Learning Degradation Uncertainty for Unsupervised Real-world Image Super-resolution

Qian Ning[1] , **Jingzhu Tang**[1] , **Fangfang Wu**[1] , **Weisheng Dong**[1*] , **Xin Li**[2] and **Guangming Shi**[1]
[1]Xidian University
[2]West Virginia University
ningqian@stu.xidian.edu.cn, tangjingzhu@stu.xidian.edu.cn, ffwu_xd@163.com,
wsdong@mail.xidian.edu.cn, xin.li@mail.wvu.edu, gmshi@xidian.edu.cn

Abstract

Acquiring degraded images with paired high-resolution (HR) images is often challenging, impeding the advance of image super-resolution in real-world applications. By generating realistic low-resolution (LR) images with degradation similar to that in real-world scenarios, simulated paired LR-HR data can be constructed for supervised training. However, most of the existing work ignores the degradation uncertainty of the generated realistic LR images, since only one LR image has been generated given an HR image. To address this weakness, we propose learning the degradation uncertainty of generated LR images and sampling multiple LR images from the learned LR image (mean) and degradation uncertainty (variance) and construct LR-HR pairs to train the super-resolution (SR) networks. Specifically, uncertainty can be learned by minimizing the proposed loss based on Kullback-Leibler (KL) divergence. Furthermore, the uncertainty in the feature domain is exploited by a novel perceptual loss; and we propose to calculate the adversarial loss from the gradient information in the SR stage for stable training performance and better visual quality. Experimental results on popular real-world datasets show that our proposed method has performed better than other unsupervised approaches.

1 Introduction

Single image super-resolution (SISR) aims to recover high-resolution (HR) images from their corresponding degraded low-resolution (LR) images. Following the pioneering study SRCNN [Dong *et al.*, 2014], many deep learning-based works employ convolutional neural networks in the SISR domain with promising results - e.g., Enhanced Deep residual networks for SR (EDSR) [Lim *et al.*, 2017], Enhanced SR Generative Adversarial Networks (ESRGAN) [Wang *et al.*, 2018], Residual Channel Attention Networks (RCAN) [Zhang *et al.*, 2018], deep unfolding network (MoG-DUN) [Ning *et al.*, 2021b].

Despite the extensive studies that have been published, most of those deep learning-based methods [Dong *et al.*, 2014; Lim *et al.*, 2017; Zhang *et al.*, 2018; Ning *et al.*, 2021b; Wang *et al.*, 2018] are trained by synthetic datasets such as bicubic degradation, suffering from the poor generation property of real-world datasets. To alleviate this situation, Cai *et al.* built a real-world dataset called RealSR [Cai *et al.*, 2019], consisting of LR-HR pairs by modifying the focal length of the cameras to obtain different resolution image pairs. Along this line of research, several works [Chen *et al.*, 2019; Ignatov *et al.*, 2017] have constructed new real-world datasets to meet the realistic requirement.

However, it is still cumbersome and labor intensive to obtain the corresponding LR-HR pairs on a large scale. More recently, some work [Fritsche *et al.*, 2019; Wei *et al.*, 2021; Son *et al.*, 2021] has proposed to train SR networks unsupervised by generating LR images from HR training images using unpaired datasets. Specifically, to obtain realistic LR images, a content loss with bicubic downsampled LR images has been used to maintain the same content, while a GAN loss with real-world LR images has been used to obtain realistic textures and details. Then, the SR or SRGAN networks can be trained by using the generated LR-HR pairs. FSSR [Fritsche *et al.*, 2019] proposed a frequency separation strategy to apply the GAN loss only to high-frequency components. Furthermore, DASR [Wei *et al.*, 2021] proposed reconstructing more realistic images by domain gap-sensitive training. Deflow [Wolf *et al.*, 2021] proposed to model the degradation process with conditional flows.

Although promising results have been achieved with these unsupervised approaches [Fritsche *et al.*, 2019; Wei *et al.*, 2021], these works only learned a single deterministic mapping model, completely ignoring the uncertainty of degradation in real LR images caused by the inevitable randomness of real degradation. Some recent work such as FSSR [Fritsche *et al.*, 2019] and DASR [Wei *et al.*, 2021] only generated one LR image instead of multiple different LR images with a single corresponding HR image, which is not in accordance with real-world situation considering more influence factors such as stochastic noise, etc. The question of how to generate multiple different but similar LR images from single HR images remains to be studied.

In this paper, we propose a novel approach called USR-DU for unsupervised real-world image SR with learned degrada-

*Corresponding Author

tion uncertainty. First, we propose learning realistic LR images and the corresponding degradation uncertainty simultaneously and sampling multiple LR images from the learned LR image (mean) and degradation uncertainty (variance). Then, any existing SR/SRGAN networks that require supervised training can be trained with newly constructed LR-HR pairs. Specifically, uncertainty can be learned by minimizing the proposed $\mathcal{L}_{kl}$ based on Kullback-Leibler (KL) divergence. Meanwhile, we propose to exploit the uncertainty in the feature domain, resulting in a novel perceptual loss function. Finally, we propose to calculate the adversarial loss from the gradient information in the SR stage for more stable training and better visual quality. We have conducted experiments on both RealSR [Cai *et al.*, 2019] and NTIRE2020 [Lugmayr *et al.*, 2020] real-world super-resolution challenging datasets. Experimental results on those real-world datasets show that our proposed method has performed better than other competing approaches based on unsupervised learning.

2 Related Work

2.1 Uncertainty in Computer Vision

The uncertainty of the data describes the noise inherent in observed data, which has been widely studied in computer vision. The performance and robustness of deep networks can be improved by modeling the uncertainty in the observation data [Kendall and Gal, 2017]. In [Chang *et al.*, 2020], they modeled the uncertainty of the data with estimated mean and variance in face recognition, achieving stronger performance in noisy training data. Ning *et al*. [Ning *et al.*, 2021a] proposed an adaptive weighted loss for SISR by assigning higher uncertainty pixels with higher weights during training. In common, those works mainly focus on the uncertainty of pair data in supervised training. Unlike those works, we propose to model the degradation uncertainty of unpaired data. With the estimated degradation uncertainty, multiple LR images can be sampled from learned LR images (mean) and uncertainty (variance) for training super-resolution (SR) networks.

2.2 Unsupervised Real-world Image SR

The difficulty of obtaining paired HR and degraded images has facilitated the advancement of unsupervised real-world image SR. FSSR [Fritsche *et al.*, 2019] and DASR [Wei *et al.*, 2021] proposed to first generate realistic LR images and train the SR network on generated paired data by domain-based adversarial loss. Those two works only learn a single deterministic mapping model without realizing the uncertainty of degradation. DeFlow [Wolf *et al.*, 2021] proposed learning the LR images with the conditional flow by an invertible network. Although multiple LR images can be sampled by giving different random initial Gaussian noises, complex degradation learning is constrained by the limited representation ability of the invertible network. DAP [Wang *et al.*, 2021a] proposed to align the degradation distribution in the feature domain with several regularization losses. However, DAP suffers from convergence and mode collapse issues and requires careful tuning of multiple losses. Different from those works, we propose learning the realistic LR images and degradation uncertainty simultaneously. Sampling from learned LR images (means) and uncertainty (variance), more robustly paired training data can be constructed, leading to better real SR performance.

3 Methodology

In unsupervised SR, only unpaired real LR and real HR images are provided for training. Let $\boldsymbol{X}$ represent the real HR images dataset and $\boldsymbol{x}$ denote samples. Similarly, the real LR image dataset can be expressed by $\boldsymbol{Y}_r$ and samples can be represented by $\boldsymbol{y}_r$. Note that the HR version of the corresponding LR image $\boldsymbol{y}_r$ is not provided. In unsupervised SR learning [Fritsche *et al.*, 2019; Wei *et al.*, 2021; Wolf *et al.*, 2021], the training process can be divided into two steps. The first step aims to generate realistic degraded LR images $\boldsymbol{y}_g$ from HR images $\boldsymbol{x}$. In this way, paired training data can be constructed and any existing SR networks requiring pair training can be trained. Therefore, the key to unsupervised SR lies in the generation of realistic LR images. In the first step, an LR version of the HR image $\boldsymbol{x}$ is needed to maintain the consistency of the content. In this paper, we simply adopt bicubic downsampling to generate the LR images denoted as $\boldsymbol{y}_b$. With the above conditions, the proposed unsupervised real-world image super-resolution approach (USR-DU) with learned degradation uncertainty will be introduced in this section. The framework of the proposed method is shown in Fig. 1.

In the next parts of this section, we first present how we train a downsampling network (DSN) to generate the realistic LR and estimate the degradation uncertainty. Then, we introduce our super-resolution network (SRN) training strategy with learned degradation uncertainty.

3.1 Learning Degradation Uncertainty in DSN

Learning degradation uncertainty in pixels domain. Generally, three types of losses, that is, content loss, perceptual loss, and GAN loss, will be used to generate realistic LR images during the training downsampling network. Content loss is used $\mathcal{L}_{1/2}$ to maintain content consistency in the downsampling step.

Taking the $\mathcal{L}_1$ loss as an example, the likelihood of a generated LR image can be formulated as

$$p(\boldsymbol{y}_g|\boldsymbol{x}, \boldsymbol{W}) = \exp(-\frac{||f^{(\boldsymbol{W})}(\boldsymbol{x}) - \boldsymbol{y}_b||_1}{c}), \qquad (1)$$

where $f^{(\boldsymbol{W})}(\cdot)$ denotes a parameterized downsampling network of $\boldsymbol{W}$ and c denotes the variance, which is a spatially invariant constant. In this way, the degradation process was modeled as a deterministic mapping, ignoring the uncertainty of degradation of the degraded LR images. This means that only a single degraded LR image $\boldsymbol{y}_g$ was generated when a clean HR image was given $\boldsymbol{x}$. To address this issue, we propose learning the degraded LR images (means) and the degradation uncertainty (variance) simultaneously. Our empirical study shows that the differences between the real LR and the bicubic downsampled LR images can be well characterized by a Laplacian distribution (while the Gaussian distribution

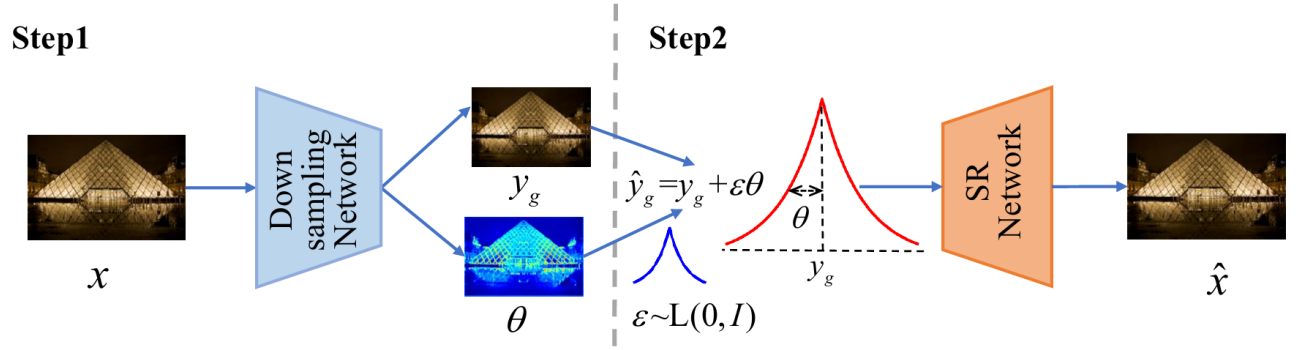

Figure 1: The framework of learning degradation uncertainty for unsupervised real image super-resolution. The whole training process can be divided into two steps. The first step estimates the uncertainty of degradation $\boldsymbol{\theta}$ (variance) of the learned LR images (means) in a downsampling network. In the second step, an HR image is paired with multiple LR images, which are sampled from the learned LR image (mean) and corresponding degradation uncertainty (variance). Those LR-HR pairs will be used to train the SR network.

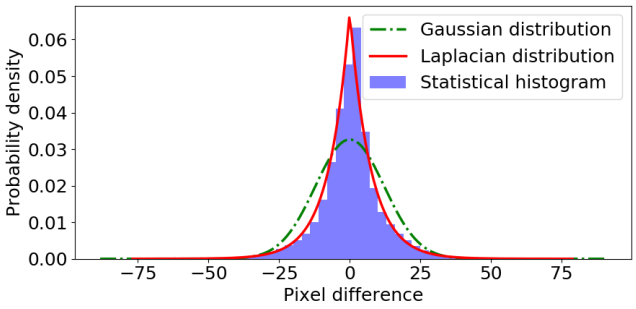

Figure 2: The statistical histogram of difference between real and bicubic LR images and the corresponding fitting distributions.

failed), as shown in Fig. 2. Based on this observation, we propose to learn the degradation uncertainty by minimizing the KL divergence of two Laplacian distributions. By explicitly constraining $L(\boldsymbol{y}_g, \boldsymbol{\theta})$ to be close to a Laplacian distribution $L(\boldsymbol{y}_b, \boldsymbol{I})$, the uncertainty can be learned by minimizing the Kullback-Leibler (KL) divergence of two Laplacian distributions, which can be formulated as

$$\begin{aligned}\mathcal{L}_{kl} &= \mathbb{E}_{\boldsymbol{y}_g}\left\{KL\left[L(\boldsymbol{y}_g, \boldsymbol{\theta})||L(\boldsymbol{y}_b, \boldsymbol{I})\right]\right\} \\ &= \mathbb{E}_{\boldsymbol{y}_g}\left[\boldsymbol{\theta}\exp(-\frac{||\boldsymbol{y}_g - \boldsymbol{y}_b||_1}{\boldsymbol{\theta}}) \right. \\ &\qquad \left. + ||\boldsymbol{y}_g - \boldsymbol{y}_b||_1 - \log\boldsymbol{\theta} - 1\right].\end{aligned} \tag{2}$$

By learning the degradation uncertainty, multiple different but similar degraded versions of a single clean HR image can be sampled according to the learned LR image $\boldsymbol{y}_g$ (mean) and degradation uncertainty $\boldsymbol{\theta}$ (variance). The construction of LR-HR training pairs will be illustrated in Sec. 3.2.

Learning uncertainty in features domain. In the above discussion, we study the uncertainty of degradation in the pixel domain. Still, there is uncertainty in the feature domain which can be explored in perceptual loss. Normally, the perceptual loss can be formulated as

$$\mathcal{L}_{per} = \mathbb{E}_{\boldsymbol{y}_g}||\phi(\boldsymbol{y}_g) - \phi(\boldsymbol{y}_b)||_1, \tag{3}$$

where $\phi(\cdot)$ denotes the feature extractor. In this paper, we propose learning the degradation in both the pixel and feature domains. Similarly to the formulation of content loss, the uncertainty $\boldsymbol{\sigma}$ in the feature domain can be learned by minimizing the KL divergence of feature distributions (e.g., $L(\phi(\boldsymbol{y}_g), \boldsymbol{\sigma})$ and $L(\phi(\boldsymbol{y}_b), \boldsymbol{I})$), which can be expressed by

$$\begin{aligned}\mathcal{L}_{kl\_per} &= \mathbb{E}_{\boldsymbol{y}_g}\left\{KL\left[L(\phi(\boldsymbol{y}_g), \boldsymbol{\sigma})||L(\phi(\boldsymbol{y}_b), \boldsymbol{I})\right]\right\} \\ &= \mathbb{E}_{\boldsymbol{y}_g}\left[\boldsymbol{\sigma}\exp(-\frac{||\phi(\boldsymbol{y}_g) - \phi(\boldsymbol{y}_b)||_1}{\boldsymbol{\sigma}}) \right. \\ &\qquad \left. + ||\phi(\boldsymbol{y}_g) - \phi(\boldsymbol{y}_b)||_1 - \log\boldsymbol{\sigma} - 1\right].\end{aligned} \tag{4}$$

In our implementation, we calculated the perceptual loss on VGG19 features of the $conv5\_4$ convolutional layer. Exploring the uncertainty in the feature domain further facilitates the better generation of realistic LR images $\boldsymbol{y}_g$.

Adversarial loss. For more stable training and visually pleasant results, we adopt the FSSR training strategy [Fritsche *et al.*, 2019], which only calculates the adversarial loss from the high-frequency information that is filtered by the Gaussian blur kernel. The adversarial loss for the training generator can be formulated as follows.

$$\mathcal{L}^G_{adv} = -\mathbb{E}_{\boldsymbol{y}_g}[log(D(F(\boldsymbol{y}_g)))], \tag{5}$$

and the loss for the discriminator can be formulated as

$$\mathcal{L}^D_{adv} = -\mathbb{E}_{\boldsymbol{y}_r}\left[log(D(F(\boldsymbol{y}_r)))\right] - \mathbb{E}_{\boldsymbol{y}_g}\left[log(1 - D(F(\boldsymbol{y}_g)))\right], \tag{6}$$

where $D(\cdot)$ denotes the discriminator and $F(\cdot)$ denotes the Gaussian high-frequency filter sized by 5×5 in our implementation. Calculating adversarial loss from high-frequency information ignores low-frequency information, which is less relevant to the degradation and focuses more on the details of image degradation. Furthermore, such a training strategy reduces the difficulty of GAN training [Fritsche *et al.*, 2019].

Training details. In our implementation, the DSN generator network consists of 8 residual blocks to extract information from the HR image, as shown in Fig. 3. Each residual block contains two convolutional layers and a PReLU activation between them. Then, the convolutional layer with step 2 is employed to reduce the spatial resolution of the features. In the end, three different heads are adopted to transform the features into LR-degraded images, as well as the degradation uncertainty in pixels and feature domains, respectively,

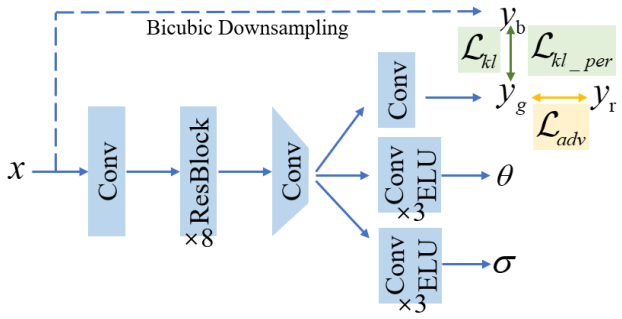

Figure 3: The proposed DSN architecture and KL-based training losses for learning degradation uncertainty.

as shown in Fig. 3. We adopt U-net [Wang *et al.*, 2021b] as a DSN discriminator. The whole DSN is trained by the combination of three losses:

$$\mathcal{L}_{DSN} = \alpha_1 \mathcal{L}_{kl} + \alpha_2 \mathcal{L}_{kl\_per} + \alpha_3 \mathcal{L}_{adv}, \tag{7}$$

where $\alpha_1 = 1, \alpha_2 = 0.01, \alpha_3 = 0.01$. We randomly select 16 RGB HR patches sized by 256×256 as batch input. The initial learning rate is 0.0001 and decreases by half for every 100 epochs. We train the model for 500 epochs.

3.2 Training SRN with Degradation Uncertainty

Constructing LR-HR training pairs. Differently from previous work [Fritsche *et al.*, 2019; Wei *et al.*, 2021] that only learned a single deterministic mapping, we propose to generate multiple realistic LR images from a single HR image by sampling multiple degraded LR images from the learned LR image $\boldsymbol{y}_g$ (mean) and uncertainty of degradation (variance). The sampling process can be expressed as

$$\hat{\boldsymbol{y}}_g = \boldsymbol{y}_g + \epsilon \boldsymbol{\theta}, \tag{8}$$

where $\epsilon = L(0, \boldsymbol{I})$ denotes the standard Laplace distribution. In this way, multiple realistic LR images $\hat{\boldsymbol{y}}_g$ are generated when given each HR image $\boldsymbol{x}$. Thus, during different training batches , different realistic LR images can be generated adaptively from a single HR image with learned degradation uncertainty.

Calculating adversarial loss from gradient information. Inspired by DSN that only calculates adversarial loss from high-frequency information, we propose to calculate adversarial loss from gradient information in the SRN stage for stable training performance and better visual quality. Calculating the adversarial loss from the gradient information enables SRGAN networks to ignore the low-frequency information, which is less relevant to visual effects, and focus more on the image texture and edge details. The adversarial loss for the training generator in SRN can be formulated as

$$\mathcal{L}_{adv}^{SR\_G} = -\mathbb{E}_{\hat{\boldsymbol{x}}}[log(D(G(\hat{\boldsymbol{x}})))], \tag{9}$$

and the loss for discriminator in SRN can be formulated as

$$\mathcal{L}_{adv}^{SR\_D} = -\mathbb{E}_{\boldsymbol{x}}\left[log(D(G(\boldsymbol{x})))\right] - \mathbb{E}_{\hat{\boldsymbol{x}}}\left[log(1 - D(G(\hat{\boldsymbol{x}})))\right], \tag{10}$$

where $D(\cdot)$ denotes the discriminator and $G(\cdot)$ denotes the calculation of the gradient, which can be formulated as

$$G(i,j) = ||I(i+1,j) - I(i,j), I(i,j+1) - I(i,j)||_2, \tag{11}$$

where I denotes the value of pixels and i, j denotes the position of pixels. When comparing the Gaussian high-frequency filter $F(\cdot)$, the gradient calculator $G(\cdot)$ has a stronger high-frequency filtering effect and explores more details of the texture and edges. Therefore, we use the gradient calculator $G(\cdot)$ in SRN for better SR performance and use the Gaussian high-frequency filter $F(\cdot)$ to preserve more degradation information in DSN.

Training details. The whole SRN is trained by the combination of three losses:

$$\mathcal{L}_{SRN} = \beta_1 \mathcal{L}_{con}^{SR} + \beta_2 \mathcal{L}_{per}^{SR} + \beta_3 \mathcal{L}_{adv}^{SR}, \tag{12}$$

where $\beta_1 = 0.01, \beta_2 = 1, \beta_3 = 0.005$. We randomly select 16 RGB LR patches sized by 64×64 as batch input. The initial learning rate is 0.0001. We adopt an exponential moving average (EMA) for more stable training and better performance. We train the model for 1000 epochs.

Note that our approach can be applied to any existing SR or SRGAN network architecture. Following [Fritsche *et al.*, 2019; Wei *et al.*, 2021; Wolf *et al.*, 2021], we adopt the **same** network architecture of ESRGAN [Wang *et al.*, 2018] as SRN in this paper. We adopt the same losses $\mathcal{L}_{con}^{SR}$ and $\mathcal{L}_{per}^{SR}$ as ESRGAN [Wang *et al.*, 2018].

4 Experiments and Results

4.1 Experimental Settings

Datasets. RealSR [Ji *et al.*, 2020] provided LR-HR pairs captured by adjusting the focal length of the camera. Due to cumbersome and labor intensive image collection and challenging post-processing, only a limited number of (about 200 pairs) training pairs have been provided. In our experiments, only 200 degraded LR images collected by the Canon camera are used as real LR images, while the DIV2K dataset provides HR images to train the DSN. Additionally, the RealSR validation set is used for evaluation. NTIRE 2020 RWSR [Lugmayr *et al.*, 2020] challenge offers two tracks for unsupervised SR training. The HR images from DIV2K are used as HR images in both tracks. In Track1, the synthetic degraded Flickr2K dataset is treated as real LR images. Furthermore, Track1 provides a validation dataset for quantitative comparison, which contains 100 images with the same degradation as the training LR images. In Track2, the real degraded LR images are derived from the DPED dataset, which consists of real low-quality images from iphone3. Note that Track2 does not provide reference images for evaluation.

Evaluation metrics. For the dataset that provides HR reference images in the test set, we adopt reference-based evaluation metrics for assessment, such as PSNR, SSIM. Furthermore, the Learning Perceptual Image Patch Similarity (LPIPS) has been used as our perceived quality assessment metric. Since LPIPS correlates well with human perception quality, it is considered the most important metric among the three reference metrics. For the Track2 data set of NTIRE, for which HR reference images are not available, we adopted NIQE as our evaluation metric.

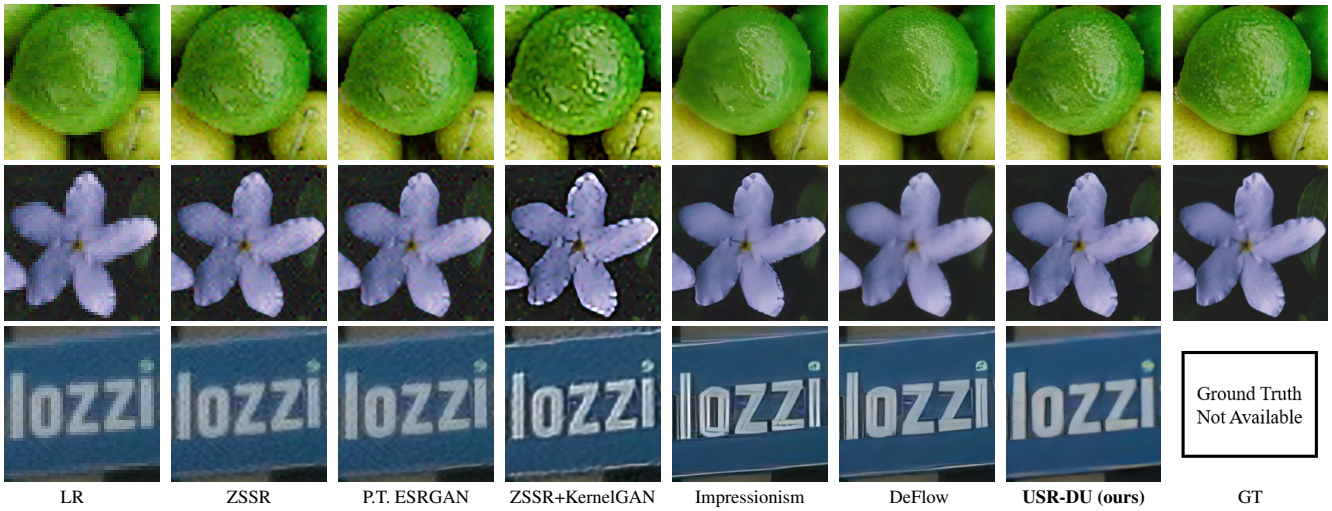

Figure 4: SISR visual quality comparisons of different methods on testing images from NTIRE2020 Track1 and Track2. The first and second rows show the comparisons of Track1, and the last row shows the comparisons of Track2.

	$\boldsymbol{\theta}$	$\boldsymbol{\sigma}$	Sampling	Gradient	PSNR↑	SSIM↑	LPIPS↓
1				✓	25.82	0.7045	0.2305
2†			✓	✓	25.63	0.6913	0.2323
3	✓		✓	✓	25.77	0.7068	0.2101
4	✓	✓	✓		**25.99**	0.7106	0.2097
5	✓	✓	✓	✓	**25.99**	**0.7141**	**0.2033**

Table 1: Ablation study of the proposed method on Track1 of the NTIRE2020 dataset. The best performance is shown in **bold**.

4.2 Ablation Study

To further verify the effectiveness of the proposed approach, we have conducted an ablation study by comparing the final results of the SRN PSNR/SSIM/LPIPS. Our ablation study has been carried out in the NTIRE 2020 Track1 dataset [Lugmayr *et al.*, 2020], and the results are shown in Table 1. The $\boldsymbol{\theta}$ and $\boldsymbol{\sigma}$ denote the uncertainty of learning degradation in pixels and feature domain, respectively, in the DSN stage. The *Sampling* denotes the application of the sampling strategy according to the learned degradation uncertainty $\boldsymbol{\theta}$ and *Gradient* denotes the calculation of the adversarial loss from the gradient information in the SRN stage. It should be noted that 2† denotes the sampling of the LR images by $\hat{\boldsymbol{y}}_g = \boldsymbol{y}_g + \epsilon \boldsymbol{I}$ since no uncertainty has been learned in the DSN stage. The investigation can be divided into two parts as follows:

The effectiveness of learning the degradation uncertainty in pixels and features domains. The corresponding results are shown in Table 1. First, comparing situations 1 and 3, learning degradation uncertainty in the pixels domain and sampling the LR images with learned uncertainty can improve the performance of LPIPS, which is very relevant for the quality of human visual perception. Comparing situations 1, 3, and 5, learning degradation uncertainty in both pixels and features domain (marked as 5) can improve the performance in terms of PSNR, SSIM, and LPIPS, achieving the best performance. Additionally, to investigate the influence of sampling LR images on learned degradation uncertainty, we have conducted the experiment shown in 2† that setting the variance to $\boldsymbol{I}$ and the sampling process can be expressed by $\hat{\boldsymbol{y}}_g = \boldsymbol{y}_g + \epsilon \boldsymbol{I}$. When comparing Situation 1, 2†, it can be observed that adding spatially invariant variance randomness cannot improve the performance and robustness of SR networks, while adopting the learned degradation uncertainty as spatially adaptive variance (marked as 3/5) can improve the performance and robustness of SR networks.

The effectiveness of Calculating adversarial loss from gradient information. We have conducted experiments that calculate adversarial loss from gradient information or pixel values. The corresponding experimental results are shown as situations 4 and 5 in Table 1. Calculating the adversarial loss from the gradient information achieves better performance in terms of both SSIM and LPIPS (the lower, the better), demonstrating the effectiveness of the proposed approach.

4.3 Comparison with SOTA on NTIRE2020

In this section, we have compared our proposed approach with other SR methods on two tracks of the NTIRE2020 challenge. Comparison methods include zero-shot SR (ZSSR) [Shocher *et al.*, 2018] and real unsupervised SR methods such as Impressionism [Cai *et al.*, 2019] (the winner of the NTIRE 2020 RWSR challenge), Deflow [Wolf *et al.*, 2021] and DAP [Wang *et al.*, 2021a]. The pre-trained ESRGAN (denoted as P.T. ESRGAN), which was trained on the bicubic degraded dataset, is also provided as a reference.

The quantitative results of the comparison of Track1 and Track2 are shown in Table 2. In general, our proposed method achieved the best performance in terms of all metrics. Compared to the most competitive DeFlow method [Wolf *et al.*, 2021], our proposed method consistently achieved the best result in the NTIRE2020 Track1 and Track2 datasets. Visual comparisons are shown in Fig. 4. It can be seen that our method has recovered the best results with the most realistic textures and details, such as the lime surface (shown in the first row of Fig. 4) and the edges of the petals (shown in

Figure 5: SISR visual quality comparisons of different methods on testing images from RealSR.

	Track1			Track2
	PSNR↑	SSIM↑	LPIPS↓	NIQE↓
ZSSR	25.01	0.6221	0.6366	5.7652
P.T. ESRGAN	25.40	0.6429	0.5712	5.5913
ZSSR+KernelGAN	21.93	0.5838	0.5970	6.4925
Impressionism	24.82	0.6619	0.2271	4.1326
DeFlow	25.88	0.7002	0.2184	3.4082
DAP	25.40	0.7070	0.2520	-
USR-DU (ours)	**25.99**	**0.7141**	**0.2033**	**3.3426**

Table 2: Comparison with the state-of-the-art on NTIRE2020. The best performance is shown in **bold**.

the second row of Fig. 4). The visual results of Track2 are shown in the third row of Fig. 4. Although no GT image is provided as a reference, it can still be found that our recovered image contains fewer undesirable artifacts. We have also shown that the generated LR images and the learned degradation uncertainty in Fig. 6. The generated realistic LR images have degradation similar to that of real LR images, and the edges and texture areas tend to have a higher degradation uncertainty.

4.4 Comparison with SOTA on RealSR dataset

The RealSR dataset provides real LR-HR pairs for validation. We have compared our approach with several state-of-the-art methods, including ZSSR [Shocher *et al.*, 2018], pre-trained ESRGAN (P.T. ESRGAN) which is trained on synthetic bicubic degradation, ZSSR+KernelGAN [Bell-Kligler *et al.*, 2019] and DASR [Wei *et al.*, 2021]. We also provide the results of the supervised trained ESRGAN (S.T. ESRGAN) for reference, which is trained by the real paired training data of RealSR.

The quantitative comparison is shown in Table 3, from which we can see that our proposed method achieves the best performance compared to other state-of-the-art methods. As shown in Fig. 5, our proposed approach recovers sharp images with pleasing details, clearly outperforming all other competing methods. Compared to DASR [Wei *et al.*, 2021], images are recovered with fewer visible artifacts. When compared with S.T. ESRGAN which is trained in a supervised manner, our method produces sharper and clearer textures with less blurring effect.

	PSNR↑	SSIM↑	LPIPS↓
ZSSR	26.01	0.7485	0.3827
P.T. ESRGAN	26.06	0.7542	0.4370
ZSSR+KernelGAN	24.00	0.7157	0.3069
DASR	26.23	0.7656	0.2507
USR-DU (ours)	**26.56**	**0.7736**	**0.2289**
S.T. ESRGAN	25.68	0.7271	0.2085

Table 3: Comparison with the state-of-the-art on RealSR. The best performance is shown in **bold**.

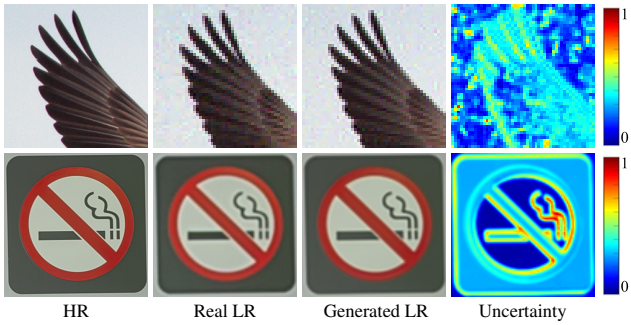

Figure 6: Visualization of generated LR images and uncertainty. The first and second rows show the images from the validation sets of NTIRE Track1 and RealSR, respectively. **Best viewed in color.**

5 Conclusion

In this paper, we propose a novel approach called USR-DU for SR in unsupervised real-world images with learned degradation uncertainty. Given unpaired data, realistic LR images and degradation uncertainty are learned first simultaneously. Then, we sample multiple LR images from the learned LR images (mean) and adaptive degradation uncertainty estimation (variance) to construct LR-HR pairs to train the SR reconstruction networks. Additionally, we propose to calculate the adversarial loss from the gradient information in the SR stage for stable training performance and better visual quality. Experimental results on popular real-world datasets demonstrate the effectiveness of our approach for real-world SR - when compared with other competing methods, ours often achieves sharper SR-resolved images with fewer artifacts.

Acknowledgements

This work was supported in part by the National Key R&D Program of China under Grant 2018AAA0101400 and the China Natural Science Foundation under Grant 61991451, Grant 61632019, Grant 61621005 and Grant 61836008. Xin Li's work is partially supported by the NSF under grants IIS-1951504 and OAC-1940855, the DoJ/NIJ under grant NIJ 2018-75-CX-0032, and the WV Higher Education Policy Commission Grant (HEPC.dsr.18.5).

References

[Bell-Kligler *et al.*, 2019] Sefi Bell-Kligler, Assaf Shocher, and Michal Irani. Blind super-resolution kernel estimation using an internal-gan. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[Cai *et al.*, 2019] 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.

[Chang *et al.*, 2020] Jie Chang, Zhonghao Lan, Changmao Cheng, and Yichen Wei. Data uncertainty learning in face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[Chen *et al.*, 2019] Chang Chen, Zhiwei Xiong, Xinmei Tian, Zheng-Jun Zha, and Feng Wu. Camera lens super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1652–1660, 2019.

[Dong *et al.*, 2014] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *European Conference on Computer Vision*, pages 184–199, 2014.

[Fritsche *et al.*, 2019] Manuel Fritsche, Shuhang Gu, and Radu Timofte. Frequency separation for real-world super-resolution. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, 2019.

[Ignatov *et al.*, 2017] 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*, pages 3277–3285, 2017.

[Ji *et al.*, 2020] 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*, 2020.

[Kendall and Gal, 2017] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[Lim *et al.*, 2017] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1132–1140, 2017.

[Lugmayr *et al.*, 2020] 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*, pages 494–495, 2020.

[Ning *et al.*, 2021a] Qian Ning, Weisheng Dong, Xin Li, Jinjian Wu, and Guangming Shi. Uncertainty-driven loss for single image super-resolution. *Advances in Neural Information Processing Systems*, 34, 2021.

[Ning *et al.*, 2021b] Qian Ning, Weisheng Dong, Guangming Shi, Leida Li, and Xin Li. Accurate and lightweight image super-resolution with model-guided deep unfolding network. *IEEE Journal of Selected Topics in Signal Processing*, 15(2):240–252, 2021.

[Shocher *et al.*, 2018] Assaf Shocher, Nadav Cohen, and Michal Irani. "zero-shot" super-resolution using deep internal learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018.

[Son *et al.*, 2021] Sanghyun Son, Jaeha Kim, Wei-Sheng Lai, Ming-Hsuan Yang, and Kyoung Mu Lee. Toward real-world super-resolution via adaptive downsampling models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[Wang *et al.*, 2018] 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) workshops*, 2018.

[Wang *et al.*, 2021a] Wei Wang, Haochen Zhang, Zehuan Yuan, and Changhu Wang. Unsupervised real-world super-resolution: A domain adaptation perspective. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4318–4327, 2021.

[Wang *et al.*, 2021b] 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.

[Wei *et al.*, 2021] 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.

[Wolf *et al.*, 2021] Valentin Wolf, Andreas Lugmayr, Martin Danelljan, Luc Van Gool, and Radu Timofte. Deflow: Learning complex image degradations from unpaired data with conditional flows. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[Zhang *et al.*, 2018] 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.