPR (2015)
41. Shocher, A., Cohen, N., Irani, M.: "zero-shot" super-resolution using deep internal learning. In: CVPR (2018)
42. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR (2015)
43. Son, C.H.: Inverse halftoning through structure-aware deep convolutional neural networks. Signal Processing (2020)
44. Stevenson, R.L.: Inverse halftoning via map estimation. IEEE TIP (1997)
45. Timofte, R., Agustsson, E., Van Gool, L., Yang, M.H., Zhang, L.: Ntire 2017 challenge on single image super-resolution: Methods and results. In: CVPRW (2017)
46. Wang, L., Wang, Y., Dong, X., Xu, Q., Yang, J., An, W., Guo, Y.: Unsupervised degradation representation learning for blind super-resolution. In: CVPR (2021)
47. Wang, X., Li, Y., Zhang, H., Shan, Y.: Towards real-world blind face restoration with generative facial prior. In: CVPR (2021)
48. Wang, X., Xie, L., Dong, C., Shan, Y.: Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In: ICCVW (2021)
49. Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y., Change Loy, C.: Esrgan: Enhanced super-resolution generative adversarial networks. In: ECCVW (2018)
50. Wei, P., Xie, Z., Lu, H., Zhan, Z., Ye, Q., Zuo, W., Lin, L.: Component divide-and-conquer for real-world image super-resolution. In: ECCV (2020)
51. Xia, M., Wong, T.T.: Deep inverse halftoning via progressively residual learning. In: ACCV (2018)

52. Xiao, Y., Pan, C., Zhu, X., Jiang, H., Zheng, Y.: Deep neural inverse halftoning. In: International Conference on Virtual Reality and Visualization. IEEE (2017)
53. Yang, T., Ren, P., Xie, X., Zhang, L.: Gan prior embedded network for blind face restoration in the wild. In: CVPR (2021)
54. Yuan, Y., Liu, S., Zhang, J., Zhang, Y., Dong, C., Lin, L.: Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks. In: CVPRW (2018)
55. Zhang, H., Goodfellow, I., Metaxas, D., Odena, A.: Self-attention generative adversarial networks. In: ICML (2019)
56. Zhang, K., Liang, J., Van Gool, L., Timofte, R.: Designing a practical degradation model for deep blind image super-resolution. In: ICCV (2021)
57. Zhang, K., Zuo, W., Zhang, L.: Deep plug-and-play super-resolution for arbitrary blur kernels. In: CVPR (2019)
58. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018)
59. Zhou, R., Susstrunk, S.: Kernel modeling super-resolution on real low-resolution images. In: ICCV (2019)
60. Zhu, S., Liu, S., Loy, C.C., Tang, X.: Deep cascaded bi-network for face hallucination. In: ECCV (2016)