Unpaired Real-World Super-Resolution with Pseudo Controllable Restoration

Andrés Romero[1] Luc Van Gool[1,2] Radu Timofte[1,3]
[1]Computer Vision Lab, ETH Zürich [2]KU Leuven [3]University of Würzburg
{roandres, vangool, timofter}@vision.ee.ethz.ch

Abstract

Current super-resolution methods rely on the bicubic down-sampling assumption in order to develop the ill-posed reconstruction of the low-resolution image. Not surprisingly, these approaches fail when using real-world low-resolution images due to the presence of artifacts and intrinsic noise absent in the bicubic setup. Consequently, attention is increasingly paid to techniques that alleviate this problem and super-resolve real-world images. As acquiring paired real-world datasets is a challenging problem, real-world super-resolution solutions are traditionally tackled as a blind problem or as an unpaired data-driven problem. The former makes assumptions about the down-sampling operations, the latter uses unpaired training to learn the real distributions. Recently, blind approaches have dominated this problem by assuming a diverse bank of degradations, whereas the unpaired solutions have shown under-performance due to the two-staged training. In this paper, we propose an unpaired real-world super-resolution method that performs on par, or even better than blind paired approaches by introducing a pseudo-controllable restoration module in a fully end-to-end system.

1. Introduction

Super Resolution (SR) is the task of recovering the sharp and detailed information of a high-resolution (HR) image from its low-resolution (LR) counterpart. Although SR is a very active topic [13–15, 27, 34, 35], its usage in realistic applications - Real-World Super Resolution (RWSR) - is in its infancy. RWSR differs from traditional Super-Resolution (SR) approaches, also known as Single-Image Super-Resolution (SISR), in a subtle yet radical way. On the one hand, SISR methods are trained in a paired fashion by generating downsampled (typically bicubic interpolation) LR images from their HR counterparts, thus resolving a simplified version of the challenging ill-posed SR problem, *i.e*. increasing the resolution of a noise-free, artifact-free, and corrupted-free low-resolution image. On the other hand, RWSR methods aim at super-resolving real-world images, that is, images with noise, sensor corruptions, and compression artifacts. Due to the inherent data acquisition, it is difficult to have LR-HR alignment.

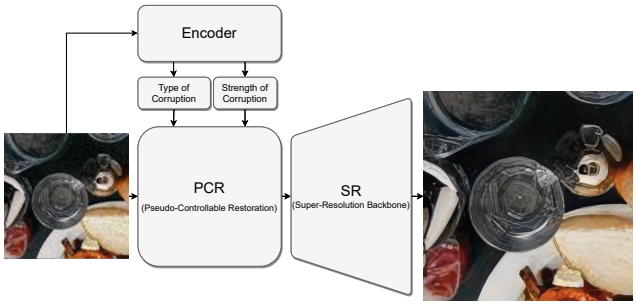

Figure 1. **PCR-ESRGAN –** Our solution first self-estimates a notion of the type of noise and strength corruptions in the low-resolution image, so it can be clean by a domain adaptor generator, which output is subsequently fed to a Super-Resolution backbone for upsampling.

Given the unsupervised nature of this problem, most RWSR methods [7, 17–19, 29] rely on a two-stage approach. Firstly, learning to degrade the HR images in order to resemble the noise and corruptions in the LR domain. Secondly, artificially generate LR-HR pairs in order to learn a new SR method with such paired supervision. There exist different alternatives to model the degradation process, such as using fixed prior assumptions regarding the degradation kernel [11, 36], or using Generative Adversarial Networks [8] to learn a generator without prior assumption on the data or the kernel [2, 17]. Nonetheless, two-stage approaches for real-world super-resolution share an important issue: as they lack end-to-end training, and given that there is no LR-HR alignment, learning the degradation model involves heuristic decisions on the model convergence, which causes discrepancies with the real LR domain. Moreover, as the super-resolution alignment in the second stage relies on the degradation procedure, this heuristic decision and domain gap dramatically affects the performance at test time.

There has been little attention to one-stage approaches that aim at jointly learning the degradation and the up-

sampling operations in an end-to-end manner. Recent efforts [20, 30] in this direction extended the two domain learning (real-world corrupted domain and high-resolution domain) by adding a new low-resolution clean domain that leverages on the downsampling of the high-resolution image, which then uses it to first learn to "clean" the real-world image before super-resolving it. However, this approach makes a strong assumption in the new clean domain, namely that cleaning a corrupted image is a deterministic operation.

Our solution builds on top of the fully end-to-end idea, yet introduces a notion of stochasticity in the transformations. As a consequence, it produces a relatively diverse super-resolved output. Our rationale is that cleaning an image is not a deterministic process, instead depending on the type and amount of noise. Therefore, we first aim at estimating a pseudo-noise and a pseudo-strength from the low-resolution image, which act as condition to a clean generator. Subsequently, we use an off-the-shelf SISR backbone to upsample our clean image (see Figure 1).

Specifically, we propose a pseudo-controllable restoration (PCR) technique that controls how much of the corrupted image should be restored before passing it to the SR network. Additionally, both the clean domain and the corrupted domain are modeled as non-deterministic, which allows us to perform more accurate cycle-consistency reconstructions, which are crucial in unpaired settings [37]. Our results suggest that introducing more control in the unpaired training outperforms the state-of-the-art in RWSR [7, 11, 29], and produces very competitive results with respect to blind approaches [26, 32].

See a simplified overview of the PCR training system in Figure 2. *Code and pre-trained models will be available.*

2. Related Work

SISR techniques where the downsampling operation is known are the most popular choice in the SR landscape. However, it lacks generalization due to the domain gap between real-world LR images and LR images used for training (typically clean bicubically downsampled images). There are two alternatives to deal with this mismatch. Firstly, blind techniques focus on learning a generalized downsampling representation (*i.e*. blur kernel) in order to impose a more realistic downscaling operation, which can be applied for super-resolution beyond the bicubic degradation. Secondly, unpaired real-world problems focus on data-driven solutions, *i.e*. super-resolving low-resolution images that lie inside a distribution (typically using a dataset that comes from a small/corrupted device sensor).

2.1. Blind SR

Despite their success, paired SISR methods [13, 27] fail when applied to LR images from a source different than that used for training, which hinders their applicability to more realistic applications. As a result, trying to model more realistic downsampling operators, Michaeli and Irani [21] exploited the internal dependencies within an image to extract the most optimal non-parametric kernel from the low-resolution image. This idea is further developed by Shocher *et al*. [24] and Bell-Kligler *et al*. [1], leveraging the recurrent information within an image to develop an image-specific network using different blur kernels. While there exist several methods in blind SR [9, 22, 25, 31], they mostly focus on learning the blur kernel and do not consider compression artifacts or sensor corruptions. Recently, a new family of blind approaches [26, 32] focuses on producing a high variety of random degradations that resemble the ones present in images in the wild, producing impressive results for RWSR. Despite their success, they require heavy manual tuning of the degradation parameters.

2.2. Unpaired SR

Unpaired real-world super-resolution has been traditionally tackled by using an ensemble of two approaches: learn to degrade in order to learn how to upsample, where the main technical improvements are usually in the degradation stage. We categorize them into two strands.

The first strand employs end-to-end training schemes. CinCGAN [30] pioneered this problem by using an end-to-end approach, which extends the two involved domains (real-world LR and HR) by adding an additional LR-clean (cleanLR) domain, where the main purpose is to remove the artifacts and noises before super-resolving the image. Similarly, Maeda [20] further extended this idea by using a pseudo-supervision HR-cleanLR-LR-cleanLR-HR that yielded better performance. As these approaches require learning 3 different generators, *i.e*. to degrade, to clean, and to upsample in an unsupervised non-fully stochastic manner, the framework is very unstable.

The second strand decouples the degradation and the super-resolution as two independent problems. Bulat *et al*. [2] and Lugmayr *et al*. [17] leveraged on a cycle-consistency approach [37] to learn the real-world generator in a fully adversarial way without prior assumptions, which is consequently used to produce LR-HR pairs to learn a specialized SR network. As real-world corruptions can be naively modeled as high-frequency perturbations, the winner of the AIM19 Real-world SR challenge [19], ESRGAN-FS [7], extended the work in [17] to focus the corrupted images discriminator on the high-frequency spectrum. Following a different trend, several methods [11, 36] borrowed insights from blind SR systems to introduce a kernel degradation pooling, which set Ji *et al*. [11] as the winner of the NTIRE 2020 - Real-World SR Challenge [18]. However, this solution employs an empirical and handcrafted non-parametric kernel pooling which limits its scalability. Although these two-staged solutions yield impressive re-

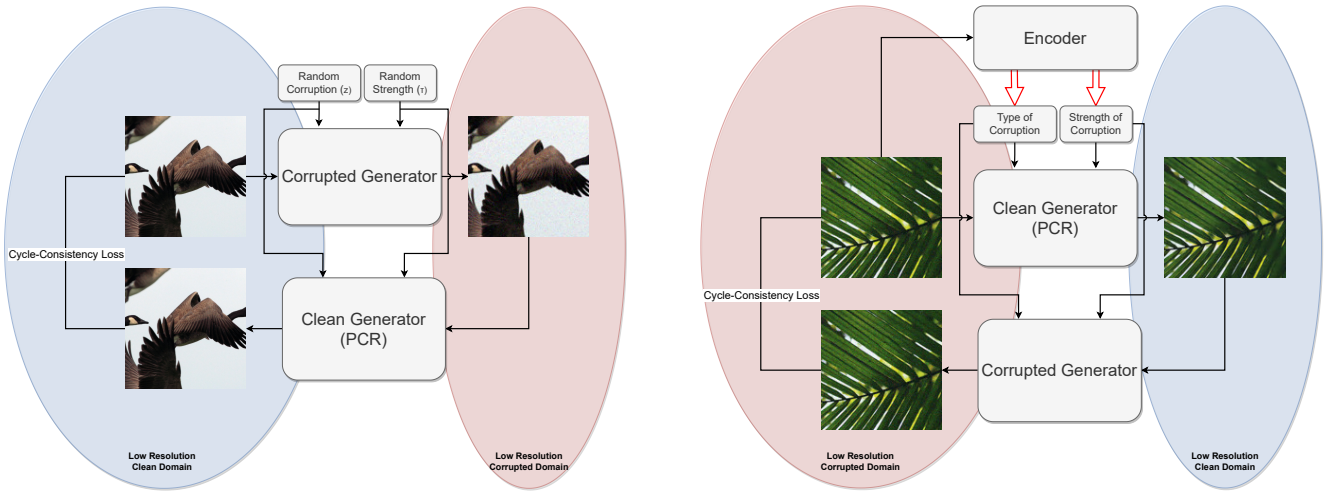

Figure 2. **Overview of our training scheme –** We introduce a pseudo-controllable restoration scheme, which for the cycle-consistency of the clean images (*left*), we sample a random noise and a random strenght for the corrupted generation stage. This fake corrupted image is subsequently cleaned using the same previously injected noise values. For the corrupted images (*right*), we resort to a trainable controllable encoder that extracts the type of corruptions from the low-resolution image, and use it as pseudo-guidance for both generation and reconstruction. Note that $\Downarrow$ refers to detaching the gradients from the graph.

sults and are overall more stable to train than end-to-end approaches, they rely on several heuristic decisions for the degradation learning scheme, which harms generalization. Moreover, using a degradation network to model RWSR corruptions independently introduces a domain shift [17]. To overcome the domain shift, Wei *et al*. [29] propose a domain-gap aware weight to balance this information during the SR training.

Both Blind SR and Unpaired SR approaches share a strong assumption, which is the determinism in the SR transformation. It is common to assume non-deterministic transformations for the corrupted domain, yet not so for the "clean" one. Our solution goes in this direction. We propose a non-deterministic transformation as an intermediate step to pseudo-clean the LR image, before feeding it to off-the-shelf SR networks, *e.g*. ESRGAN [27]. Importantly, we produce two different solutions: a fully end-to-end trainable system PCR-ESRGAN, and a Plug&Play (PP-ESRGAN) cleaning module that adapts to any existing bicubic SR network, (*e.g*. RCAN [35], ESRGAN [27], and RankSRGAN [34]) for the purpose of real-world unpaired scenery.

3. Proposed Approach

Our goal is to learn a mapping function $\mathbb{G}$ that reconstructs a HR image $\mathcal{Y}_{hr} \in \mathbb{R}^{H\times W\times 3}$ by using the real-world LR counterpart $\mathcal{X}_{lr} \in \mathbb{R}^{h\times w\times 3}$ without having access to the joint distribution $(\mathcal{X}_{lr}, \mathcal{X}_{hr})$, where $(h, w) = (H/n, W/n)$ and n is a predefined down-scaling factor. Using real-world LR images implies that traditional SR methods [27, 34] do not perform well on this task due to the inherent artifacts and noise absent in such training frameworks. With this in mind, and in the same vein as recent works [20, 30], we use an unsupervised generative adversarial approach where we learn to degrade and clean HR and LR images, respectively. Importantly, we are leveraging off-the-shelf SR techniques to learn a domain adaptor (clean network) model that produces clean images suitable for such SR solutions. In addition to the off-the-shelf network, the proposed system is coupled with 2 more generators, 3 discriminators, and 1 controllable encoder. We depict a simplified overview of our system in Figure 2, which we explain in the following section.

3.1. Networks

Generators. In order to create HR images from real-world LR ones in an unsupervised fashion, we define a domain adaptation generator $\mathbb{G}_{clean}$ that aims at removing the corruptions and artifacts present in the LR images (x_{lr}), while maintaining the same resolution (y_{lr}). As our system is fully unsupervised, we also employ a corrupted generator $\mathbb{G}_{corrupted}$ that transforms from the bicubically downsampled clean image ($y_{hr}\downarrow_s$ - s is the downscaling factor) domain to the real-world domain.

Additionally, as the generic off-the-shelf network ($\mathbb{G}_{up}$) expects clean LR images in order to perform the upsampling, our mapping function $\mathbb{G}$ is redefined as $\mathbb{G} = \mathbb{G}_{up} \circ \mathbb{G}_{clean}$. There are two alternatives to our solution: in the plug & play version (PP-ESRGAN), $\mathbb{G}_{up}$ is not trained, and

instead it uses pretrained SR weights. The second alternative is the full retraining of $\mathbb{G}_{up}$, which corresponds to our main system PCR-ESRGAN.

Moreover, the corrupted generator and the clean generator are conditioned with a random variable to insert ($\hat{z}$) or remove corruptions ($\tilde{z}$), respectively, in order to produce a plurality of outputs. In detail, $\hat{z}$ aims at modeling the resulting type of degradation the corrupted generator should produce, and $\tilde{z}$ aims at modeling what kind of degradation the clean generator should remove. Moreover, to also model a notion of strength, we include a temperature variable that scales $\hat{z}$ and $\tilde{z}$, independently. Formally, the output of each generator is defined as follows:

$$\hat{x}_{lr} = \mathbb{G}_{corrupted}(y_{lr}, z), \text{ where } z = \hat{z} \cdot \hat{\tau} \tag{1}$$

$$\tilde{y}_{lr} = \mathbb{G}_{clean}(x_{lr}, z), \text{ where } z = \tilde{z} \cdot \tilde{\tau} \tag{2}$$

Controllable Encoder. During $\hat{z}$ and $\tilde{z}$ training, $\hat{\tau}$ and $\tilde{\tau}$ are randomly sampled from a normal distribution and a uniform distribution, respectively. Therefore, inspired by recent works in general image manipulation [5, 23], the controllable encoder serves as an estimator for both the type (z) and strength (τ) of corruptions. In detail, the controllable encoder has two output branches for type ($\mathbb{S}_z$) and strength ($\mathbb{S}_\tau$), respectively.

3.2. Loss Functions

At each iteration, we assume we have access to a disjoint pair of HR (y_{hr}) and LR images (x_{lr}), and by bicubically downsampling y_{hr} we obtain y_{lr}. Additionally, fake images (Equation 1 and 2) are generated at each iteration using normally distributed noise ($\hat{z}, \tilde{z} \in \mathbb{R}^{w \times h}$) and a uniformly sampled strength ($\hat{\tau}, \tilde{\tau} \in \{1, 2, \ \ldots, N-1\}$, and N is empirically set to 10).

Noise Loss. As we are interested in estimating the amount of corruption (either to remove or to inject) present in an image, from Equation 1, we aim at reconstructing the type of noise injected to the corrupted generator, as it is more difficult to estimate the amount of corruptions removed by Equation 2:

$$\mathcal{L}_{noise} = \|\hat{z} - \mathbb{S}_z(\hat{x}_{lr})\|_1 \tag{3}$$

Note that z is the same size as the input image, and the noise is modeled as pixel-independent. This assumption has a particularity that allows us to introduce our next loss function.

Strength Loss. The Noise Loss allows the model to learn a distribution over the possible noises, corruptions, and artifacts. In order to extend the diversity of the search space, we use τ different scaled Gaussian distributions (Please refer to the Supplementary Material for a visual description), where each pixel in z is randomly scaled. While the noise loss aims at always reconstructing the noise at a *unitarian* strength, we define the strength loss as the probability each pixel has to belong to the scaled distribution.

We use the cross-entropy loss at the pixel level as follows:

$$\mathcal{L}_{temp} = -p(\hat{x}_{lr}) \cdot \log \mathbb{S}_\tau(\hat{x}_{lr}) \tag{4}$$

Identity Loss. To further guide the clean network, we deploy the identity loss using a zeroed strength component as:

$$\mathcal{L}_{idt} = \|y_{lr} - \mathbb{G}_{clean}(y_{lr}, [0]_{h \times w})\|_1 \tag{5}$$

Our rationale is that clean images do not require any further cleaning.

Reconstruction Loss. To enforce that the networks do not ignore the noise and strength cues, we enforce the cycle-consistency reconstruction loss. For the clean image (y_{lr}) case (Equation 1): as the corrupted generator injects the resulting type of distribution ($\hat{z} \cdot \hat{\tau}$), and the clean generator feeds the corrupted distribution to be removed, then the same distribution ($\hat{z} \cdot \hat{\tau}$) serves in order to reconstruct the original image.

$$y_{rec-lr} = \mathbb{G}_{clean}(\hat{x}_{lr}, z), \text{ where } z = \hat{z} \cdot \hat{\tau}$$

The treatment for Equation 2 is different. In this case, we want to fully remove the corruptions in $\tilde{y}_{lr}$, and insert them again for the cycle-consistency. Yet, we do not have access to such information. Therefore, we use pseudo ground-truth that estimates the "real" corruptions in x_{lr}. We redefine Equation 2 with the following conditions: $\tilde{z} = sg(\mathbb{S}_z(x_{lr}))$ and $\tilde{\tau} = sg(\arg\max_\tau(\mathbb{S}_\tau(x_{lr})))$, where $sg(\cdot)$ denotes the *stop-gradient* operation, so we can pseudo-remove corruptions and pseudo-include them again for the loss function.

$$x_{rec-lr} = \mathbb{G}_{corrupted}(\tilde{y}_{lr}, z), \text{ where } z = \tilde{z} \cdot \tilde{\tau}$$

The reconstruction loss is defined as:

$$\mathcal{L}_{rec-lr} = \|x_{lr} - x_{rec-lr}\|_1 + \|y_{lr} - y_{rec-lr}\|_1 \tag{6}$$

Furthermore, we also use the reconstruction loss for the high-resolution cycle-reconstructed image.

$$\mathcal{L}_{rec-hr} = \|y_{hr} - \mathbb{G}_{up}(y_{rec-lr})\|_1 \tag{7}$$

In addition to the standard pixel-wise loss and perceptual VGG19 loss [12] for the high-resolution image, we introduce a novel perceptual loss that uses a stronger prior. We use VQGAN [6] pretrained with Gumbel Quantization [16]

on Imagenet, which allows us to apply a general-purpose feature-based loss function. We refer to it as VQLoss.

Total Loss. In addition to the adversarial losses for each domain (HR domain, LR clean domain, and LR real-world domain), our total loss is a weighted linear combination of the above-mentioned losses, where each loss contributes equally.

Inference. During inference, we use the controllable encoder to self-estimate the type and strength of corruptions in the low-resolution image: $\mathbb{G}_{up} \circ \mathbb{G}_{clean}(x_{lr}, z)$, where $z = \mathbb{S}_z(x_{lr}) \cdot \arg\max_\tau(\mathbb{S}_\tau(x_{lr}))$.

3.3. VQGAN as Prior

We observed that the VQGAN [6] internal representation is robust enough to successfully reconstruct an LR blurry/corrupted/degenerated image as well as a pristine high-resolution image, which makes the VQGAN encoder very suitable as a powerful prior for downstream tasks. In this paper, we confirm this for perceptual SR tasks. In addition to the standard VGG as *de facto* perceptual loss, we use the proposed VQLoss as follows. Similar to a Feature Matching loss, we use the L1 loss for the deep feature representations of a pretrained Gumbel Quantized VQGAN encoder (E):

$$\mathcal{L}_{\text{VQ-FM}} = \sum_{i=1}^{T} \frac{1}{N_i}[||E^{(i)}(y_{hr}) - E^{(i)}(\mathbb{G}_{up}(y_{rec-lr}))||_1],$$

Where T are the number of layers in the encoder.

Moreover, as each ground-truth image can be quantized $q(\cdot)$ as a discrete vocabulary from a rich codebook $q(E(y_{hr}))$, we use the standard cross-entropy loss ($\mathcal{L}_{\text{VQ-CE}}$) to force the generated image $E(\mathbb{G}_{up}(y_{rec-lr}))$ to also encode for the same vocabulary. Therefore, the final $\mathcal{L}_{\text{VQ}}$ loss is a weighed sum between $\mathcal{L}_{\text{VQ-FM}}$ and $\mathcal{L}_{\text{VQ-CE}}$.

Note that $\mathcal{L}_{\text{VQ}}$ is not inherent to unpaired SR tasks, yet it can be extended for perceptual SISR, Blind SR, Video SR, and it can be potentially used outside SR problems. Please refer to the Supplementary Material for more details, and comprehensive results of $\mathcal{L}_{\text{VQ}}$ applied to SISR, Blind SR, and Video SR pre-existing solutions.

4. Experimental Setup

4.1. Datasets

We use the datasets provided in two different challenges in Real-World Super-Resolution: NTIRE20 [18] and AIM19 [19]. Additionally, we also use two datasets with less visual artifacts, yet real-world camera and cellphone corruptions: RealSR [3] modeling DLSR camera corruptions, and DPED [10] modeling cellphone corruptions. In all cases we use $\times 4$ upsampling.

4.2. Evaluation Framework

Since AIM19, NTIRE20, and RealSR datasets provide a paired validation set, we to compute the learned perceptual image patch similarity (LPIPS) metric, as well as, fidelity metrics such as the peak signal-to-noise ratio (PSNR), and the structural similarity index (SSIM) [28].

4.3. Training Details

We train our system during 100,000 iterations with a constant learning rate of 0.001, batch size 4, and high-resolution crops of 256×256 pixels. In order to keep our system easily reproducible and architecture agnostic, we use pre-existent architectures for all our generators, discriminators, and controllable encoder. For instance, both the clean and corrupted generator is a light-weighted version of an RRDB [27] network with no upsampling layers. Please refer to the Supplementary Material for more training and architecture details.

5. Results

5.1. Ablation Study

To validate each section of our system, and considering that our solution is inspired by the one presented by Maeda [20], we select Maeda as our baseline. We study our three main contributions, namely the pseudo-ground-truth for cycle-consistency, the controllable strength, and the effect of an additional perceptual loss. By using a generic pretrained SR network (ESRGAN [34]) as upsampling network, Table 1 shows the quantitative evaluation of each part of our system over the AIM19 dataset.

Although Maeda source code is not publicly available, our baseline ⓪ behaves similarly to the scores reported in the main paper [20] for AIM19, and it overall has similar ingredients. ①, Our first experiment (PP-ESRGAN) is to exploit the unsupervised nature of the problem in order to force the clean generator to produce bicubically downsampled clean images that can be super-resolved by an off-the-shelf pretrained yet frozen SR network *e.g.*, RCAN [35], ESRGAN [27], and RankSRGAN [34]. This model can be used as a plug & play clean module during inference. In detail, we train the whole system by leveraging on the gradients of a fixed SR Network, and during inference, we can test over all possible upsampling networks. Please refer to the Supplementary Material for more results over different architectures. Using a P&P approach is actually very beneficial with respect to the baseline.

Different to prior methods, Pseudo-GT ② assumes that the clean domain is inherently stochastic due to the multiple representations for the clean generator. This prevents us to extend the P&P approach that has a deterministic assumption. By introducing our Pseudo-GT, we get an important boost over the performance. Moreover, ③, scaling (τ) the

		Finetuning	Pseudo-GT	Strength	VQLoss	PSNR↑	SSIM↑	LPIPS↓
Baseline	⓪ →	✓				19.500	0.529	0.442
PP-ESRGAN	① →					21.029	0.552	0.406
	② →	✓	✓			20.213	0.538	0.364
	③ →	✓	✓	✓		21.492	0.592	0.334
PCR-ESRGAN	④ →	✓	✓	✓	✓	21.590	0.610	0.321

Table 1. **Ablation study –** We validate our system using three different components and compared to our baseline that shares similarities with Maeda [20]. *Finetuning* refers to updating the weights of the upsampling network that starts from a pretrained SISR network.

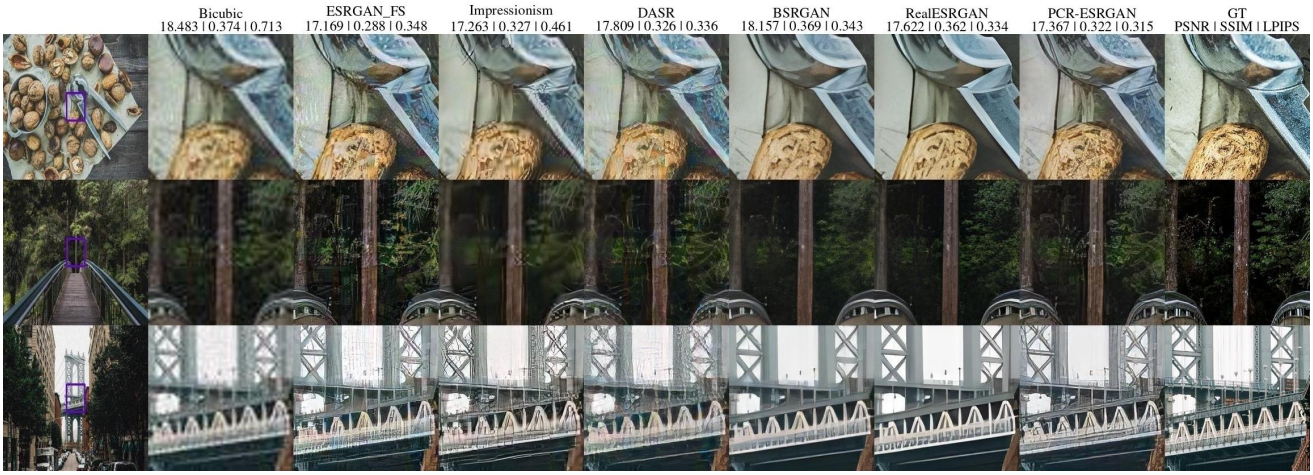

Figure 3. **AIM19 [19] Qualitative Results –** We compare against unsupervised approaches (ESRGAN-FS [7], Impresionism [11], DASR [29]), and two blind approaches (BSRGAN [32] and Real-ESRGAN [26]). We compute the averaged PSNR, SSIM, and LPIPS for the three displayed images. Note that in contrast to blind approaches, our solution produce consistent high-frequency details.

noise distribution allows us to have a notion of controllable strength that also strengthen our performance.

Finally ④, we observed that using a powerful prior as a loss can help to yield better perceptual results in the RWSR task. We further evaluate this observation over different strategies, *i.e.*, retraining ESRGAN [27], RealESRGAN [26], and BasicVSR [4] in a GAN setup with VQLoss. In all cases, there is a consistent improvement in both pixel-wise metrics (PSNR and SSIM) and perceptual metric (LPIPS). Please refer to the Supplementary Material for this evaluation.

Furthermore, we empirically found that the PCR training scheme does not further contribute to the PP-ESRGAN scheme.

5.2. Comparison with State-Of-The-Art

We compare our system against the two winners of the last two editions of the Real-World Super-Resolution Challenge (ESRGAN-FS [7] for AIM19 [19] and Impressionism [11] for Ntire20 [18]), DASR [29], and two [26, 32] blind approaches.

Table 2 shows quantitative comparison over three different datasets. Our method, PCR-ESRGAN, consistently outperforms the unpaired solutions in both AIM19 and NTIRE20 datasets, and it even performs better than blind approaches on the NTIRE20 dataset. However, for the RealSR dataset, our solution is not champion. In contrast to AIM19 and NTIRE20, the RealSR dataset is built by using similar DSLR camera sensors for both LR and HR datasets with different focal lengths, which leads to very subtle corruptions on the LR counterpart. We argue that our method is more suitable for those datasets that visually require a clean adaptation module before super-resolving, hence failing in those that do not require it. Additionally, we see the DASR method as an orthogonal solution to this problem, as they introduce a domain-gap aware training scheme plus wavelet bands for the discriminator, which can also be incorporated in our solution.

Figure 3, 4, and 5 show qualitative results for AIM19 [19], NTIRE20 [18], RealSR [3], and DPED [10] datasets. Note that Real-ESRGAN and BSRGAN achieve impressive performance across the three datasets. However, their generated images tend to be over-smoothed in the high-frecuency regions. Our solution, and in general

	Method	AIM19 [19]			NTIRE20 [18]			RealSR [3]		
		PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
	Bicubic	22.360	0.618	0.683	25.520	0.673	0.634	25.620	0.729	0.481
Blind	BSRGAN [32]	22.475	0.623	0.299	24.563	0.669	0.265	24.878	0.739	**0.268**
	Real-ESRGAN [26]	22.190	0.624	**0.296**	24.678	0.687	**0.251**	24.310	0.737	0.273
Unsupervised	ESRGAN-FS [7]	20.832	0.511	0.390	24.599	0.689	0.253	23.992	0.709	0.295
	Impressionism [11]	21.901	0.600	0.411	24.831	0.662	0.228	22.453	0.639	0.309
	DASR [29]	21.788	0.573	0.347	-	-	-	25.786	0.753	**0.267**
	Maeda [20]	19.500	0.529	0.442	21.109	0.596	0.321	21.770	0.637	0.336
	PP-ESRGAN (**Ours**)	21.029	0.552	0.406	23.984	0.662	0.267	21.781	0.667	0.331
	PCR-ESRGAN (**Ours**)	21.590	0.610	**0.321**	24.970	0.682	**0.223**	25.174	0.714	0.305

Table 2. **Quantitative Comparison –**. We compare against recent works using unsupervised training schemes as well as blind approaches. We compute both fidelity (PSNR and SSIM) and perceptual (LPIPS [33]) metrics for three different datasets AIM19 [19], NTIRE20 [18], and RealSR [3].

Figure 4. **NTIRE20 [18] Qualitative Results –** We compare against unsupervised approaches (ESRGAN-FS [7], Impresionism [11], and two blind approaches (BSRGAN [32] and Real-ESRGAN [26]). We compute the averaged PSNR, SSIM, and LPIPS for the three displayed images. Similar to the AIM19 case, blind approaches tend to over-smooth high frequency regions.

unpaired solutions, are overall sharper in those regions.

Furthermore, it is important to mention that both ESRGAN-FS and Impressionism require special hand-crafted tuning during the first stage of corruption generation. Visual analysis on the type of noise leads to hard-coded assumptions, which can be very difficult to scale to different datasets with different artifacts. Our method has no assumption about the data and it is trained in an end-to-end manner.

Although the PP-ESRGAN performance is far from the PCR-ESRGAN one, it proves, to some extent, that bicubic SISR networks can be incorporated into both training and inference stages of real-world super-resolution systems.

Overall, blind approaches do not widely outperform unsupervised approaches. In detail, for a given LR dataset built from a particular sensor yet unknown corruptions and artifacts, results on NTIRE20 suggest that it might be beneficial to use unpaired learning.

5.3. Note on Controllable Restoration

Figure 6 shows results using different values for both type of noise and strengths, *i.e.*, using different random vectors (z) and varying τ. Note that, as we are using a pseudo-controllable approach, we are not directly estimating the real type of noise and the real strength from the LR image, which result in not having a zero-to-one restoration scheme,

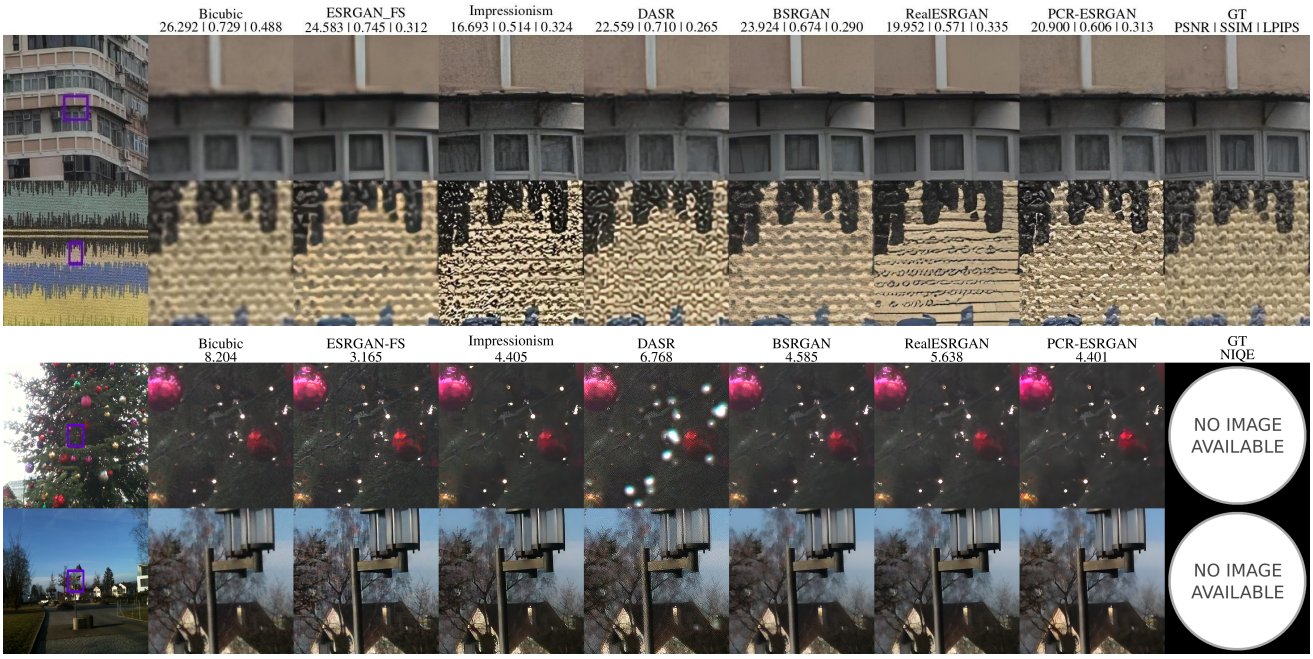

Figure 5. **RealSR [3] (*top*) and DPED [10] (*bottom*) Qualitative Results –** We compare against unsupervised approaches (ESRGAN-FS [7], Impresionism [11], DASR [29]), and two blind approaches (BSRGAN [32] and Real-ESRGAN [26]). We compute the averaged PSNR, SSIM, and LPIPS for RealSR, and NIQE for DPED. For both datasets, unsupervised approaches tend to preserve sharp details.

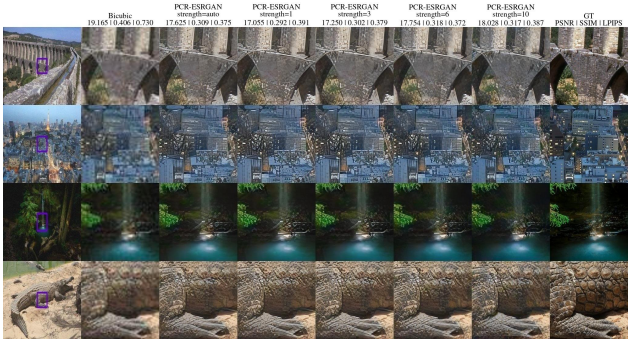

Figure 6. **PCR Diversity –** We show that using different values of strength or type of noise yield different transformations. As we are using a pseudo-controllable restoration, our strength does not correlate with the zero-to-one strength removal. Additionally, self-estimating (*auto*) the noise and strength does not always lead to the best performance.

and it can also be seen that sometimes random values can have better performance than the self-estimated one (third column).

5.4. Limitations

Although our method produces good performance on most of the datasets, it is still far from solving the real-world super-resolution problem, and we argue there are at least two areas where our method can be improved. (1) We introduced the concept of pseudo-controllable restoration. However, as Figure 6 shows, our method does not fully restore the image in a zero to one fashion, and instead, it is a notion of controllable restoration, hence the pseudo. (2) One strong assumption about our system is the cleaning strategy, *i.e.*, we require there are visible corruptions or artifacts to clean before the upsampling layers and it is not always the case in less corrupted real-world images such as RealSR.

5.5. Potential Negative Societal Impact

As a generative method, our solution is not excepted for misusing it in low-resolution images in-the-wild such as face or object hallucination. However, we hope the reader and final user of our code see our solution as a scientific research.

6. Conclusions

We have presented a method for Unpaired Real-World Super Resolution. We introduced a novel framework that leverages the full cycle-consistency of the corrupted images by using a pseudo-controllable restoration. The effectiveness of our framework is validated over the AIM19 dataset and generalized to NTIRE20, RealSR, and DPED datasets. Moreover, compared with blind approaches, we show that in some cases using unpaired learning leads to better performance.

Acknowledgments This work was support by Huawei, ETH Zurich General Fund and Humboldt Foundation.

References

[1] Sefi Bell-Kligler, Assaf Shocher, and Michal Irani. Blind super-resolution kernel estimation using an internal-gan. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 2

[2] Adrian Bulat, Jing Yang, and Georgios Tzimiropoulos. To learn image super-resolution, use a gan to learn how to do image degradation first. In *Proceedings of the European conference on computer vision (ECCV)*, pages 185–200, 2018. 1, 2

[3] Jianrui Cai, Hui Zeng, Hongwei Yong, Zisheng Cao, and Lei Zhang. Toward real-world single image super-resolution: A new benchmark and a new model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3086–3095, 2019. 5, 6, 7, 8

[4] Kelvin CK Chan, Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Basicvsr: The search for essential components in video super-resolution and beyond. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4947–4956, 2021. 6

[5] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. StarGAN v2: diverse image synthesis for multiple domains. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 4

[6] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12873–12883, 2021. 4, 5

[7] Manuel Fritsche, Shuhang Gu, and Radu Timofte. Frequency separation for real-world super-resolution. In *Proceedings of the IEEE International Conference on Computer Vision Workshop (ICCVW)*, pages 3599–3608. IEEE, 2019. 1, 2, 6, 7, 8

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2672–2680, 2014. 1

[9] Jinjin Gu, Hannan Lu, Wangmeng Zuo, and Chao Dong. Blind super-resolution with iterative kernel correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1604–1613, 2019. 2

[10] Andrey Ignatov, Nikolay Kobyshev, Radu Timofte, Kenneth Vanhoey, and Luc Van Gool. Dslr-quality photos on mobile devices with deep convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3277–3285, 2017. 5, 6, 8

[11] Xiaozhong Ji, Yun Cao, Ying Tai, Chengjie Wang, Jilin Li, and Feiyue Huang. Real-world super-resolution via kernel estimation and noise injection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 466–467, 2020. 1, 2, 6, 7, 8

[12] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 694–711. Springer, 2016. 4

[13] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017. 1, 2

[14] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1833–1844, 2021. 1

[15] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 136–144, 2017. 1

[16] Christos Louizos, Matthias Reisser, Tijmen Blankevoort, Efstratios Gavves, and Max Welling. Relaxed quantization for discretized neural networks. *arXiv preprint arXiv:1810.01875*, 2018. 4

[17] Andreas Lugmayr, Martin Danelljan, and Radu Timofte. Unsupervised learning for real-world super-resolution. In *Proceedings of the IEEE International Conference on Computer Vision Workshop (ICCVW)*, pages 3408–3416. IEEE, 2019. 1, 2, 3

[18] Andreas Lugmayr, Martin Danelljan, and Radu Timofte. Ntire 2020 challenge on real-world image super-resolution: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 494–495, 2020. 1, 2, 5, 6, 7

[19] Andreas Lugmayr, Martin Danelljan, Radu Timofte, Manuel Fritsche, Shuhang Gu, Kuldeep Purohit, Praveen Kandula, Maitreya Suin, AN Rajagoapalan, Nam Hyung Joon, et al. Aim 2019 challenge on real-world image super-resolution: Methods and results. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 3575–3583. IEEE, 2019. 1, 2, 5, 6, 7

[20] Shunta Maeda. Unpaired image super-resolution using pseudo-supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 291–300, 2020. 2, 3, 5, 6, 7

[21] Tomer Michaeli and Michal Irani. Nonparametric blind super-resolution. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 945–952, 2013. 2

[22] Dongwei Ren, Kai Zhang, Qilong Wang, Qinghua Hu, and Wangmeng Zuo. Neural blind deconvolution using deep priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[23] Andrés Romero, Luc Van Gool, and Radu Timofte. Smile: Semantically-guided multi-attribute image and layout editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1924–1933, 2021. 4

[24] Assaf Shocher, Nadav Cohen, and Michal Irani. "zero-shot" super-resolution using deep internal learning. In *Proceed-*

ings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3118–3126, 2018. 2

[25] Jae Woong Soh, Sunwoo Cho, and Nam Ik Cho. Meta-transfer learning for zero-shot super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[26] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1905–1914, 2021. 2, 6, 7, 8

[27] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018. 1, 2, 3, 5, 6

[28] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 5

[29] Yunxuan Wei, Shuhang Gu, Yawei Li, Radu Timofte, Longcun Jin, and Hengjie Song. Unsupervised real-world image super resolution via domain-distance aware training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13385–13394, 2021. 1, 2, 3, 6, 7, 8

[30] Yuan Yuan, Siyuan Liu, Jiawei Zhang, Yongbing Zhang, Chao Dong, and Liang Lin. Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 701–710, 2018. 2, 3

[31] Kai Zhang, Luc Van Gool, and Radu Timofte. Deep unfolding network for image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3217–3226, 2020. 2

[32] Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. *arXiv preprint arXiv:2103.14006*, 2021. 2, 6, 7, 8

[33] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pages 586–595, 2018. 7

[34] Wenlong Zhang, Yihao Liu, Chao Dong, and Yu Qiao. Ranksrgan: Generative adversarial networks with ranker for image super-resolution. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3096–3105, 2019. 1, 3, 5

[35] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 286–301, 2018. 1, 3, 5

[36] Ruofan Zhou and Sabine Susstrunk. Kernel modeling super-resolution on real low-resolution images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2433–2443, 2019. 1, 2

[37] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2223–2232, 2017. 2