

# SinIR: Efficient General Image Manipulation with Single Image Reconstruction

Jihyeong Yoo<sup>1</sup> Qifeng Chen<sup>1</sup>

## Abstract

We propose SinIR, an efficient reconstruction-based framework trained on a single natural image for general image manipulation, including super-resolution, editing, harmonization, paint-to-image, photo-realistic style transfer, and artistic style transfer. We train our model on a single image with cascaded multi-scale learning, where each network at each scale is responsible for image reconstruction. This reconstruction objective greatly reduces the complexity and running time of training, compared to the GAN objective. However, the reconstruction objective also exacerbates the output quality. Therefore, to solve this problem, we further utilize simple random pixel shuffling, which also gives control over manipulation, inspired by the Denoising Autoencoder. With quantitative evaluation, we show that SinIR has competitive performance on various image manipulation tasks. Moreover, with a much simpler training objective (*i.e.*, reconstruction), SinIR is trained 33.5 times faster than SinGAN (for  $500 \times 500$  images) that solves similar tasks. Our code is publicly available at [github.com/YooJiHyeong/SinIR](https://github.com/YooJiHyeong/SinIR).

## 1. Introduction

Researchers in image processing have increasing interests in *deep internal learning*, which can solve image manipulation problems by training a model on one single image and not relying on a large-scale dataset. Training on a single image is plausible as the statistics of a single image have abundant information that can be used as a powerful prior for solving various problems (Shaham et al., 2019). Internal learning has long been employed in prior work even before

<sup>1</sup>Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong. Correspondence to: Qifeng Chen <cqf@ust.hk>.

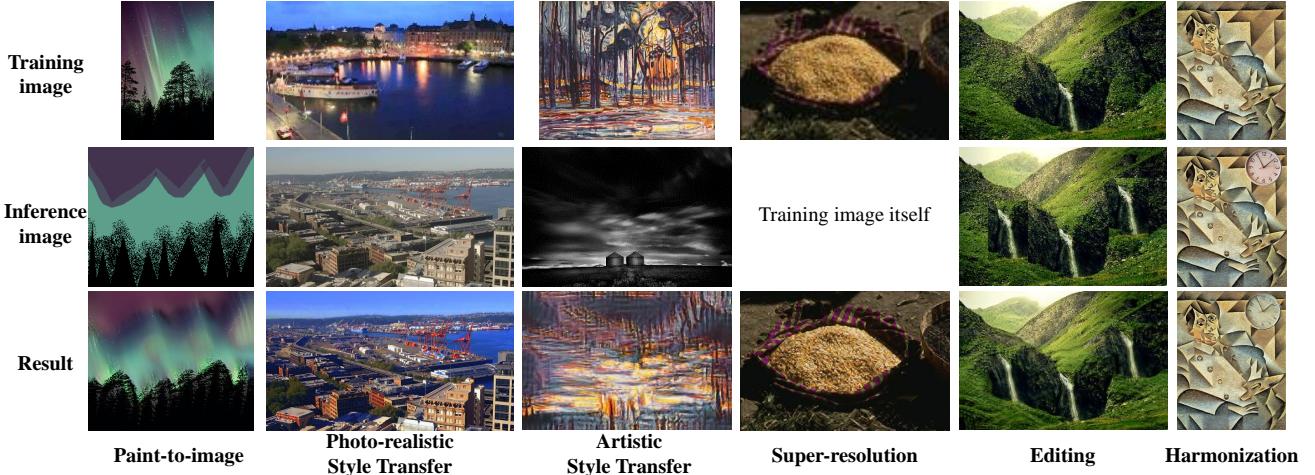
*Proceedings of the 38<sup>th</sup> International Conference on Machine Learning*, PMLR 139, 2021. Copyright 2021 by the author(s).

deep learning became popular, proving internal approaches can be successfully applied to several image manipulation tasks (*e.g.*, super-resolution (Glasner et al., 2009), editing (He & Sun, 2012), dehazing (Bahat & Irani, 2016), texture synthesis (Efros & Leung, 1999), and segmentation (Bagon et al., 2008)).

Recently, several *deep internal learning* methods (Shocher et al., 2018) are proposed and achieve remarkable performance that is comparable to that of external methods trained on large-scale datasets. For example, they solve super-resolution (Shocher et al., 2018; Ulyanov et al., 2018; Bell-Kligler et al., 2019), restoration (Zhang et al., 2019b; Mastan & Raman, 2020), reflection removal (Fan et al., 2019), deblur (Ren et al., 2019), segmentation and dehazing (Gandelsman et al., 2019), denoising, and inpainting (Ulyanov et al., 2018; Zhang et al., 2019a) tasks. Utilizing Generative Adversarial Networks (GAN) (Goodfellow et al., 2014), several approaches (Li & Wand, 2016; Jetchev et al., 2016; Bergmann et al., 2017; Zhou et al., 2018) solve texture generation from a single texture image. Expanding the capacity of GANs, InGAN (Shocher et al., 2019) and DCIL (Mastan & Raman, 2020) solve retargeting with a single natural image.

However, these methods have critical problems that prevent the practical application of deep internal learning. First, most of these methods are image-specific in terms of manipulation, except for MGANs (Li & Wand, 2016). This means that the trained models can only manipulate the training images, and for other images, separate models have to be trained. Second, most of these methods are task-specific. This means that the trained models can only perform one specific image manipulation. Although DIP (Ulyanov et al., 2018), Double-DIP (Gandelsman et al., 2019), and DCIL (Mastan & Raman, 2020) can solve several tasks, these models can only conduct one specific manipulation at a time.

SinGAN (Shaham et al., 2019), one of the most recent works in deep internal learning, was the first to achieve general image manipulation, where one trained model can solve various problems, including super-resolution, editing, paint-to-image, and harmonization. SinGAN learns unconditional generation (*i.e.*, mapping noise to images, not images to images) and then uses the generative power for image ma-



**Figure 1. General image manipulation of SinIR.** SinIR is trained on a *single natural image* with *reconstruction loss*, cascaded multi-scale learning, and random pixel shuffling. Once trained, SinIR can manipulate any image utilizing internal information learned from the training image. Best viewed when zoomed.

nipulation. However, this leads to long training time (1.5 and 4.5 hours for  $250 \times 250$  and  $500 \times 500$  images on RTX 2080 Ti respectively. See Table 1). As unconditional image generation is a relatively difficult problem, SinGAN uses a sophisticated loss function (*i.e.*, WGAN-GP loss (Gulrajani et al., 2017)) for better convergence and trains multiple GANs for a large number of iterations. Although SinGAN is not an image-specific and task-specific framework, this prolonged training time still hinders practical usage of deep internal learning.

In this work, we circumvent this problem with **SinIR**, a reconstruction-based framework trained on a single natural image for general image manipulation. Our model learns **image reconstruction**, which is a much simpler problem compared to the unconditional generation of SinGAN (Shaham et al., 2019). This allows SinIR to achieve a vastly shorter training time (33.49 times faster than SinGAN for  $500 \times 500$  images). To this end, **cascaded multi-scale learning** is employed to learn robust cross-scale representations. However, due to the innate property of reconstruction, simply applying cascaded multi-scale learning leads to poor manipulation quality. Thus, inspired by denoising autoencoder (Vincent et al., 2008), we introduce **random pixel shuffling** that effectively mitigates the problem without a significant increase in computational cost. Furthermore, we show that random pixel shuffling gives additional control over manipulation.

However, we want to make it clear that although SinIR provides a simpler solution to most of the tasks SinGAN can handle, SinIR cannot completely replace SinGAN. This is because SinIR and SinGAN show clearly different capabilities for some tasks. Particularly, SinIR shows less capa-

bility for random image generation but performs well on photo-realistic and artistic style transfer, and vice versa. To our best knowledge, SinGAN is the only well-known deep internal learning framework that performs general image manipulation. Therefore, SinIR and SinGAN are compared to provide meaningful comparisons in solving similar tasks and not for the purpose of replacing the latter. Additionally, recently introduced GAN-prior methods (Pan et al., 2020; Gu et al., 2020) also tackle general image manipulation. However, they are not internal learning methods, and their outputs are obtained directly from the given images. (*i.e.*, there is no separation between training and inference images like SinIR and SinGAN). This results in essential differences in methodology and solvable tasks.

Our main contributions can be summarized as follows.

- To the best of our knowledge, SinIR is the first reconstruction-based deep internal learning framework for general image manipulation.
- By dint of a much simpler training objective (*i.e.*, reconstruction), the training time of SinIR is vastly reduced compared to SinGAN (Shaham et al., 2019) that solves similar problems (trained **33.49 times faster** for  $500 \times 500$  images) as discussed in Section 3.1. This makes deep internal learning more plausible and practical.
- We show that random pixel shuffling enables successful manipulation of SinIR. Moreover, owing to this, SinIR obtains additional controllability over manipulation results. We give analyses with several tasks.
- As depicted in Figure 1 and discussed in Section 3.2 with quantitative evaluations, SinIR has competitive

performance for various image manipulation tasks. Even though SinIR is trained with a single image, it produces visually pleasing results comparable to those of dedicated methods trained on large-scale datasets.

## 2. Method

Our goal is, given *one single natural image*, to train a model for general image manipulation in a much faster way. Although SinGAN (Shaham et al., 2019) solves similar problems, it suffers from prolonged training time with a complex GAN-based objective. Instead of using GANs, we start from a well-known unsupervised representation learning framework, autoencoders (Hinton & Salakhutdinov, 2006), because of its *reconstruction* objective that is much simpler than unconditional image generation.

However, the autoencoder was originally designed to be trained on large-scale datasets, and when trained on a single image, autoencoders may learn trivial identity mapping and reproduce meaningless samples if random inference images are provided for manipulation. Thus, for our model to learn better representations from a single image and achieve general image manipulation, we expand the capability of autoencoders with two methods: **cascaded multi-scale learning** and **random pixel shuffling**.

### 2.1. Training

#### 2.1.1. CASCADED MULTI-SCALE LEARNING

A single natural image often contains various structures across different scales. To successfully learn these cross-scale visual properties, our model learns to **refine a down-sampled training image to obtain the original image in a cascaded manner across multiple scales**. Our model consists of multiple networks that are responsible for refinement at each scale. Outputs of each network will be upsampled and fed into a network at a one-level finer scale (thus, inputs and outputs have the same resolution, but inputs are blurry). The networks at coarser scales will learn more to refine overall structures, while networks at finer scales will learn more to refine detailed textures.

Although there are subtle methodological differences, this multi-scale approach is a well-explored practice (Burt & Adelson, 1983; Denton et al., 2015; Huang et al., 2017b; Chen & Koltun, 2017; Zhang et al., 2017; Li et al., 2017; Wang et al., 2018). Especially to overcome the shortage of large-scale datasets, many deep internal learning frameworks such as ZSSR (Shocher et al., 2018), INGAN (Shocher et al., 2019) and SinGAN (Shaham et al., 2019) adopted this approach. Note that SinIR’s architecture follows that of SinGAN in this work for simplicity.

To train multi-scale networks, we opt for progressively grow-

ing learning (Karras et al., 2018), as it is well-known as one of the most typical frameworks for multi-scale learning. Progressively growing multi-scale learning is widely used by many image synthesis methods (Aigner & Körner, 2018; Karras et al., 2019; Zhang et al., 2019c; Chen & Koltun, 2017; Qi et al., 2018), including SinGAN (Shaham et al., 2019). They train multiple networks one by one from the coarsest scale while freezing previously trained networks. Such methods ease the difficulty of training by solving easier problems one by one.

Specifically, we downsample the training image and obtain ground-truth images for  $N + 1$  scales,  $\{X_N, X_{N-1}, \dots, X_1\}$  and  $X_0 = X$ , where  $X$  is the original training image. Then, at each  $n$ -th scale, we train  $n$ -th network ( $F_n$ ) to reconstruct the one-level finer image. Thus, denoting  $\hat{X}_n = F_n(\hat{X}_{n+1} \uparrow^r)$  and  $\hat{X}_N = F_N(X_N)$  for the coarsest scale ( $\uparrow^r$  means upsampling by a factor  $r$ ), our objective is

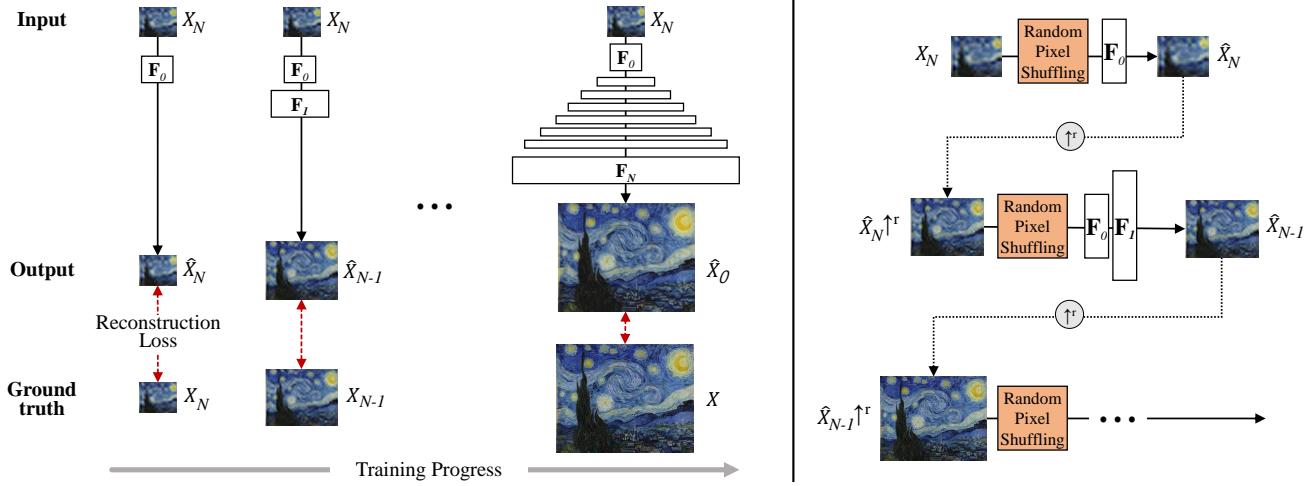
$$\min_{F_n} \mathcal{L}_{\text{rec}}(X_n, \hat{X}_n), \quad (1)$$

where  $\mathcal{L}_{\text{rec}}$  is some reconstruction loss. Once we finish training one network, we freeze this network and add new network to be trained for a one-level finer scale.

#### 2.1.2. RANDOM PIXEL SHUFFLING

Although cascaded multi-scale learning gives a great chance to learn cross-scale representations, it is insufficient to achieve successful manipulation (see visual artifacts in the last row of Figure 3). This is likely because of our training objective, reconstruction. As aforementioned, we use outputs of coarser scales as inputs of finer scales in a cascaded manner. Thus, if outputs of coarser scales show less diversity, it has the effect of showing limited variations in training samples to the next networks. Then it is likely that networks of finer scales will learn trivial or even identity mapping and fail to learn robust representations. In the case of GANs, this problem is naturally avoided as they use random noise (or latent code) for inputs. Besides, because of adversarial loss, they are less constrained in terms of outputs giving diverse samples, compared to the reconstruction loss. The effect of limited diversity is even more exacerbated when we use progressive growing of networks (Karras et al., 2018) for reconstruction because each network is trained only with *fixed* outputs from *frozen* networks. In other words, there is *zero* diversity in inputs of each network.

However, this problem can be effectively mitigated with a simple technique inspired by denoising autoencoder (Vincent et al., 2008): **random pixel shuffling**. An autoencoder may end up learning identity mapping and fail to learn robust representations. However, simply by imposing corruption to inputs, deterministic mapping of the autoencoder can be replaced with stochastic mapping, even giving the



**Figure 2. Training of SinIR.** We train a reconstruction network at each resolution level to refine images in a cascaded manner (Right). However, without random pixel shuffling, it is hard for SinIR to obtain meaningful results as explained in Section 2.1.2 and Figure 3. The outputs of each network are upsampled ( $\uparrow^r$ ) and used as inputs at a one-level finer scale. SinIR is trained with progressively growing learning (Karras et al., 2018) (Left). Please see Section 2.1 for details. Note that SinIR is fully-convolutional, and thus images of any size can be used for training and inference. The network details are in the supplement.

autoencoder a generative property, although it is trained with reconstruction loss. With this, the network can capture the main variation of the images, learning robust representations.

Noticing that the problem posed by denoising autoencoder is very analogous to ours, we apply a similar idea to SinIR. Instead of randomly setting some pixels to complete black as denoising autoencoder does, considering that we are using more complicated natural images, we randomly shuffle some pixels in the given single image so that the network can learn a more robust relationship between adjacent pixels. The effect of different types of corruption is explored in the supplement. But for simplicity, we use random pixel shuffling in this paper.

Rows of Figure 3 illustrates the effect of random pixel shuffling. When we do not use random pixel shuffling (last row), it shows severe artifacts that significantly harm manipulation quality. However, when we randomly shuffle 0.05% to 5% pixels, it shows much better manipulation quality. Interestingly, depending on the percentage, SinIR produces very different results. This behavior has various usages, as discussed in detail later. For example, this is utilized as a tool to control perception-distortion tradeoff (Blau & Michaeli, 2018) of super-resolution (Figure 8). Also, for artistic style transfer, this property can be used to control style-content tradeoff (Figure 5). It can also be used to adjust smoothness as it stands out in Figure 3.

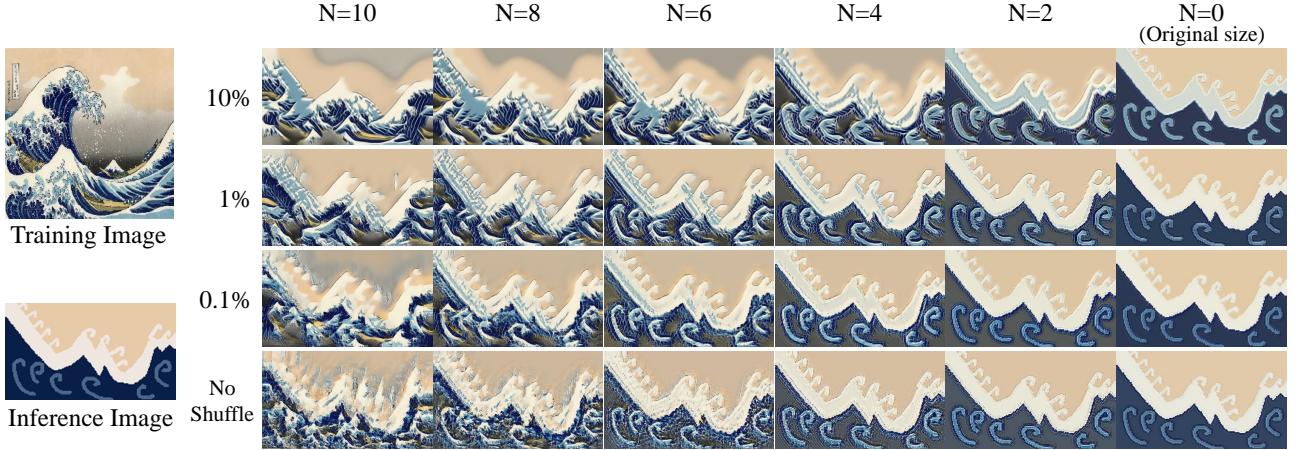
### 2.1.3. OPTIMIZATION

We use the Adam optimizer ( $\beta_1 = 0.5, \beta_2 = 0.999$ ) (Kingma & Ba, 2015) and the learning rate is 1e-4, unless mentioned otherwise. For the reconstruction loss, we use MSE (mean squared error) loss. However, it is well-known that MSE loss produces blurry images (Zhao et al., 2017), thus we combine MSE loss and SSIM (structural similarity) loss (Wang et al., 2004).  $\mathcal{L}_{\text{rec}}$  in Equation 1 is

$$\mathcal{L}_{\text{rec}}(A, B) = \text{MSE}(A, B) + (1 - \text{SSIM}(A, B)). \quad (2)$$

## 2.2. Inference: General Image Manipulation

The columns of Figure 3 depict the effect of different inference starting scales. After deciding on the starting scale, we properly downsample and forward inference images so that when it reaches the top scale ( $n = 0$ ), the size of the output becomes the same as the original. If the inference starting scale is close to the coarsest ( $n = N$ ), the image is manipulated more globally and vice versa. Thus, depending on the inference image and the starting scale, we can achieve general image manipulation with a single trained model. For example, if we feed a clip-art to coarser scales, it performs the paint-to-image task. If we feed a classical painting to the same scales, it performs artistic style transfer. If we feed some images to the finest scale, it performs photo-realistic style transfer as color and tone are manipulated. A similar multi-scale inference scheme can be found in WCT (Li et al., 2017) and SinGAN (Shaham et al., 2019), which also employ multi-scale learning.



**Figure 3. Effects of percentages of randomly shuffled pixels and different inference starting scales.** Each row and column corresponds to different percentages and different inference starting scales. Also, note when we do not shuffle pixels of the training image, speckle-like or pecky stains are observed in manipulated inference images. This is not desirable as it hampers better reflection of original textures from the training image. On the other hand, when we randomly shuffle some pixels, such artifacts are effectively reduced. See Section 2.1 and 2.2 for details.

**Table 1. Comparative analysis of training time.** All results are obtained by averaging ten rounds on a single NVIDIA RTX 2080 Ti GPU.

| Image size<br>(# scales) | SinGAN   | SinIR<br>(Ours) | Speedup |
|--------------------------|----------|-----------------|---------|
| 125px (8)                | 36m 29s  | <b>1m 53s</b>   | ×19.37  |
| 250px (11)               | 90m 39s  | <b>3m 39s</b>   | ×24.84  |
| 500px (13)               | 267m 57s | <b>8m 0s</b>    | ×33.49  |

**Table 2. Results of the user study.** The numbers indicate preference rates. Photo ST and Artistic ST means photo-realistic and artistic style transfer.

|                | SinIR (Ours)  | SinGAN | Indecisive |
|----------------|---------------|--------|------------|
| Photo ST       | <b>90.00%</b> | 6.75%  | 3.25%      |
| Artistic ST    | <b>76.95%</b> | 18.70% | 4.35%      |
| Paint-to-image | <b>64.50%</b> | 28.00% | 8.50%      |
| Editing        | <b>51.25%</b> | 37.00% | 11.75%     |
| Harmonization  | <b>56.90%</b> | 33.33% | 9.76%      |

### 3. Experiments

All experiments are conducted on the same machine with a single NVIDIA RTX 2080 Ti GPU. For a fair comparison, in the same way with SinGAN (Shaham et al., 2019), we set scale factor  $r$  closely to 4/3, minimum and maximum dimension to 25px, 250px, and the number of scale  $N + 1$  is calculated from these parameters. Note that the maximum dimension constraint of SinGAN (250px) due to impractically long training time can be greatly relieved by SinIR’s faster training speed (Table 1). Moreover, for the same reason, all results from SinIR in this section are obtained **in a few minutes**. In terms of iteration number, SinIR requires much fewer iterations as its training objective (*i.e.*, reconstruction) is much simpler than unconditional image generation. Thus, SinIR is trained for 500 iterations at every scale, whereas SinGAN is trained for 6,000 iterations at every scale following the authors’ best practice. For example, when we have 10 scales, SinGAN is trained for 60,000 iterations, and SinIR is trained for 5,000 iterations. As explained in Section 1, although comparisons are mainly

conducted with SinGAN (Shaham et al., 2019), this does not mean SinIR can completely replace SinGAN. The percentage of random pixel shuffling is set to 5e-4. For resampling, a bicubic kernel is used.

#### 3.1. Analysis

Table 1 shows the training time of SinIR and SinGAN (Shaham et al., 2019). SinIR is trained **24.84 times faster** compared to SinGAN, with 250 × 250 images, which is the maximum dimension of training images presented in SinGAN paper. Moreover, SinIR is trained **33.49 times faster** with 500 × 500 images. Owing to these results, SinIR makes deep internal learning for general image manipulation more plausible and practical.

#### 3.2. Applications

We explore the capacity of SinIR for various image manipulation tasks. Inference images are not limited to 250px. All dedicated methods used for comparisons are not internal learning methods, thus require large-scale datasets. More

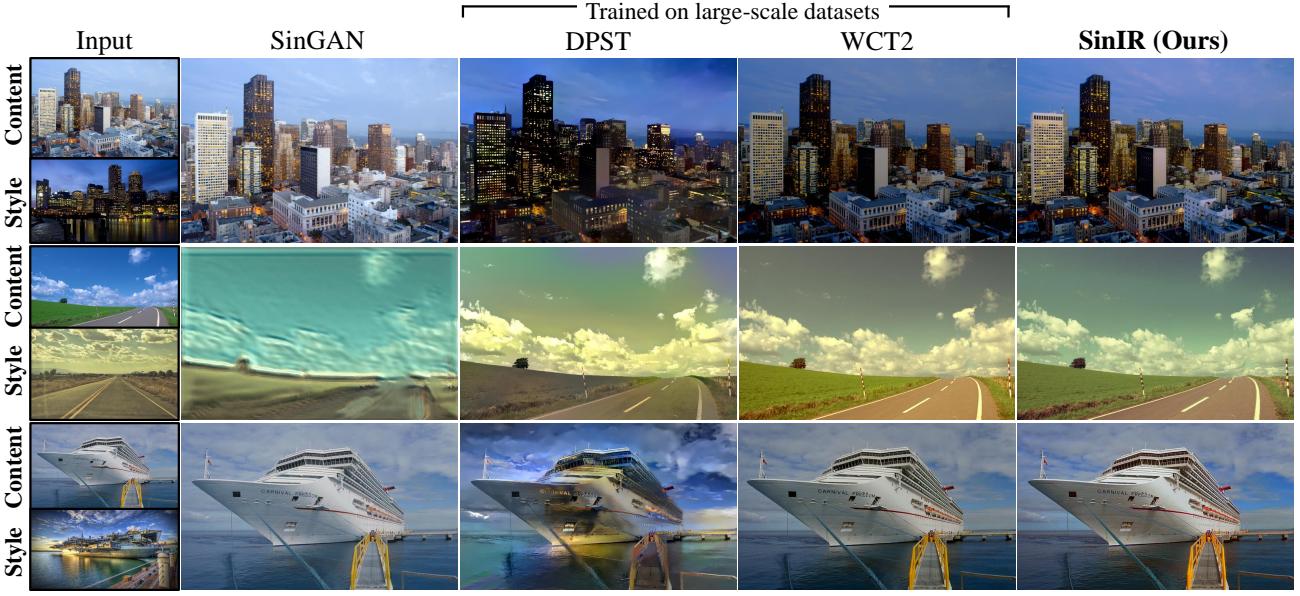


Figure 4. Visual results of **Photo-realistic Style Transfer**. DPST and WCT2 are the results from (Luan et al., 2017) and (Yoo et al., 2019).

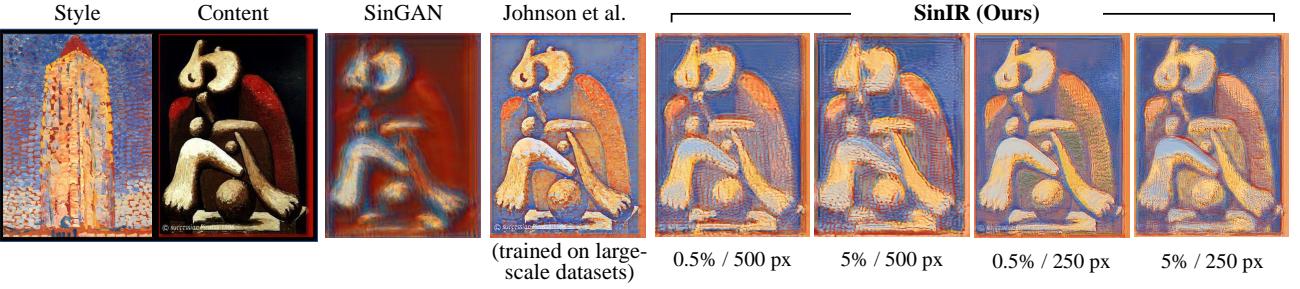


Figure 5. Visual results of **Artistic Style Transfer**. The numbers below the results of SinIR indicate the percentages of randomly shuffled pixels and the maximum dimension of the training image. These results show that SinIR has control over manipulation using random pixel shuffling. The inference starting scale for all SinIR results is  $n = N - 4$ . Note that Johnson et al. (Johnson et al., 2016) is a dedicated method trained externally. Best viewed when zoomed in.

results are included in the supplement. For application, a SinIR model is trained on an image with desired textures (*e.g.*, style image for style-transfer), and then inference images (*e.g.*, content image for style-transfer) are fed to this model for manipulation.

**Photo-realistic Style Transfer.** Photo-realistic style transfer demands strict preservation of the original content while transferring the color and tone of a given style image. To this end, we feed content images to finer scales so that the model does not manipulate overall structures except for the color and tone. In Figure 4, we can observe that SinIR successfully transfers color and tone while not touching structures. On the other hand, SinGAN shows less capability for photo-realistic style transfer. For quantitative evaluation, we conducted a user study using randomly sampled images from a dedicated dataset provided by (Luan

et al., 2017). We showed 20 samples of SinIR and SinGAN to 20 subjects experienced in computer vision and asked to choose better samples. The preference rate of SinIR in Table 2 was significantly higher than that of SinGAN with a considerable margin of 83.25%, which aligns with the qualitative evaluation.

Note that the results of SinIR are similar to those of dedicated methods. Moreover, to further reduce training and inference time, we can optionally set scale factor  $r$  to 1 and the number of scales to 2 or so, as we do not need manipulation of overall structures. With this setting, we averaged 10 inferences of  $1024 \times 1024$  images. Impressively, SinIR took less than a second taking 0.189 s and 0.381 s for starting scales  $n = 0$  and  $n = 1$ , while WCT2 took 2.030 s and 4.799 s with the fastest and slowest methods. Considering WCT2 was 830 times faster than previous methods,

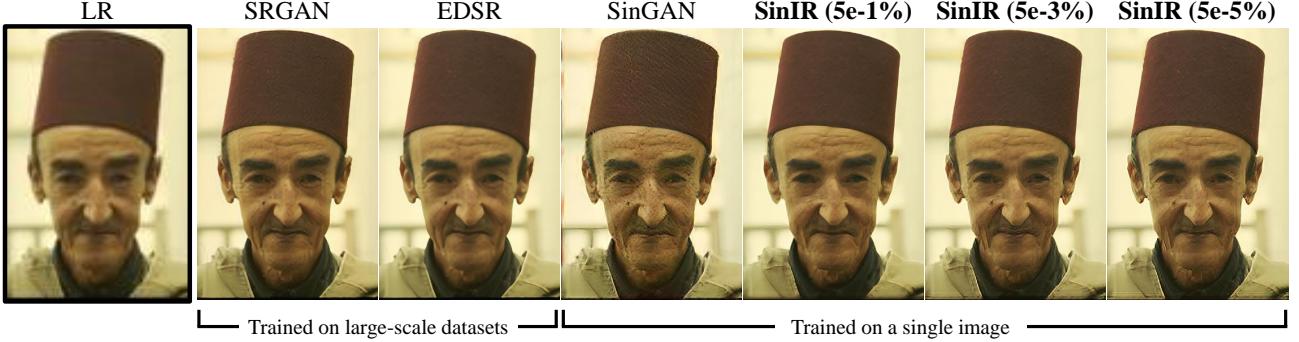


Figure 6. Visual results of **4X Super-resolution**. SRGAN (Ledig et al., 2017) and EDSR(Lim et al., 2017) are dedicated methods trained on large-scale datasets. See Section 3.2 for detailed analysis. Best viewed when zoomed in.

this result makes SinIR close to state-of-the-art, achieving **real-time photo-realistic style transfer**. However, the best scenarios for each method may be different. SinIR may be preferred when real-time inference with a fixed style is desirable and training samples are limited. In contrast, WCT2 may be preferred when training on large-scale datasets is feasible and transferring arbitrary styles is required.

**Artistic Style Transfer.** Artistic style transfer requires the successful blending of style and content. As Figure 5 shows, SinIR is able to produce visually pleasing results, successfully blending the style and the content. Also, SinIR has additional controllability over manipulation results by adjusting the percentage of randomly shuffled pixels, as explained in Section 2.1.2 and Figure 3. In particular, when we increase the percentage, the given style is more aggressively blended, and vice versa (0.5% vs. 5% in Figure 5). If users want to transfer larger textures, a larger maximum dimension of the training images may be preferred (250px vs. 500px in Figure 5). Please zoom and see the blob-like textures which became larger, for example). This option is less feasible with SinGAN (Shaham et al., 2019), because it requires greatly prolonged training time with larger maximum dimension (*e.g.*, 4.5 hours for  $500 \times 500$  images. See Section 3.1 and Table 1). Also, SinGAN shows less capability on this task, obtaining heavily distorted images or limited difference from the content image. For quantitative evaluation, we created 20 samples from SinIR and SinGAN using images collected from the Web and showed them to 23 subjects experienced in computer vision. Aligning with our qualitative evaluation, the preference rate of SinIR was significantly higher than that of SinGAN **with a considerable margin of 58.26%** (Table 2).

**Super-resolution.** For super-resolution, we need to refine detailed textures, but overall structures should largely remain unchanged as it would otherwise lead to unnecessary distortion and poor quality. For this reason, SinIR generally shows better performance with fewer scales. For example, we train SinIR using 2 scales with a scale factor of 2. In

Table 3. **Results of 4X Super-resolution on the BSD100 benchmark.** The percentages of randomly shuffled pixels of SinIR in parentheses. For example, *SinIR (5e-2)* means 0.05% of pixels are shuffled during training.

|              | MS-SSIM↑     | SSIM↑        | RMSE↓        | NIQE↓        |
|--------------|--------------|--------------|--------------|--------------|
| SRGAN        | 0.933        | 0.640        | 16.33        | 3.407        |
| EDSR         | <b>0.963</b> | <b>0.743</b> | <b>12.29</b> | 6.498        |
| SinGAN       | 0.913        | 0.612        | 16.21        | 3.709        |
| SinIR (5e-1) | 0.915        | 0.635        | 16.28        | 4.155        |
| SinIR (5e-2) | 0.920        | 0.632        | 16.56        | 3.647        |
| SinIR (5e-3) | 0.918        | 0.623        | 17.14        | 3.410        |
| SinIR (5e-4) | 0.917        | 0.621        | 17.23        | <b>3.402</b> |
| SinIR (5e-5) | 0.916        | 0.619        | 17.34        | 3.409        |

stead, we set the number of kernels to 256, the iteration number at each scale to 1000 (2000 iterations in total), and the learning rate to 0.001 so that the overall model can still be trained sufficiently with enough capacity. Also, anti-aliasing is recommended when downsampling the training image to suppress undesirable artifacts that may lead to critical degradation in case of super-resolution. During inference, similarly to SinGAN (Shaham et al., 2019), the last network is used for all scales as it better handles detailed textures. The LR image is properly upsampled before it is fed to the model to obtain the final HR image with the target size.

Table 3 shows 4X super-resolution scores using BSD100 dataset (Martin et al., 2001), in terms of distortion (RMSE, SSIM, and MS-SSIM) and perception (NIQE) quality. These are two conflicting criteria for super-resolution often found in a trade-off relationship (Blau & Michaeli, 2018). In general, SinIR shows better perceptual scores that are similar to or even better than SRGAN (Ledig et al., 2017) that is externally trained on a large-scale dataset. Compared to SinGAN (Shaham et al., 2019) that is also trained on a single image, SinIR shows better perceptual quality. In terms of distortion score, SinIR was slightly better in MS-SSIM

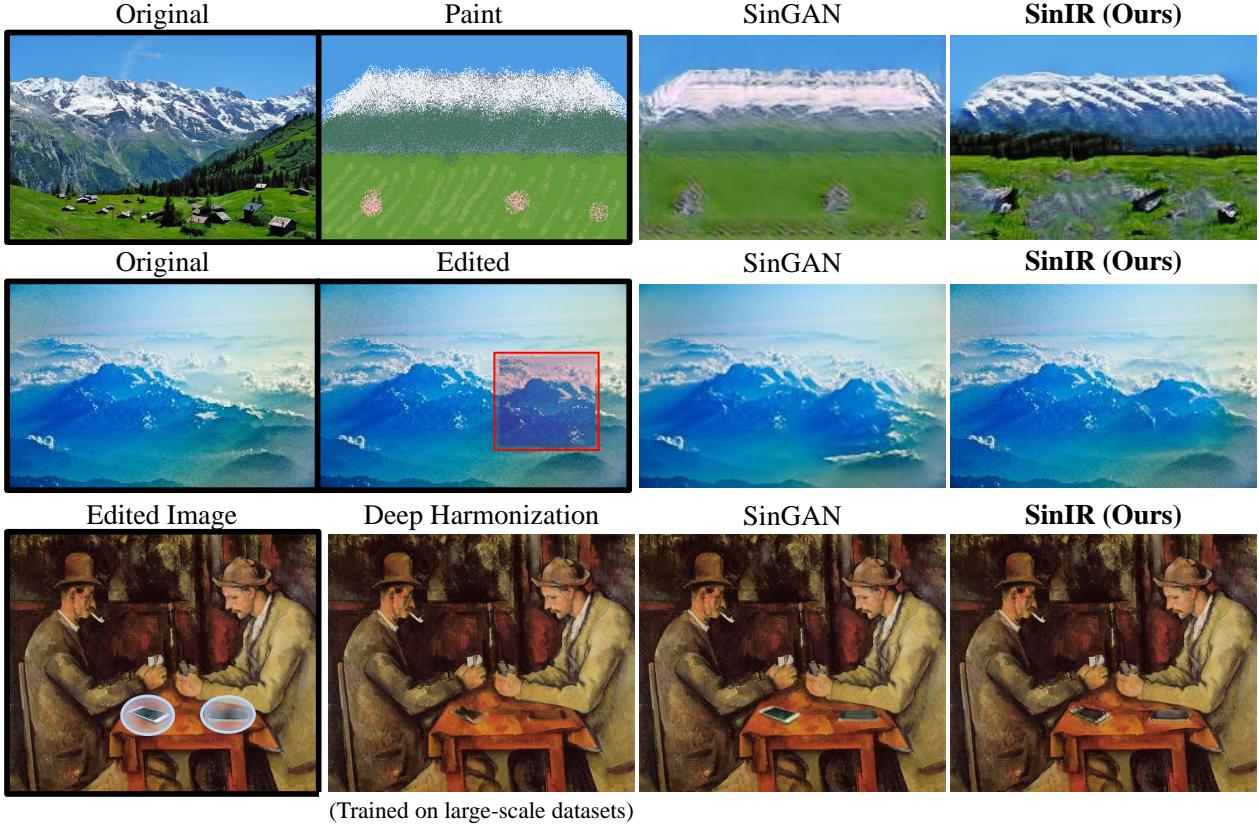


Figure 7. Visual results of **Paint-to-image** (Top), **Editing** (Middle. Edited region in a square), and **Harmonization** (Bottom. Pasted objects in circles). *Deep Harmonization* indicates Deep Paint Harmonization (Luan et al., 2018), a dedicated external method.

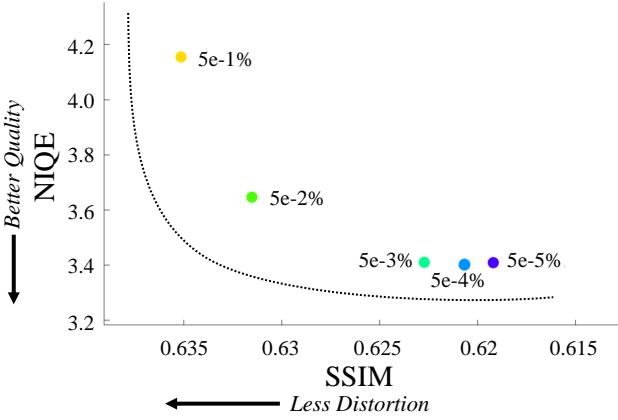


Figure 8. Controllability over perception-distortion trade-off (Blau & Michaeli, 2018) of SinIR. As the percentage of randomly shuffled pixels increases, it shows less distortion but worse perceptual quality, vice versa. The numbers indicate the percentage of randomly shuffled pixels during training.

and SSIM and slightly inferior in RMSE than SinGAN.

Interestingly, by adjusting the percentage of randomly shuffled pixels during training (Section 2.1.2), the perception-distortion trade-off can be controlled to some extent by SinIR. When we increase the percentage, the distortion

score improves while the perception score becomes worse, and vice versa, as depicted in Figure 8.

**Paint-to-Image.** Paint-to-image aims to obtain natural-looking images from simple drawings. As shown in the first row of Figure 7, SinIR produces naturally textured images using the visual characteristics of the training image. For quantitative evaluation, we collected images from the Web and created simple drawings to obtain 20 samples from SinIR and SinGAN (Shaham et al., 2019). Then we showed them to 20 subjects experienced in computer vision and asked them to choose better results. SinIR was preferred to SinGAN **with a margin of 36%** as shown in Table 2.

**Editing.** For this task, we aim to obtain natural-looking images from edited images by blending an edited region with adjacent parts. We first manipulate edited images from SinIR, then combine them with the original using masks like SinGAN (Shaham et al., 2019). The second row of Figure 7 shows that SinIR obtains visually consistent results. For quantitative evaluation, we created 20 edited images from the images collected from the Web. Then we showed the results of SinIR and SinGAN to 20 subjects, and SinIR was preferred to SinGAN **with a margin of 14.25%**.

**Harmonization.** Image harmonization aims at the natural blending of pasted alien objects and original images. We manipulate edited images and combine them with the original using masks like SinGAN (Shaham et al., 2019). The last row of Figure 7 shows that the results of SinIR are comparable to those of SinGAN and Deep Paint Harmonization (Luan et al., 2018), a dedicated method. We conducted a user study using randomly sampled images from a dedicated dataset provided by (Luan et al., 2018). We showed 20 samples to 21 subjects experienced in computer vision. As Table 2 shows, SinIR was preferred to SinGAN **with a margin of 23.57%**.

## 4. Discussion

We introduce SinIR, an internal learning framework trained on a single image for general image manipulation. SinIR learns reconstruction instead of utilizing the adversarial loss in SinGAN (Shaham et al., 2019). Owing to this, SinIR is trained much faster (Table 1). However, to obtain visually pleasing manipulation results, we need to train SinIR with cascaded multi-scale learning and random pixel shuffling. Especially, random pixel shuffling is necessary as our training objective is reconstruction. Achieving general image manipulation with reduced training time, SinIR makes the real-world application of deep internal learning more practical with faster training speed and better visual results.

For future work, although we used progressively growing learning (Karras et al., 2018) for multi-scale learning as it is a well-known technique, more dedicated methods can be explored. Also, while we use only one training image in this work, the effect of training with multiple images can be further explored. Effective methods to handle extreme cases where there are fewer internal references in training images can be explored. Lastly, considering recent findings on memorization and generalization of neural networks with experiments with a single image (Zhang et al., 2020), we find that sophisticated corruption techniques as in (Yu et al., 2019) and (Krull et al., 2019) can be considered instead of simple random pixel shuffling for more robust texture transfer.

## References

- Aigner, S. and Körner, M. Futuregan: Anticipating the future frames of video sequences using spatio-temporal 3d convolutions in progressively growing autoencoder gans. *arXiv: 1810.01325*, 2018.
- Bagon, S., Boiman, O., and Irani, M. What is a good image segment? A unified approach to segment extraction. In *ECCV*, 2008.
- Bahat, Y. and Irani, M. Blind dehazing using internal patch recurrence. In *ICCP*, 2016.
- Bell-Kligler, S., Shocher, A., and Irani, M. Blind super-resolution kernel estimation using an internal-gan. In *NeurIPS*, 2019.
- Bengio, Y., Lamblin, P., Popovici, D., and Larochelle, H. Greedy layer-wise training of deep networks. In *NeurIPS*, 2006.
- Bergmann, U., Jetchev, N., and Vollgraf, R. Learning texture manifolds with the periodic spatial GAN. In *ICML*, 2017.
- Blau, Y. and Michaeli, T. The perception-distortion tradeoff. In *CVPR*, 2018.
- Burt, P. J. and Adelson, E. H. The laplacian pyramid as a compact image code. *IEEE Trans. Commun.*, 1983.
- Chen, Q. and Koltun, V. Photographic image synthesis with cascaded refinement networks. In *ICCV*, 2017.
- Denton, E. L., Chintala, S., Szlam, A., and Fergus, R. Deep generative image models using a laplacian pyramid of adversarial networks. In *NeurIPS*, 2015.
- Efros, A. A. and Leung, T. K. Texture synthesis by non-parametric sampling. In *ICCV*, 1999.
- Fan, Q., Yin, Y., Chen, D., Wang, Y., Avilés-Rivero, A. I., Li, R., Schönlieb, C., Lischinski, D., and Chen, B. Deep reflection prior. *arXiv: 1912.03623*, 2019.
- Gandelsman, Y., Shocher, A., and Irani, M. "double-dip": Unsupervised image decomposition via coupled deep-image-priors. In *CVPR*, 2019.
- Glasner, D., Bagon, S., and Irani, M. Super-resolution from a single image. In *ICCV*, 2009.
- Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. C., and Bengio, Y. Generative adversarial nets. In *NeurIPS*, 2014.
- Gu, J., Shen, Y., and Zhou, B. Image processing using multi-code GAN prior. In *CVPR*, 2020.
- Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C. Improved training of wasserstein gans. In *NeurIPS*, 2017.
- He, K. and Sun, J. Statistics of patch offsets for image completion. In *ECCV*, 2012.
- He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *CVPR*, 2016.
- Hinton, G. E. and Salakhutdinov, R. R. Reducing the dimensionality of data with neural networks. *Science*, 313: 504–507, 2006.

- Huang, G., Liu, Z., van der Maaten, L., and Weinberger, K. Q. Densely connected convolutional networks. In *CVPR*, 2017a.
- Huang, X., Li, Y., Poursaeed, O., Hopcroft, J. E., and Belongie, S. J. Stacked generative adversarial networks. In *CVPR*, 2017b.
- Jetchev, N., Bergmann, U., and Vollgraf, R. Texture synthesis with spatial generative adversarial networks. *arXiv: 1611.08207*, 2016.
- Johnson, J., Alahi, A., and Fei-Fei, L. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016.
- Karras, T., Aila, T., Laine, S., and Lehtinen, J. Progressive growing of gans for improved quality, stability, and variation. In *ICLR*, 2018.
- Karras, T., Laine, S., and Aila, T. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019.
- Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *ICLR*, 2015.
- Krull, A., Buchholz, T., and Jug, F. Noise2void - learning denoising from single noisy images. In *CVPR*, 2019.
- Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, 1998.
- Ledig, C., Theis, L., Huszar, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A. P., Tejani, A., Totz, J., Wang, Z., and Shi, W. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, 2017.
- Li, C. and Wand, M. Precomputed real-time texture synthesis with markovian generative adversarial networks. In *ECCV*, 2016.
- Li, Y., Fang, C., Yang, J., Wang, Z., Lu, X., and Yang, M. Universal style transfer via feature transforms. In *NIPS*, 2017.
- Lim, B., Son, S., Kim, H., Nah, S., and Lee, K. M. Enhanced deep residual networks for single image super-resolution. In *CVPR*, 2017.
- Luan, F., Paris, S., Shechtman, E., and Bala, K. Deep photo style transfer. In *CVPR*, 2017.
- Luan, F., Paris, S., Shechtman, E., and Bala, K. Deep painterly harmonization. *Comput. Graph. Forum*, 37(4): 95–106, 2018.
- Maas, A. L., Hannun, A. Y., and Ng, A. Y. Rectifier nonlinearities improve neural network acoustic models. In *ICML Workshop on Deep Learning for Audio, Speech and Language Processing*, 2013.
- Martin, D., Fowlkes, C., Tal, D., and Malik, J. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *ICCV*, 2001.
- Mastan, I. D. and Raman, S. DCIL: deep contextual internal learning for image restoration and image retargeting. In *WACV*, 2020.
- Pan, X., Zhan, X., Dai, B., Lin, D., Loy, C. C., and Luo, P. Exploiting deep generative prior for versatile image restoration and manipulation. In *ECCV*, 2020.
- Qi, X., Chen, Q., Jia, J., and Koltun, V. Semi-parametric image synthesis. In *CVPR*, 2018.
- Ren, D., Zhang, K., Wang, Q., Hu, Q., and Zuo, W. Neural blind deconvolution using deep priors. *arXiv: 1908.02197*, 2019.
- Shaham, T. R., Dekel, T., and Michaeli, T. Singan: Learning a generative model from a single natural image. In *ICCV*, 2019.
- Shocher, A., Cohen, N., and Irani, M. "zero-shot" super-resolution using deep internal learning. In *CVPR*, 2018.
- Shocher, A., Bagon, S., Isola, P., and Irani, M. Ingan: Capturing and retargeting the "dna" of a natural image. In *ICCV*, 2019.
- Ulyanov, D., Vedaldi, A., and Lempitsky, V. S. Instance normalization: The missing ingredient for fast stylization. *arXive: 1607.08022*, 2016.
- Ulyanov, D., Vedaldi, A., and Lempitsky, V. S. Deep image prior. In *CVPR*, 2018.
- Vincent, P., Larochelle, H., Bengio, Y., and Manzagol, P. Extracting and composing robust features with denoising autoencoders. In *ICML*, 2008.
- Wang, T., Liu, M., Zhu, J., Tao, A., Kautz, J., and Catanzaro, B. High-resolution image synthesis and semantic manipulation with conditional gans. In *CVPR*, 2018.
- Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, 13(4): 600–612, 2004.
- Yoo, J., Uh, Y., Chun, S., Kang, B., and Ha, J. Photorealistic style transfer via wavelet transforms. In *ICCV*, 2019.

Yu, N., Barnes, C., Shechtman, E., Amirghodsi, S., and Lukáć, M. Texture mixer: A network for controllable synthesis and interpolation of texture. In *CVPR*, 2019.

Zhang, C., Bengio, S., Hardt, M., Mozer, M. C., and Singer, Y. Identity crisis: Memorization and generalization under extreme overparameterization. In *ICLR*, 2020.

Zhang, H., Xu, T., and Li, H. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *ICCV*, 2017. URL <https://doi.org/10.1109/ICCV.2017.629>.

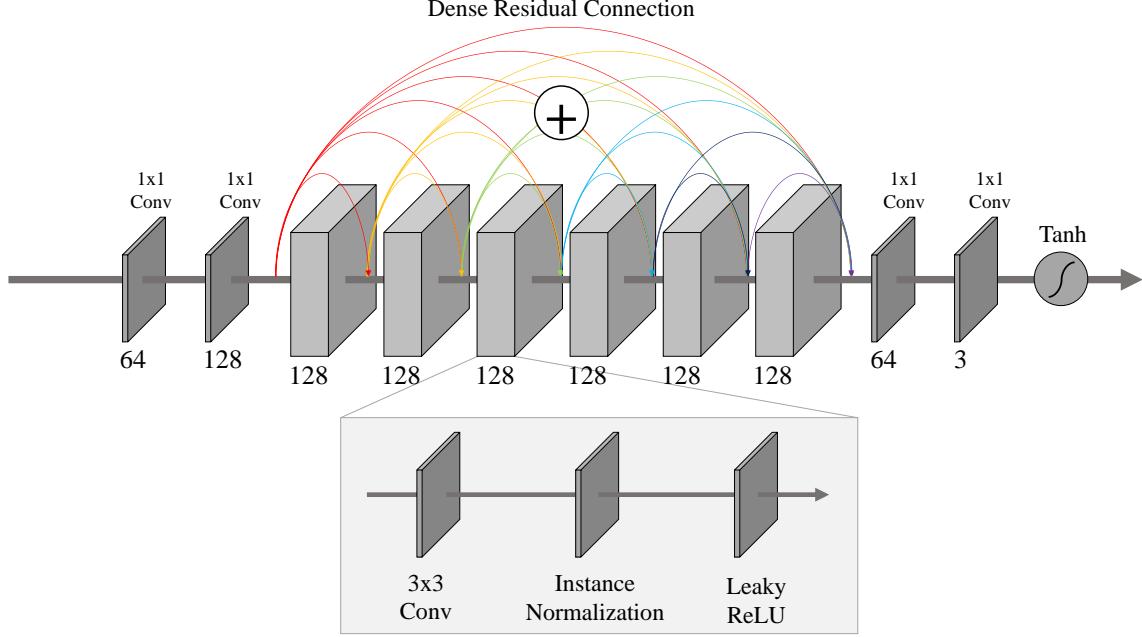
Zhang, H., Mai, L., Jin, H., Wang, Z., Xu, N., and Colomosse, J. P. An internal learning approach to video inpainting. In *ICCV*, 2019a.

Zhang, L., Zhang, L., Liu, X., Shen, Y., Zhang, S., and Zhao, S. Zero-shot restoration of back-lit images using deep internal learning. In *ACM Multimedia*, 2019b.

Zhang, Z., Li, M., and Yu, J. D2PGGAN: two discriminators used in progressive growing of GANS. In *ICASSP*, 2019c.

Zhao, H., Gallo, O., Frosio, I., and Kautz, J. Loss functions for image restoration with neural networks. *IEEE Trans. Computational Imaging*, 3(1):47–57, 2017.

Zhou, Y., Zhu, Z., Bai, X., Lischinski, D., Cohen-Or, D., and Huang, H. Non-stationary texture synthesis by adversarial expansion. *ACM Trans. Graph.*, 37(4):49:1–49:13, 2018.



**Figure 9. Network Architecture of SinIR.** The numbers below each layer indicate the number of convolutional kernels. Here we assume that the input and the output are RGB images.

## A. Network Architecture

All the networks at every scale use the same architecture described in Figure 9. Each network consists of two  $1 \times 1$  convolutional layers which map RGB images to feature space, six convolutional blocks which are densely connected (Huang et al., 2017a) with residual operation (He et al., 2016) (not concatenation), and two  $1 \times 1$  convolutional layers which render features to RGB images. Each convolutional block has one  $3 \times 3$  convolutional layer, an instance normalization layer (Ulyanov et al., 2016) and a LeakyReLU activation layer (negative slope = 0.2) (Maas et al., 2013). We do not use pooling or unpooling inside a network, and thus the inputs and the outputs of each network have the same spatial dimension. Reflection padding is used before  $3 \times 3$  convolutional layer. Tanh function is used to obtain the final output.

## B. Additional Results

We show more results for all tasks presented in the original paper. The setting described in the paper is used for all results. While SinIR and SinGAN (Shaham et al., 2019) are internal methods trained on a single image, all dedicated methods used for comparisons are external methods trained on large-scale datasets.

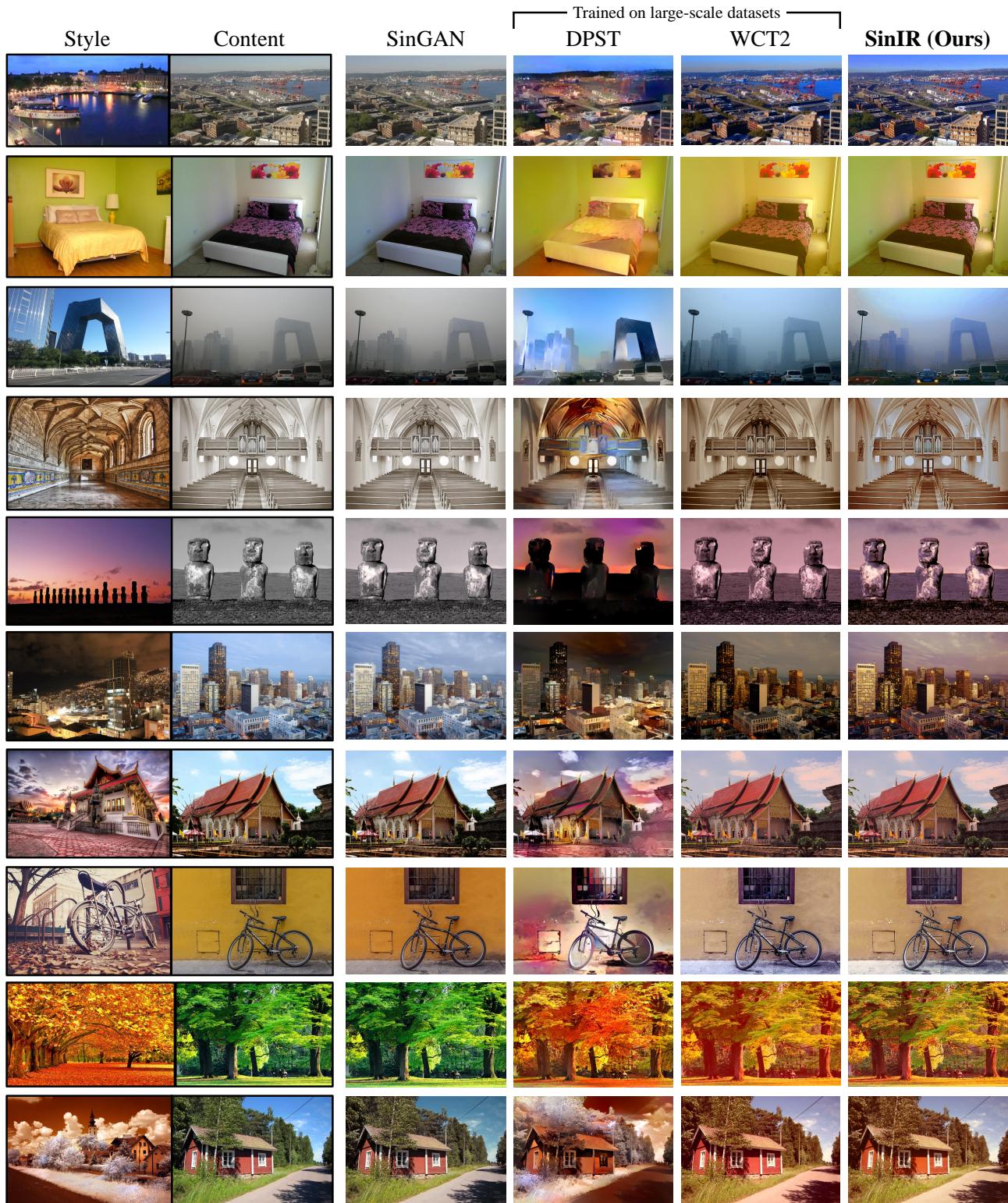


Figure 10. Photo-realistic Style Transfer. DPST (Luan et al., 2017) and WCT2 (Yoo et al., 2019) are dedicated methods.

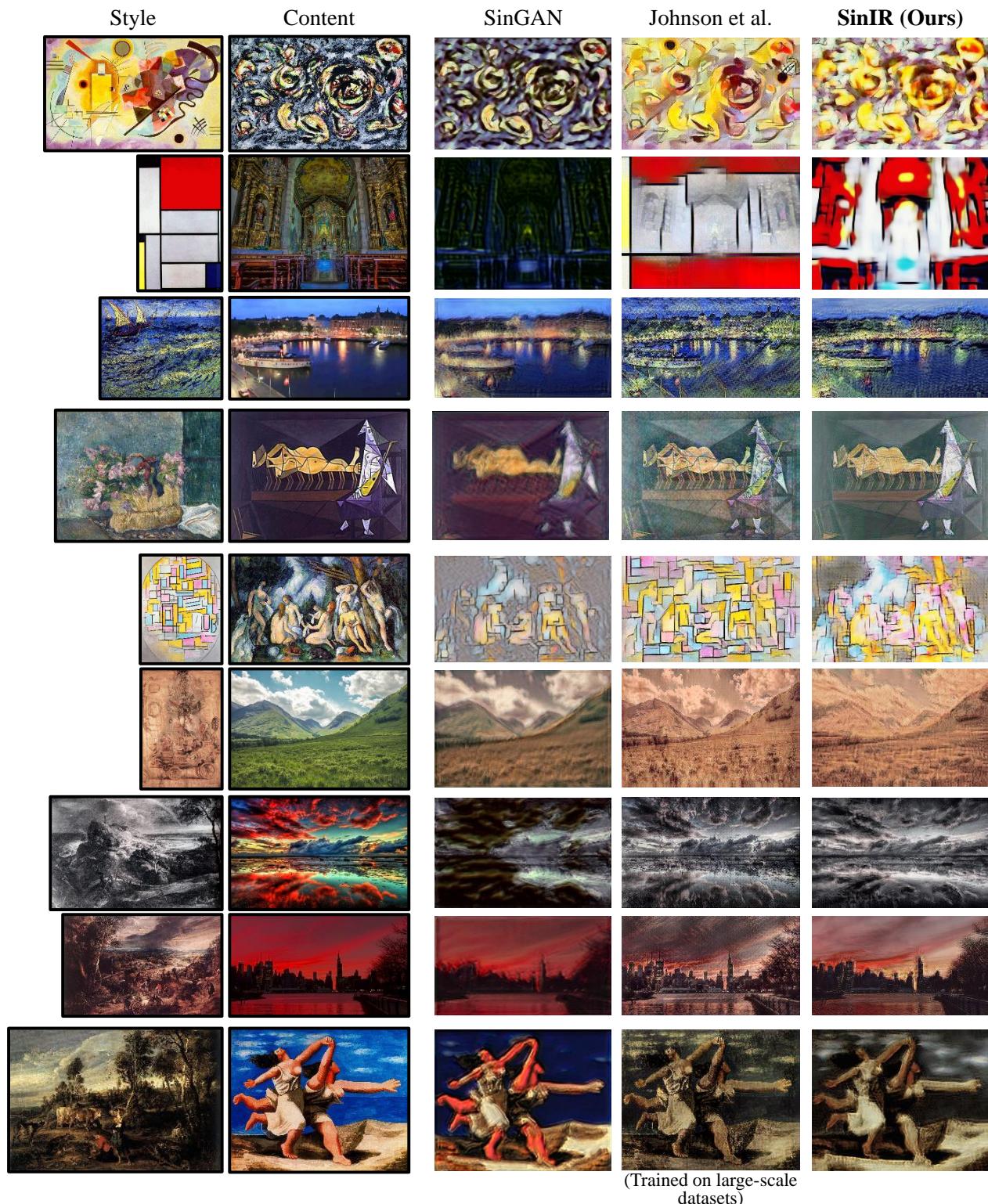
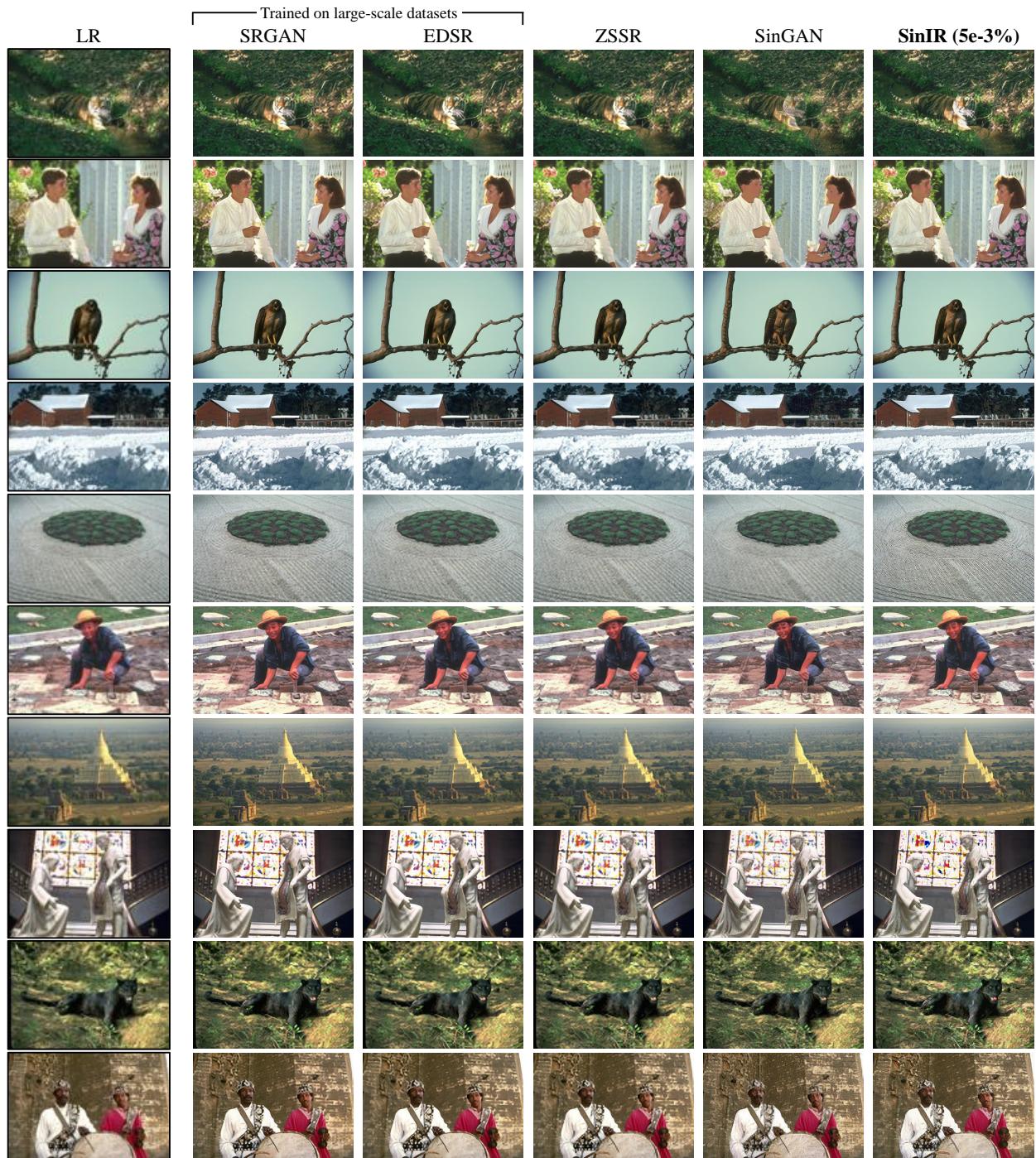


Figure 11. Artistic Style Transfer. (Johnson et al., 2016) is a dedicated method.



**Figure 12. Super-resolution.** SRGAN (Ledig et al., 2017), EDSR (Lim et al., 2017), ZSSR (Shoher et al., 2018) are dedicated methods. For simplicity, only results from SinIR with 5e-3% of random pixel shuffling are shown. (Please see super-resolution section in the original paper for details.)

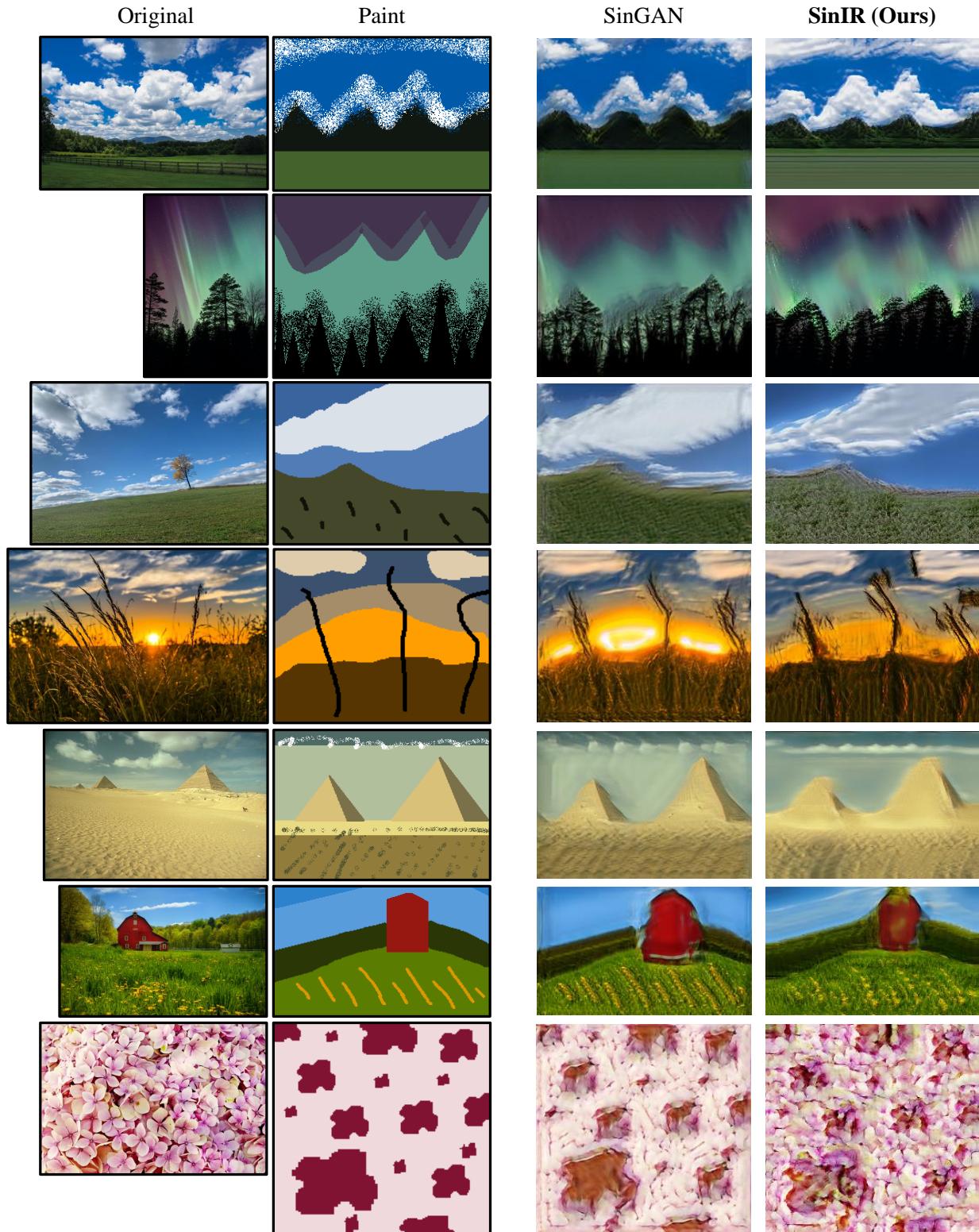


Figure 13. Paint-to-Image.

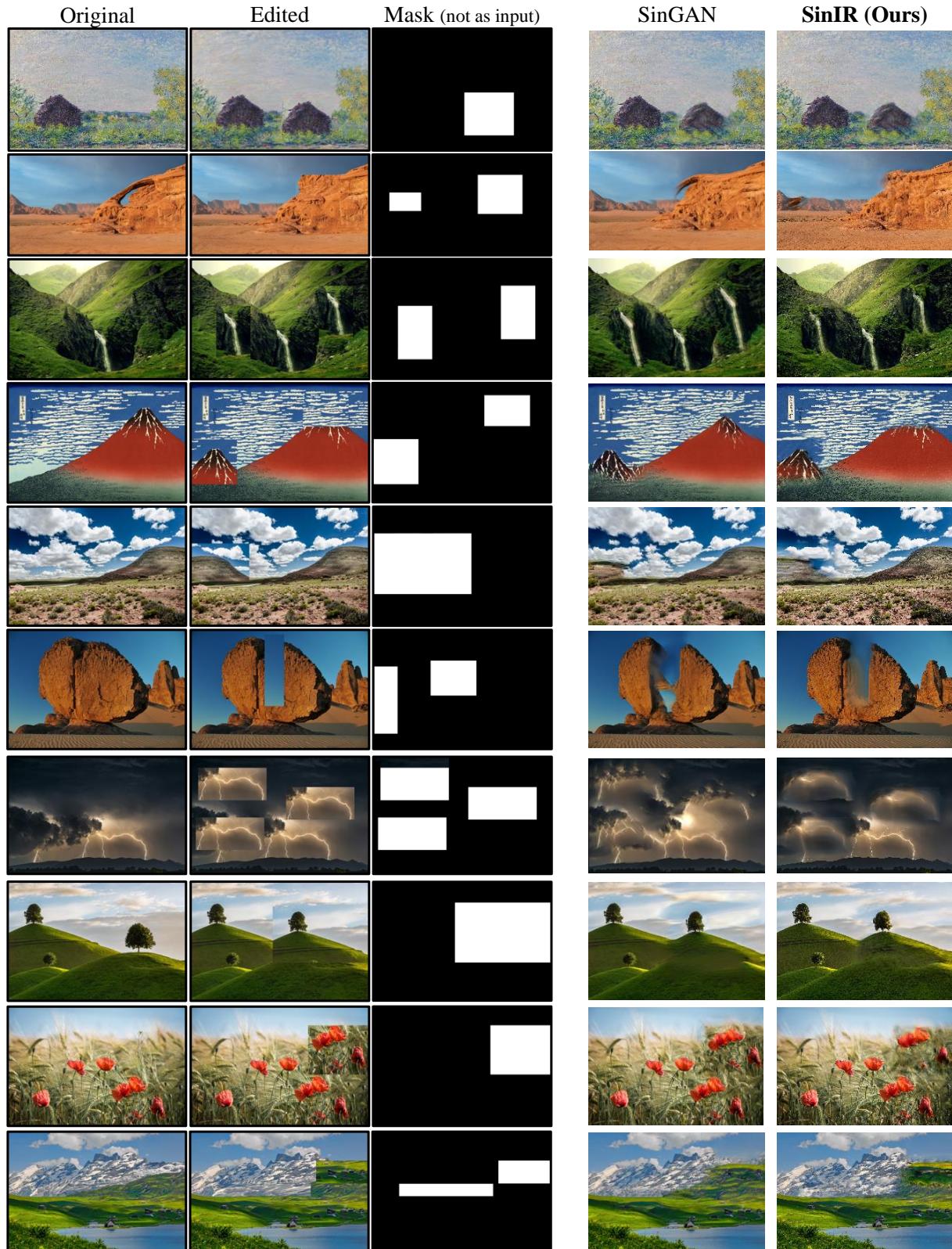
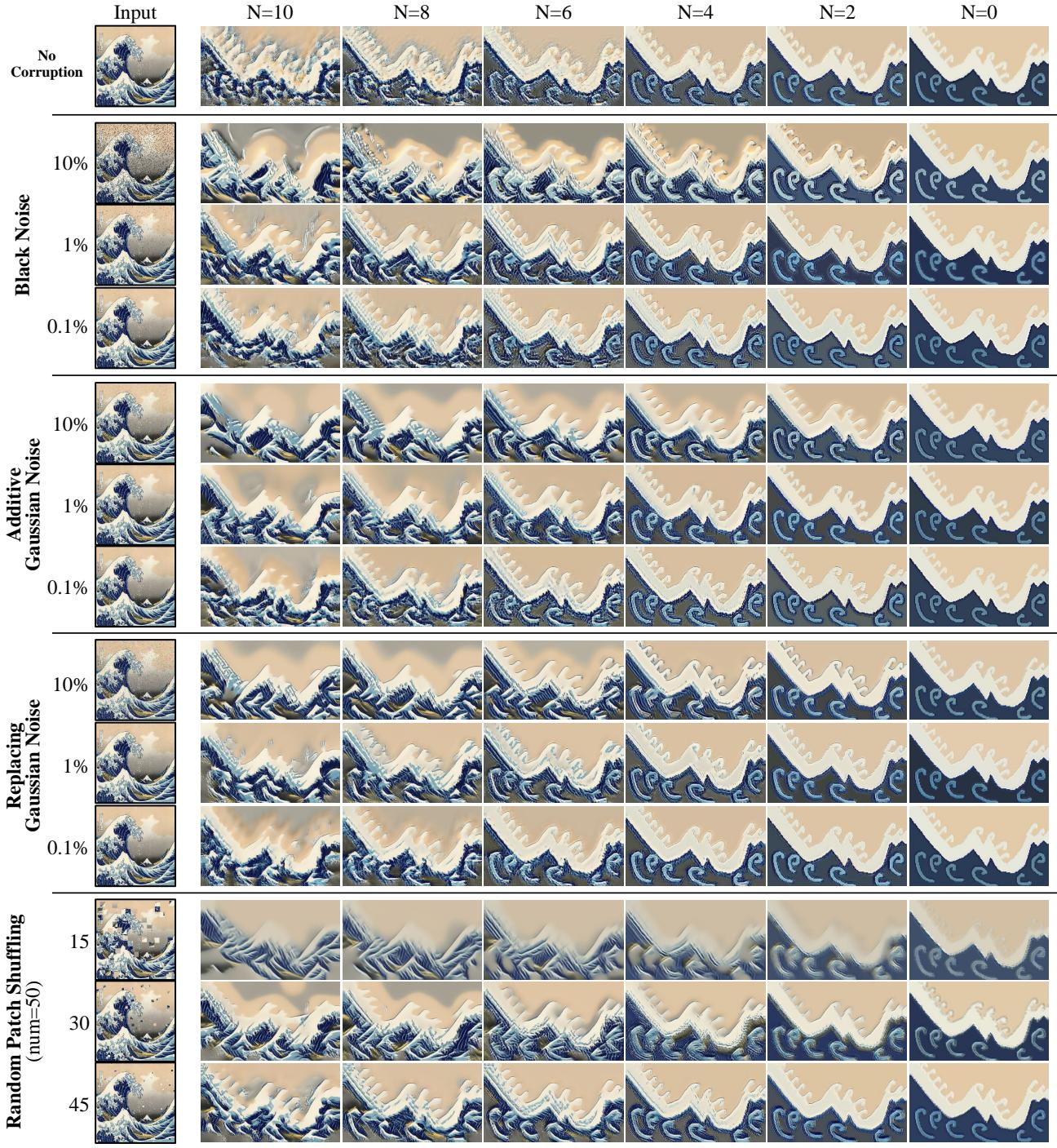


Figure 14. Editing.



Figure 15. **Harmonization.** Deep Painterly Harmonization (Luan et al., 2018) is a dedicated method.



**Figure 16. Effect of different corrupting noise.** We explore 4 types of corruption. The numbers at the top indicate inference starting scales. Please see Section 2.1 of the original paper. The numbers on the left-hand side of inputs indicate the intensity of corruption. For random patch shuffling, they are the ratio of (a longer side of the original image) to (a side of a patch). For example, in case of the 30, (a longer side of the original image) / 30 = (a side of a patch). For other corruption schemes, the numbers are the percentage of randomly sampled pixels to be corrupted. Please see Section C for details.

## C. Effect of Different Corrupting Scheme

Although we use only *random pixel shuffling* in the original paper, alternative corrupting schemes can be considered. Figure 16 illustrates the effect of different corruption. Here we explore 4 types of corruption. From Figure 16, we can see that applying mild corruption to the input of SinIR generally gives better results regardless of corruption schemes, compared to those of no corruption (the first row in Figure 16). Also, consistent with the findings discussed in Section 2.1.2 of the original paper, regardless of the corrupting schemes, when the intensity of the corruption becomes high, the results become smoothed and vice versa.

**Black Noise** *Force randomly sampled pixels to be completely black.* This corrupting scheme is originally used in denoising autoencoder (Vincent et al., 2008) with MNIST dataset (Bengio et al., 2006; Lecun et al., 1998). Compared to other corrupting schemes, this corruption sometimes produces unnatural textures (*e.g.*, floating objects with black edges in Figure 16). This is probably because we are using natural images, not the MNIST dataset. Considering that natural images often include more complex structures, simply *turning off* pixels ignores such visual properties and may prevent learning better relationship between adjacent pixels.

**Additive Gaussian Noise** *Add gaussian noise to randomly sampled pixels.* For gaussian noise, we set mean and variance to 0 and 0.5 with the pixel value of [-1, 1]. The corrupted pixel values are clipped at -1 and 1. Compared to *Black Noise*, this corrupting scheme generally produces better results that are close to random pixel shuffling that is used in the original paper. A possible reason is that now we are corrupting the input based on its original pixel values.

**Replacing Gaussian Noise** *Replace randomly sampled pixels with gaussian noise.* For gaussian noise, we set mean and variance to 0 and 0.5 with the pixel value of [-1, 1]. The corrupted pixel values are clipped at -1 and 1. The results are similar to *Additive Gaussian*, but it sometimes produces unnatural objects as *Black Noise* does. It is probably because both of them corrupt pixels not based on its original values.

**Random Patch Shuffling** *Shuffle randomly sampled patches.* We shuffle 50 patches for Figure 16. As we directly use internal patches, this scheme produces well-textured results (*e.g.* wave bubbles). These results are closest to those of *random pixel shuffling* used in the original paper.