PDA-RWSR: Pixel-Wise Degradation Adaptive Real-World Super-Resolution

Andreas Aakerberg[1], Majed El Helou[2], Kamal Nasrollahi[1,3], Thomas Moeslund[1]
[1] Aalborg University, Denmark, [2] ETH Zürich, Switzerland, [3] Milestone Systems, Denmark
anaa,tbm,kn@create.aau.dk, melhelou@ethz.ch

Abstract

While many methods have been proposed to solve the Super-Resolution (SR) problem of Low-Resolution (LR) images with complex unknown degradations, their performance still drops significantly when evaluated on images with challenging real-world degradations. One often overlooked factor contributing to this, is the presence of spatially varying degradations in real LR images. To address this issue, we propose a novel degradation pipeline capable of generating paired LR/High-Resolution (HR) images with spatially varying noise, a key contributor to reduced image quality. Furthermore, to fully leverage such training data, we novelly propose a Pixel-Wise Degradation Adaptive Real-World Super-Resolution (PDA-RWSR) framework. Specifically, we design a new Restormer-based Real-World Super-Resolution (RWSR) model capable of adapting the reconstruction process based on pixel-wise degradation features extracted by a new supervised degradation estimation model. Along with our proposed method, we also introduce a new challenging real-world Spatially Variant Super-Resolution (SVSR) benchmarking dataset, where the images are degraded by complex noise of varying intensity and type, to evaluate the robustness of existing RWSR methods. Comprehensive experiments on synthetic and the proposed challenging real dataset demonstrates the superiority of our method over the current State-of-The-Art (SoTA). The SVSR dataset is available at https://doi.org/10.5281/zenodo.10044260.

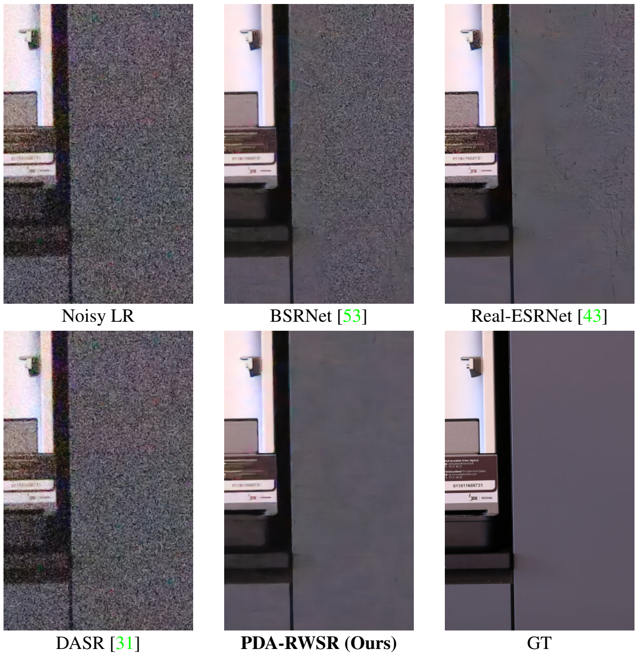


Figure 1. The assumption of uniform degradations in State-of-The-Art (SoTA) Real-World Super-Resolution (RWSR) methods [31, 43, 53] limits the reconstruction performance of a real noisy Low-Resolution (LR) image from our proposed Spatially Variant Super-Resolution (SVSR) dataset. On the contrary, our proposed PDA-RWSR results in a more faithful reconstruction with better artifact suppression.

1. Introduction

Image Super-Resolution (SR) aims to enhance the resolution and details of LR images to generate High-Resolution (HR) images, which have many practical applications. Most recent SR methods accomplish this task by learning a mapping from LR images, generated synthetically by bicubic downsampling, to the corresponding HR images [12, 14, 25, 32, 45, 58]. However, Deep Neural Network (DNN)-based SR methods often suffer from overfitting to the training data distribution, which consequently leads to decreased performance when applied to images with different degradations. Recent attempts to overcome this problem include elaborate degradation models [29, 37, 53] and network conditioning based on degradation estimation [31]. However, these methods assume uniformly distributed degradations and hereby ignore the phenomenon of spatially varying noise present in real images acquired in photon-limited situations. This key factor compromising the image quality is not contingent upon specific image sensors, but a result of the physics involved in the imaging process, such as random photon arrival and non-ideal-sensor characteristics, leading to higher Signal-to-Noise Ratio (SNR) in brighter pixels (low noise) and lower SNR in darker pixels (high noise) [22]. Since the SNR is ultimately controlled by the quantum nature of light, the noise stemming from this phenomenon is an inherent characteristic of any realizable imaging device operating under natural settings [39]. An example case illustrating the limitations of current SoTA methods when applied to images with spatially variant degradations is depicted in Figure 1. A similar challenge has been explored in the context of deblurring images affected by spatially variant blur [51].

Our main motivation lies in the observation that while SoTA RWSR methods can effectively enhance real images without severe degradations, they frequently fail in more challenging and practical scenarios, such as surveillance, where SR is most needed. To this end, we introduce a novel degradation modeling pipeline capable of introducing spatially variant degradations. More precisely, we propose a mask blending technique that synthesizes LR images with varying degrees of noise across the image to resemble the signal-dependent noise in real images. To fully leverage such complex training data, we also propose a Pixel-Wise Degradation Adaptive Real-World Super-Resolution (PDA-RWSR) framework which consists of a DNN that learns to extract pixel-wise degradation features from the LR image in a supervised manner, and a Restormer-based [49] RWSR model that conditions the reconstruction process based on the pixel-wise degradation features. One factor that has been hampering research in practical SR is the lack of sufficiently realistic and challenging real-world SR evaluation datasets. While datasets of real image pairs do exist, they either solely consider the resolution difference [8, 48, 57], or contain noisy/clean pairs without scale differences [2, 40], and hereby excluding more complex scenarios such as LR images corrupted by strong and signal dependant noise. To enable evaluation of RWSR methods in practical scenarios, we propose a new Spatially Variant Super-Resolution (SVSR) dataset, that contains LR images of multiple different scenes captured with varying noise intensity and type, and the corresponding noise-free HR Ground-Truth (GT) images. We summarize our contributions as follows:

- A novel image degradation model that enables degradation at pixel level, as opposed to existing models that mostly operate on image level.
- A new Restormer-based [49] RWSR model capable of adapting the reconstruction process based on pixel-wise degradation features extracted by a new supervised degradation estimation model.
- A novel real-world Spatially Variant Super-Resolution (SVSR) benchmarking dataset that challenges all existing SR methods.
- We highlight the importance of spatially variant degradation modeling and adaptation by demonstrating SoTA performance on the SVSR dataset with our proposed method.

2. Related Work

Single Image Super-resolution: Since the first Convolutional Neural Network (CNN) based SR network [14], a plethora of subsequent work [12, 19, 25, 32, 58, 59] have achieved promising reconstruction performance on LR images created by bicubic downsampling. Furthermore, Generative Adversarial Networks (GANs) have been used to push the SR networks to introduce realistic textures for more visually pleasing results [28, 45, 56]. However, due to the simplistic bicubic downsampling model, the classic SR methods do not generalize well to real-world scenarios [4, 16, 17]. As such, the practical applications of such methods are limited when the LR images contain complex non-uniform degradations, such as noise, blur, and compression artifacts. An overview of classic and deep-learning-based SR methods can be found in [38, 47].

Blind Super-Resolution: Classic blind SR assumes that the blur kernel for the LR image is unavailable [36]. As such, blind SR methods aim to enhance images beyond the bicubic degradation scenario, by including estimated blur kernel information either as a pre-processing step [5, 6, 41, 54], or as part of the SR pipeline [18, 34]. RWSR is a more practical version of blind SR, where the goal is to handle the many complex degradation types, and combinations hereof, present in real-world images. To address this, recent SoTA approaches rely on elaborate degradation models that introduce random combinations of blur and noise types, down-sampling operations, and JPEG compression artifacts [43, 53]. Other works try to estimate the average degradation in the input image and adapt the features in the SR network accordingly [31, 37, 61]. FeMaSR [9] formulates the SR problem as a feature matching problem between LR features and distortion-free HR priors. Other approaches to solving the RWSR problem include [1, 8, 48, 57] that collect paired real LR and HR images for supervised learning. However, these datasets do not contain strong degradations, and while [60] and [1] propose more challenging datasets for joint denoising and SR of microscopy images, and joint low-light enhancement and SR, respectively, there is still a lack of a sufficiently challenging natural image RWSR benchmarking dataset. Partially related to our work are MANet [30] and KOALAnet [26] who perform feature modulation based on spatially varying blur kernel estimations. More recently, the problem of spatially variant noise is investigated in [10], where a method for HDR imaging based on simultaneous denoising and fusion of images with spatially varying SNR ratios is proposed. However, unlike our approach, these methods do not possess the capability to effectively super-resolve real LR images degraded by spatially variant noise.

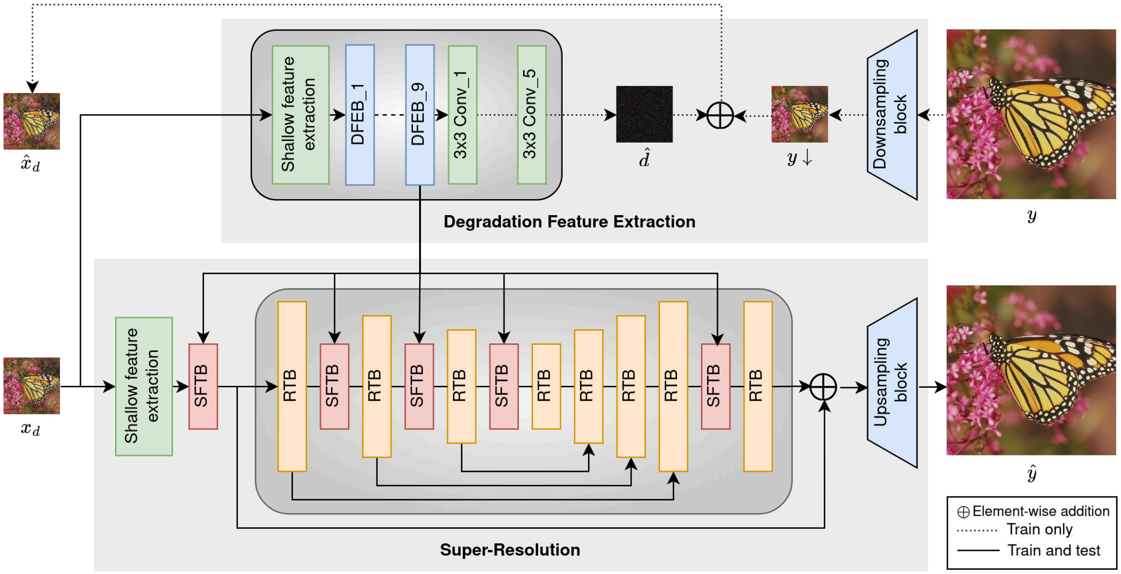


Figure 2. An overview of our proposed Pixel-Wise Degradation Adaptive Real-World Super-Resolution (PDA-RWSR). We design a Transformer-based RWSR model on the basis of Restormer [49], where Restormer Transformer Blocks (RTBs) [49] are organized in a U-Net shaped architecture together with our Spatial Feature Transformation Blocks (SFTBs) to adapt the image reconstruction process based on pixel-wise degradation features. A supervised degradation estimation model with Degradation Feature Extraction Blocks (DFEBs) learns to separate image degradations from content, for the purpose of providing degradation features for conditioning the SR model.

3. Method

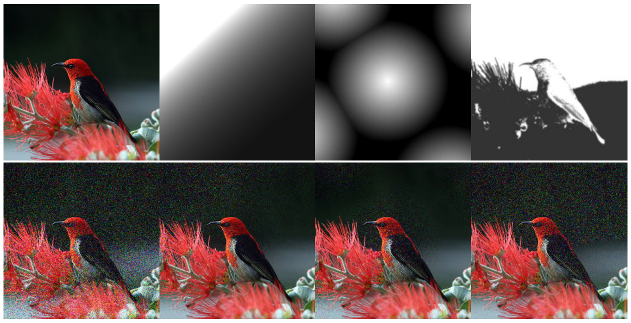
We focus on the challenging task of SR of real-world LR images with complex and non-uniformly distributed degradations, a setting where current SoTA most often fails, as seen in Figure 1. Based on this observation, we design a framework to handle images with spatially variant degradations which include both pixel-wise degradation modeling, estimation, and adaptation. An overview of our proposed method is presented in Figure 2. It consists of a SR network based on the Restormer image restoration Transformer [49], where RTBs [49] are organized in a U-Net shaped architecture together with our Spatial Feature Transformation Blocks (SFTBs), a supervised degradation estimation network with Degradation Feature Extraction Block (DFEB), and a degradation model for synthesizing LR training images with spatially variant degradations. The core novelty of our work is that the SR model is conditioned on pixel-wise degradation features provided by the degradation estimation network for improved refinement of location-specific degradations. In the following, each component in our framework is presented.

3.1. Spatially Variant Degradation Model

A necessity for a DNN based SR model to perform well on test data is prior training on equivalent training data. The classic degradation pipeline for creating realistic LR/HR training image pairs [15] involves convolution with a blur kernel k on the HR image y, followed by downsampling with scale factor s, and lastly, degradation by additive noise to produce the degraded LR image x_d. The pipeline is formally described in Equation 1.

$$x = (y \circledast k) \downarrow_s + n \tag{1}$$

More elaborate and high-order degradation models for synthesis of low-quality LR images have recently been proposed by Zhang *et al.* [53], and Wang *et al.* [43] which introduce diverse combinations of degradations by a random shuffling strategy. However, a fundamental limitation of these models is the use of spatially uniform degradations, which we hypothesize limits the generalization performance to real images. Thus, we propose a novel degradation pipeline where the noise strength varies spatially across the image. This better resembles the distribution of noise in real images, which varies naturally as a result of different SNR levels [20, 21] (See also Figure 6). More specifically, we propose to synthesize LR images with spatially varying noise with the concept of mask blending. First, we generate a mask m of the same spatial size as the LR image x, which contains either a randomly shaped and oriented gradient mask, or a mask based on the image brightness level. Next, we generate a noisy image x_n by adding spatially invariant Gaussian or Poisson noise to x. Then x and x_n are blended according to the varying intensity levels defined in the mask, to form the degraded image x_d with spatially varying noise, formally:



Input images a) Linear gradient b) Radial gradient c) Thresholding

Figure 3. Examples of a LR image from DIV2K [3] degraded by our proposed spatially variant degradation framework. Top left: clean input image. Bottom left: Input image corrupted by uniform noise. a-b: Examples of the different masks used in our framework (top row) and the corresponding output images after blending (bottom).

$$x_d = (1 - m) * x + x_n * m \tag{2}$$

Examples of different masks and the resulting noisy images can be seen in Figure 3. Note that our goal is not to accurately model camera specific noise distributions, but rather to introduce the concept of spatially variant noise into the training of the SR model, and thereby facilitate learning spatially variant noise suppression in real images. More details about the mask generation are given in the supplementary material.

3.2. Pixel-Wise Degradation Estimation

Most existing degradation estimation methods only provide a global average estimate of the degradations in the input image [31, 42]. For more fine-grained control of the reconstruction of local degradations, we propose to estimate the degradation on a pixel level. However, complex combinations of different degradations are difficult to quantify and label for supervised learning, and unsupervised learning requires elaborate frameworks with large batch sizes. As such, we propose to estimate the degradations by learning to extract them directly from a degraded image. More specifically, as shown in Figure 2, the degradation feature extraction network D takes as input an LR image x_d, which is a degraded version of y with spatially variant degradations. In D, shallow features are first extracted by a 7×7 convolutional layer. Next, these features are further processed by 9 DFEBs to extract spatially variant degradation features. Lastly, the deep degradation features are mapped to 3-channels by four 3×3 convolutional layers to form d, which are combined with a bicubicly downsampled version of y by element-wise addition to produce $\hat{x}_d$. The design of the DFEBs, illustrated in Figure 4, combines a gating mechanism and depth-wise convolutions for efficient extraction of local degradation information [49]. In each DFEB, information is first processed by one 3×3 convolutional layer with LeakyReLU followed by two parallel paths through depth-wise convolutional layers, where one is activated with a ReLU non-linearity. Lastly, the two paths are combined by taking the element-wise product followed by a 1×1 convolutional layer. An additive skip connection is used to allow direct information flow from the initial convolutional layer. D is optimized by the loss between $\hat{x}_d$ and x_d. To encourage images with similar structure and frequency distributions we use a combination of SSIM [46] and focal frequency loss [24]. The whole degradation feature extraction model has 4.6M parameters and a moderate receptive field of 51×51. During inference, we extract degradation features from the 9th DFEB for conditioning of the SR network.

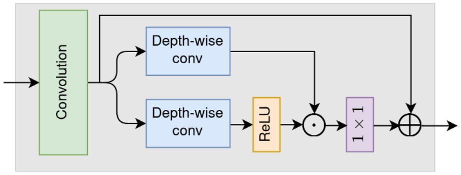


Figure 4. Details of the proposed Degradation Feature Extraction Block (DFEB).

3.3. Pixel-Wise Feature Modulation

To condition the SR model on the pixel-wise degradations estimated by the degradation feature extraction network, we design a feature modulation block that transforms the deep spatial features of the SR network adaptively and individually for each pixel accordingly. As visualized in Figure 5, the Spatial Feature Transformation Block (SFTB) takes a degradation feature map d and an image feature map f of the same spatial dimensions as input. First, channel-wise attention is applied to d, followed by two convolutional layers with LeakyReLU to reduce the channel dimension from 256 to the same dimension as the feature maps in the SR network. As each SFTB shares the same degradation map, the channel attention serves to emphasize the most relevant degradation features for each part of the SR network. Next, feature transformation is performed by two Spatial Feature Transformation (SFT)-layers [44], each followed by convolutional layers, which learn parameters for a spatially affine transformation of each feature map individually. Formally, feature maps f are conditioned on the degradation map d by a scaling and shifting operation:

$$SFT(f, d) = \gamma \odot f + \beta \tag{3}$$

where γ and β are the scaling and shifting parameters and $\odot$ represents the element-wise addition operation. To avoid mixing spatially adjacent degradations, the filter size of all convolutional layers in the feature transformation block is 1×1. Furthermore, multiple separate SFTB are inserted in the SR backbone model, as the deep features propagating through the network have different sensitivity to the degradations for each level in the network. Implementation specific details are given in Section 5.1.

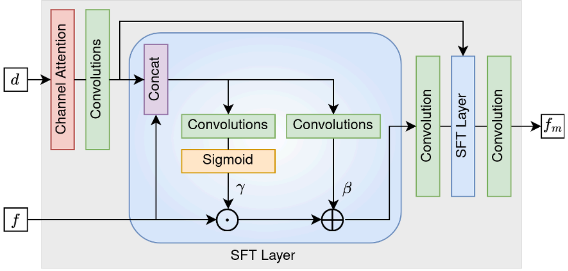


Figure 5. Details of the proposed Spatial Feature Transformation Block (SFTB) for adaptive conditional feature-wise and spatial-wise transformation.

4. SVSR Dataset

In this section, we present the data collection method for our SVSR benchmark dataset along with an analysis of its characteristics. The purpose of this dataset is to advance research in RWSR by enabling evaluation on real LR images with challenging and spatially variant degradations.

4.1. Data Collection

Our goal is to acquire high-quality HR reference images and corresponding low-quality LR images. To this end, we capture static scenes with diverse content, both in- and outdoors, using three different Canon Digital single-lens reflex (DSLR) cameras, two different zoom lenses, and three different aperture values. This ensures diverse degradations, as the noise characteristics and point-spread-function vary between the different cameras, lenses, and aperture settings. The scale difference is obtained by changing the focal length of the zoom lenses, by which we collect image pairs of both ×2 and ×4 scale difference. To obtain varying degrees of noise, we capture multiple images of the same static scene using aperture priority and change the camera's ISO setting. At low ISO settings (low signal gain) the camera will produce the most noise-free images, while at higher ISO settings, and appropriately shorter exposure times, the images will contain more noise due to the lower signal-to-noise ratio. Hence, we capture the clean images at the camera's native ISO setting (ISO100), while the noisy images are captured at incrementally higher ISO levels up to the maximum setting for each camera. We have established ISO1600 as the threshold to distinguish noisy images, as this is the point at which all three cameras introduce visible noise. Consequently, the dataset comprises a total of 978 images, with 141 noise-free images for each scale level, and 555 images as the noisy LR counterparts. For completeness, the released dataset contains the additional ISO levels which we do not consider in this work. A breakdown of the dataset can be seen in Table 1. Note that due to different technologies, images captured at the same ISO setting by different cameras do not necessarily contain similar noise levels and types. Additional details about the cameras, setup, and examples are given in the supplementary material.

Table 1. Overview of the different combinations of camera types and ISO settings and the resulting number of degraded LR images in the SVSR benchmarking dataset.

	ISO						
Camera	1600	3200	6400	12800	25600	51200	65535
Canon 6D	✓	✓	✓	✓	✓	✓	✓
Canon 600D	✓	✓	✓	✗	✗	✗	✗
Canon 1Ds Mark II	✓	✓	✗	✗	✗	✗	✗
Total noisy LR images	141	141	93	45	45	45	45

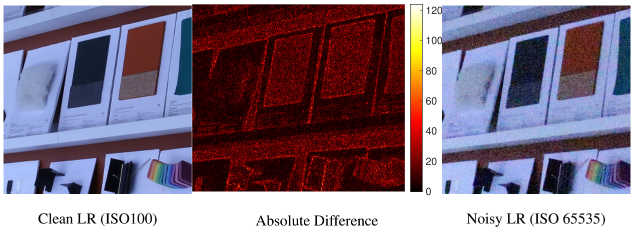


Figure 6. Visualization of the color-channel avg. absolute distance in LR space between a noisy and clean pairs from the SVSR dataset. As seen, noise is more present in the darker regions.

4.2. Data Pre-processing

Even though the image collection is done with the camera mounted on a tripod and using a remote trigger, misalignment between the LR and HR image pairs can still occur, as the different focal lengths distort the image differently. To mitigate this, we design a pre-processing pipeline: First, the lens distortion is removed using Adobe Lightroom [11], followed by center cropping to keep only the sharpest part of the images. Next, we obtain pixel-wise registration of LR and HR images using a luminance-aware iterative algorithm [8], which we empirically found to be more accurate for the highly noisy images, compared to keypoint-based algorithms. To maintain the scale difference between the LR and HR images, we perform the alignment in LR space. Finally, all image pairs are examined, and ones with misalignment, out-of-focus, or other unwanted defects are discarded. The resulting image pairs have a resolution of 640×640, 1280×1280, and 2560×2560px for the ×1, 2, and 4 scale factors, respectively.

4.3. Data Analysis

To demonstrate the spatially variant distribution of noise in the dataset, we visualize the color-channel average absolute distance between LR images of different ISO levels in Figure 6. As seen, a larger degree of noise is present in the darker regions of the image, compared to lighter regions. Figure 7 quantitatively supports this by showing the mean value of the noise in different intensity ranges. Furthermore, to quantify the effect of varying ISO levels on the image quality, we compare clean and noisy images at LR scale for the different ISO values. In Table 2 we present the average standard deviation of the noise, and the resulting change in image quality as the ISO increases. As seen, high ISO settings result in higher noise contributions, which translates to accordingly lower image quality, e.g. the Peak Signal-to-Noise Ratio (PSNR) for the highest ISO setting is 12.53dB lower than for ISO1600. Examples of the different noise levels can be seen in Figure 8, respectively.

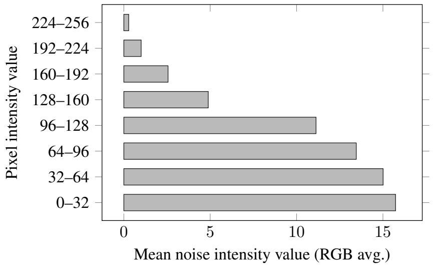


Figure 7. Mean noise intensity values for the highest ISO level of the SVSR dataset, computed at different pixel intensity ranges.

Table 2. Overview of the std. deviation σ of the noise at the different ISO levels in the SVSR benchmarking dataset, and how it affects image quality at LR scale.

ISO	σ	PSNR ↑	SSIM ↑	LPIPS ↓	DISTS ↓
100	0.0	∞	1.0	0.0	0.0
1600	4.93	34.32	0.9041	0.0305	0.0780
3200	6.43	32.06	0.8516	0.0711	0.1203
6400	8.06	30.18	0.8318	0.1061	0.1543
12800	9.38	28.77	0.7813	0.1403	0.1670
25600	12.31	26.37	0.6420	0.2693	0.2232
51200	15.15	24.58	0.5649	0.3364	0.2525
65535	20.87	21.79	0.4239	0.4562	0.3015

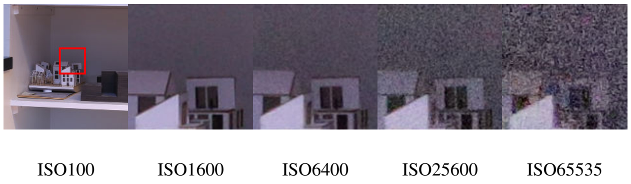


Figure 8. Examples from the SVSR benchmarking dataset illustrating how the noise level changes at different ISO settings.

5. Experiments and Analysis

5.1. Experimental Setup

Table 3. Average PSNR(dB) results of state-of-the-art methods for ×4 SR on synthetic noisy LR images. DN and SR indicates if the method has denoising and/or super-resolution capabilities, respectively. σ indicates the noise level.

			Set14 [50]		BSD100 [35]		Urban100 [23]	
DN	SR	Method	$\sigma = 15$	$\sigma = 25$	$\sigma = 15$	$\sigma = 25$	$\sigma = 15$	$\sigma = 25$
✓	✓	PDM-SR [33]	19.83	17.27	19.36	16.87	18.36	16.69
✗	✓	RRDB [45]	19.84	16.48	19.81	16.42	18.96	15.98
✓	✗	3×3 Median+Bic	20.38	19.89	21.56	21.05	19.13	18.79
✓	✓	DAN [34]	20.98	18.07	20.73	17.95	19.79	17.26
✓	✓	FeMaSR [9]	21.89	21.41	21.80	21.40	20.00	19.67
✗	✗	Bicubic	22.05	19.76	22.08	19.76	20.22	18.56
✓	✓	BSRNet [53]	22.08	19.58	22.14	19.66	20.81	18.98
✓	✓	DASR [31]	23.26	21.73	23.14	21.84	21.26	20.14
✓	✓	MM-RealSRNet [37]	23.41	22.69	23.68	23.04	21.38	20.90
✓	✓	Real-SwinIR-L [29]	23.61	22.12	23.75	22.48	21.96	20.91
✓	✓	Real-ESRNet [43]	23.93	22.74	23.90	22.97	**22.06**	21.24
✓	✓	PDA-RWSR (Ours)	**24.07**	**23.12**	**24.10**	**23.29**	22.04	**21.41**

Datasets: Following recent practice in SR research [37, 43, 45, 53], we use the DIV2K [3] and Flick2K [32] dataset for training. For evaluation on images with synthetic degradations we use Set14 [50], BSD100 [35] and Urban100 [23] which we corrupt by additive Gaussian noise with zero mean and standard deviation $\sigma = 15, 25, 50$, respectively. For evaluation on real-world degraded LR images, we use the SVSR dataset. In both cases, we experiment with ×4 upsampling as commonly used in the SR literature.

Implementation Details: We use our proposed spatially variant noise degradation model together with the degradation pipeline from [53] by replacing the degradation with uniform Gaussian noise with spatially variant Gaussian and Poisson noise. Following [52], we set the noise standard deviation to [1,50] and scale to [2,4] for Gaussian and Poisson noise, respectively. The remaining steps in the degradation pipeline include Gaussian blur, downsampling and JPEG compression noise, with the same hyperparameters as defined in [53] for comparability. As such, any performance improvements related to the degradation modeling are solely due to the introduction of spatially variant noise. We perform experiments on Restormer [49], a Transformer based image reconstruction model, where we add SFTBs for each encoder level, and before the final refinement block. We use average pooling of the degradation maps to match the spatial dimensions of the feature maps at the different encoder levels. ×4 upsampling is done as final step by nearest-neighbour interpolation + convolutional layers, as commonly used in the SR literature [29, 45, 53]. Otherwise, the architecture follows the original implementation. We train our proposed degradation estimation and

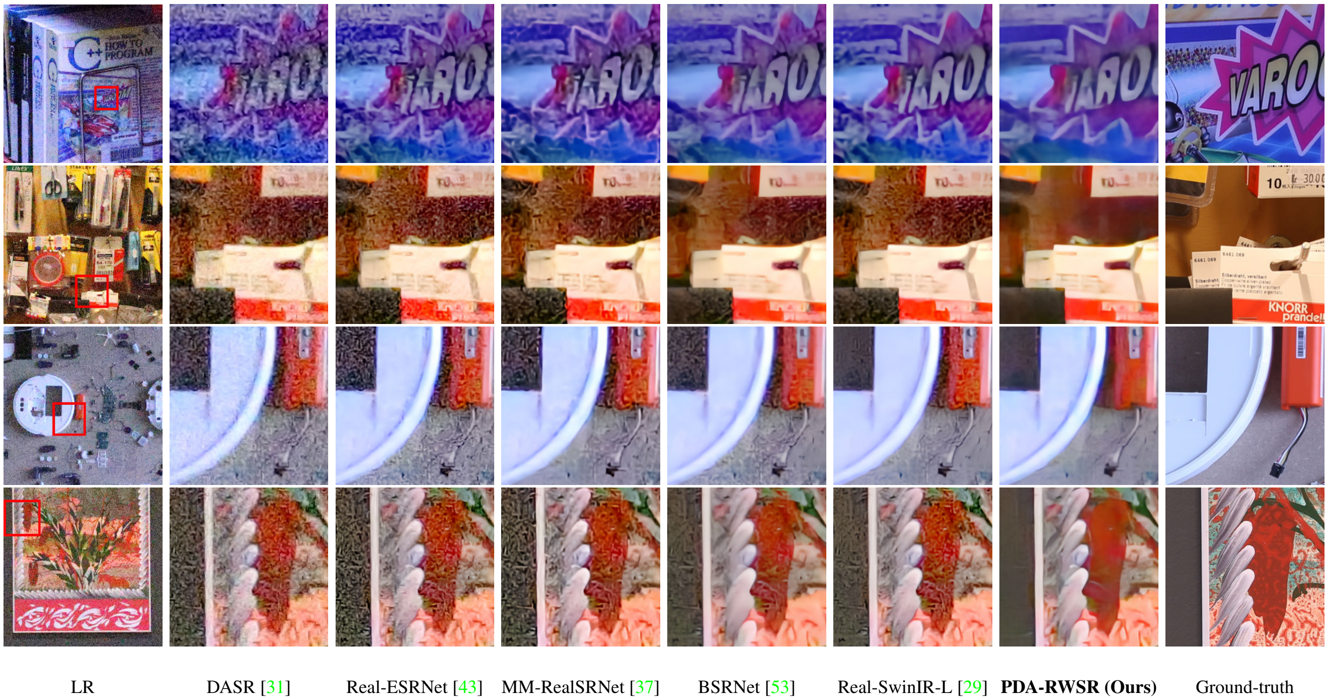


Figure 9. Visual comparison of the reconstruction performance on the SVSR dataset. In comparison to the SoTA approaches, our PDA-RWSR produces more visually faithful results with less artifacts.

Table 4. Quantitative comparison of state-of-the-art methods for ×4 SR on real noisy LR images from the SVSR benchmarking dataset (Full dataset / ISO65535). DN and SR indicate if the method has denoising and/or super-resolution capabilities, respectively.

DN	SR	Method	SSIM ↑	PSNR ↑	LPIPS ↓	DISTS ↓
✓	✓	FeMaSR [9]	0.5914 / 0.3535	22.87 / 18.66	0.2772 / 0.5176	0.1557 / 0.2813
✓	✓	PDM-SR [33]	0.6685 / 0.3891	23.82 / 18.84	0.3047 / 0.5367	0.1805 / 0.2838
✗	✓	RRDB [45]	0.7073 / 0.4380	24.69 / 20.00	0.3011 / 0.5579	0.1745 / 0.2602
✓	✓	DAN [34]	0.7085 / 0.4450	24.69 / 20.03	0.2997 / 0.5527	0.1741 / 0.2895
✓	✓	DASR [31]	0.7092 / 0.3990	24.53 / 19.79	0.2478 / 0.4917	0.1577 / 0.2779
✗	✗	Bicubic	0.7282 / 0.4920	24.84 / 20.41	0.3093 / 0.5439	0.1717 / 0.2828
✓	✓	Real-ESRNet [43]	0.7650 / 0.6043	24.32 / 20.84	0.2063 / 0.4008	0.1407 / 0.2265
✓	✗	3 × 3 Median+Bic.	0.7690 / 0.6412	25.08 / 21.51	0.2953 / 0.4350	0.1687 / 0.2587
✓	✓	MM-RealSRNet [37]	0.7708 / 0.6302	24.26 / 21.17	0.2071 / 0.3836	0.1501 / 0.2280
✓	✓	BSRNet [53]	0.7844 / 0.6707	25.13 / 21.71	0.2067 / 0.3563	0.1401 / 0.2148
✓	✓	Real-SwinIR-L [29]	0.7853 / 0.6818	25.01 / 21.96	0.1956 / 0.3395	0.1442 / 0.2074
✓	✓	PDA-RWSR (Ours)	**0.7943 / 0.7427**	**25.16 / 22.56**	**0.1916 / 0.2985**	**0.1374 / 0.2043**

SR network jointly for 1M iterations with a batch size of 16 using the ADAM [27] optimizer, a learning rate of 2×10^{-4}, LR patch sizes of 64×64, and L1-loss. Note that we do not focus on finding the optimal architecture, or training hyper-parameters, but rather on showing the importance of handling the phenomenon of spatially variant degradations. As such, the performance of our proposed method can likely be further improved.

Evaluation Metrics: We evaluate the reconstruction performance using two hand-crafted (PSNR, SSIM [46]), and two SoTA DNN-based (LPIPS [55], DISTS [13]) Full-Reference Image Quality Assessment (FR-IQA) metrics. PSNR reports the image fidelity as a measure of the peak pixel-wise error between the prediction and target, while SSIM, LPIPS, and DISTS are more focused on the perceived image quality [7].

5.2. Comparison with State-of-The-Art Methods

We compare our method with recent SoTA real-world SR methods. Specifically, we include one codebook-

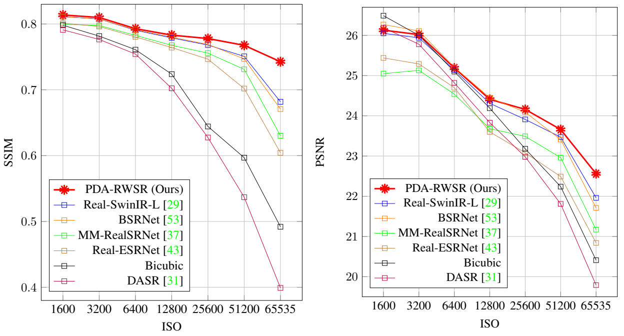


Figure 10. Plot of how the performance (SSIM/PSNR) of SoTA methods decrease as the ISO (noise levels) in the SVSR benchmarking dataset increases. On the contrary, our PDA-RWSR has a more consistent performance across the range.

Table 5. Comparison of different model types. DM, and FM denotes the degradation model, and whether the model uses feature modulation, respectively. Giga Multiply-Accumulates per Second (GMACs) are computed for an input image of 64 × 64 pixels.

Name	DM	FM	Params$\times 10^6$	GMACs	SSIM ↑	PSNR ↑
A	Bicubic	✗	26.2	12.1	0.7058	24.68
B	BSRNet [53]	✗	26.2	12.1	0.7736	25.03
C	Ours	✗	26.2	12.1	0.7906	25.12
D	Ours	✓	28.4	51.4	0.7943	25.16

based method (FeMaSR [9]), two degradation estimation and adaptation-based methods (DASR [31], DAN [34]), five methods relying on elaborate degradation modeling (Real-ESRNet [43], MM-RealSRNet [37], BSRNet [53], PDM-SR [33]) and Transformers (SwinIR [29]), and for completeness, one method trained on bicubicly down-sampled images (RRDBNet [43]). For reference, we also include a filter-based method *i.e.* 3 × 3 Median filter followed by Bicubic upsampling. For all DNN-based methods, we use the pre-trained weights provided by the authors for enhancement of real images and optimized for PSNR, rather than perceptual quality, since our goal is to restore the original image with the highest possible fidelity.

Comparison on Synthetic Data: Table 3 shows the results on synthetically degraded LR images. In this experiment, where the degradations are uniformly distributed, our method outperforms all the competing methods on both noise levels, except for $\sigma 15$ on Urban100 where our method performs comparably with Real-ESRNet [45].

Comparison on Real Data: Table 4 shows the results on real LR images with complex degradations. Contrary to the experiments on synthetic data, the SVSR dataset poses a more challenging reconstruction task, where the assumption of spatially invariant Gaussian noise employed by most of the SoTA methods will not hold. As such, the global degradation estimation-based methods (DASR [31], DAN [34]) cannot handle such real-world scenarios, resulting in low performance based on all Image Quality Assessment (IQA) metrics. Furthermore, while methods based on elaborate degradation models (Real-ESRNet [43], MM-RealSRNet [37], BSRNet [53], PDM-SR [33], SwinIR [29]) are trained on more complex degradations, their reconstruction quality is very inconsistent on images with spatially variant noise from the SVSR dataset. This can be seen visually in Figure 9, and from the plot in Figure 10 where their performance drops sharply as the ISO level increases. On the contrary, our proposed PDA-RWSR performs better and more consistently across the range. This is also reflected in Figure 9, where the reconstructions by our methods are more faithful with fewer artifacts, proving the superiority of PDA-RWSR for dealing with real-world degradations.

5.3. Ablation Studies

In this section, we empirically show the importance of our main technical contributions. As seen in Table 5 using our proposed degradation model with spatially variant noise (C) results in 0.09dB higher PSNR compared to using the degradation model from BSRNet [53] (B). Due to the complementary effect between our spatially variant degradation model and our per-pixel-based degradation feature extraction and adaptation method (D) results in the best performance, although with the cost of additional computations.

6. Conclusion

In this paper, we make significant progress towards SR of real images with complex and spatially varying degradations. Specifically, we propose to adapt the SR reconstruction process on pixel-wise degradations. To achieve this, we introduce a novel pixel-wise degradation feature extraction network that conditions the SR backbone model using pixel-wise modulation blocks. Additionally, we develop a new degradation pipeline capable of introducing spatially variant degradations to the LR training images. We further propose SVSR, a new RWSR benchmarking dataset that challenges all the existing RWSR approaches. Through experiments on synthetic and real LR images, we demonstrate that our proposed PDA-RWSR outperforms current SoTA methods.

Acknowledgments: This research was funded by Milestone Systems A/S, Brøndby Denmark through the Milestone Research Programme at Aalborg University, and the Independent Research Fund Denmark, under grant number 8022-00360B.

References

[1] Andreas Aakerberg, Kamal Nasrollahi, and Thomas B. Moeslund. RELLISUR: A real low-light image super-resolution dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. 2

[2] Abdelrahman Abdelhamed, Stephen Lin, and Michael S. Brown. A high-quality denoising dataset for smartphone cameras. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 1692–1700. Computer Vision Foundation / IEEE Computer Society, 2018. 2

[3] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017. 4, 6

[4] Andreas Lugmayr et al. Ntire 2020 challenge on real-world image super-resolution: Methods and results. *CVPR Workshops*, 2020. 2

[5] Michal Irani Assaf Shocher, Nadav Cohen. "zero-shot" super-resolution using deep internal learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2

[6] Sefi Bell-Kligler, Assaf Shocher, and Michal Irani. Blind super-resolution kernel estimation using an internal-gan. In *Advances in Neural Information Processing Systems 32*, pages 284–293. Curran Associates, Inc., 2019. 2

[7] Yochai Blau, Roey Mechrez, Radu Timofte, Tomer Michaeli, and Lihi Zelnik-Manor. The 2018 pirm challenge on perceptual image super-resolution. In *Computer Vision – ECCV 2018 Workshops*, pages 334–355, Cham, 2019. 7

[8] 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. 2, 5

[9] Chaofeng Chen, Xinyu Shi, Yipeng Qin, Xiaoming Li, Xiaoguang Han, Tao Yang, and Shihui Guo. Real-world blind super-resolution via feature matching with implicit high-resolution priors. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1329–1338, 2022. 2, 6, 7, 8

[10] Yiheng Chi, Xingguang Zhang, and Stanley H Chan. Hdr imaging with spatially varying signal-to-noise ratios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5724–5734, 2023. 2

[11] Adobe Lightroom Classic. *version 10.0)*. Adobe Inc., 2020. 5

[12] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-order attention network for single image super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 11065–11074. Computer Vision Foundation / IEEE, 2019. 1, 2

[13] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P. Simoncelli. Image quality assessment: Unifying structure and texture similarity. *CoRR*, abs/2004.07728, 2020. 7

[14] Chao Dong, C.C. Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 38(2):295–307, Feb 2016. 1, 2

[15] Michael Elad and Arie Feuer. Restoration of a single superresolution image from several blurred, noisy, and undersampled measured images. *IEEE Trans. Image Process.*, 6(12):1646–1658, 1997. 3

[16] Andreas Lugmayr 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, 2019. 2

[17] Pengxu Wei et al. AIM 2020 challenge on real image super-resolution: Methods and results. In Adrien Bartoli and Andrea Fusiello, editors, *Computer Vision - ECCV 2020 Workshops - Glasgow, UK, August 23-28, 2020, Proceedings, Part III*, volume 12537 of *Lecture Notes in Computer Science*, pages 392–422. Springer, 2020. 2

[18] Jinjin Gu, Hannan Lu, Wangmeng Zuo, and Chao Dong. Blind super-resolution with iterative kernel correction. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2

[19] Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. Deep back-projection networks for super-resolution. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 1664–1673. IEEE Computer Society, 2018. 2

[20] Glenn Healey and Raghava Kondepudy. Radiometric CCD camera calibration and noise estimation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 16(3):267–276, 1994. 3

[21] Majed El Helou, Ruofan Zhou, and Sabine Süsstrunk. Stochastic frequency masking to improve super-resolution and denoising networks. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XVI*, volume 12361 of *Lecture Notes in Computer Science*, pages 749–766. Springer, 2020. 3

[22] Bernd Jähne. Evaluation of modern image sensors using the emva 1288 standard. *Imag. Appl. Opt. Opt. Soc. Amer. Tech. Digest*, 2016. 2

[23] Narendra Ahuja Jia-Bin Huang, Abhishek Singh. Single image super-resolution from transformed self-exemplars. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5197–5206, 2015. 6

[24] Liming Jiang, Bo Dai, Wayne Wu, and Chen Change Loy. Focal frequency loss for image reconstruction and synthesis. In *ICCV*, 2021. 4

[25] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR Oral)*, June 2016. 1, 2

[26] Soo Ye Kim, Hyeonjun Sim, and Munchurl Kim. Koalanet: Blind super-resolution using kernel-oriented adaptive local adjustment. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 10611–10620. Computer Vision Foundation / IEEE, 2021. 2

[27] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. 7

[28] 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. 2

[29] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *IEEE International Conference on Computer Vision Workshops*, 2021. 1, 6, 7, 8

[30] Jingyun Liang, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Mutual affine network for spatially variant kernel estimation in blind image super-resolution. In *IEEE International Conference on Computer Vision*, 2021. 2

[31] Jie Liang, Hui Zeng, and Lei Zhang. Efficient and degradation-adaptive network for real-world image super-resolution. In *European Conference on Computer Vision*, 2022. 1, 2, 4, 6, 7, 8

[32] 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, 2, 6

[33] Zhengxiong Luo, Yan Huang, , Shang Li, Liang Wang, and Tieniu Tan. Learning the degradation distribution for blind image super-resolution. In *CVPR*, 2022. 6, 7, 8

[34] Zhengxiong Luo, Yan Huang, Shang Li, Liang Wang, and Tieniu Tan. Unfolding the alternating optimization for blind super resolution. *Advances in Neural Information Processing Systems (NeurIPS)*, 33, 2020. 2, 6, 7, 8

[35] David R. Martin, Charless C. Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, volume 2, pages 416–423 vol.2, 2001. 6

[36] Tomer Michaeli and Michal Irani. Nonparametric blind super-resolution. In *IEEE International Conference on Computer Vision, ICCV 2013, Sydney, Australia, December 1-8, 2013*, pages 945–952. IEEE Computer Society, 2013. 2

[37] Chong Mou, Yanze Wu, Xintao Wang, Chao Dong, Jian Zhang, and Ying Shan. Metric learning based interactive modulation for real-world super-resolution. In *European Conference on Computer Vision*, pages 723–740. Springer, 2022. 1, 2, 6, 7, 8

[38] Kamal Nasrollahi and Thomas B. Moeslund. Super-resolution: A comprehensive survey. *Mach. Vision Appl.*, 25(6):1423–1468, Aug. 2014. 2

[39] Russ Palum. How many photons are there? In *IS and T's PICS Conference*, pages 203–206. Society For Imaging Science & Technology, 2002. 2

[40] Tobias Plötz and Stefan Roth. Benchmarking denoising algorithms with real photographs. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 2750–2759. IEEE Computer Society, 2017. 2

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

[42] Longguang Wang, Yingqian Wang, Xiaoyu Dong, Qingyu Xu, Jungang Yang, Wei An, and Yulan Guo. Unsupervised degradation representation learning for blind super-resolution. In *CVPR*, 2021. 4

[43] 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. 1, 2, 3, 6, 7, 8

[44] Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 606–615. Computer Vision Foundation / IEEE Computer Society, 2018. 4

[45] 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 Laura Leal-Taixé and Stefan Roth, editors, *Computer Vision – ECCV 2018 Workshops*, pages 63–79, Cham, 2019. Springer International Publishing. 1, 2, 6, 7, 8

[46] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, 13(4):600–612, 2004. 4, 7

[47] Zhihao Wang, Jian Chen, and Steven C. H. Hoi. Deep learning for image super-resolution: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2020. 2

[48] Pengxu Wei, Ziwei Xie, Hannan Lu, ZongYuan Zhan, Qixiang Ye, Wangmeng Zuo, and Liang Lin. Component divide-and-conquer for real-world image super-resolution. In *Proceedings of the European Conference on Computer Vision*, 2020. 2

[49] Syed Waqas Zamir, Aditya Arora, Salman H. Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. *CoRR*, abs/2111.09881, 2021. 2, 3, 4, 6

[50] Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In *Curves and Surfaces: 7th International Conference, Avignon, France, June 24-30, 2010, Revised Selected Papers 7*, pages 711–730. Springer, 2012. 6

[51] Jiawei Zhang, Jinshan Pan, Jimmy S. J. Ren, Yibing Song, Linchao Bao, Rynson W. H. Lau, and Ming-Hsuan Yang. Dynamic scene deblurring using spatially variant recurrent neural networks. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 2521–2529. Computer Vision Foundation / IEEE Computer Society, 2018. 2

[52] Kai Zhang, Yawei Li, Jingyun Liang, Jiezhang Cao, Yulun Zhang, Hao Tang, Radu Timofte, and Luc Van Gool. Prac-

tical blind denoising via swin-conv-unet and data synthesis. *arXiv preprint*, 2022. 6

[53] Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *IEEE International Conference on Computer Vision*, pages 4791–4800, 2021. 1, 2, 3, 6, 7, 8

[54] Kai Zhang, Wangmeng Zuo, and Lei Zhang. Deep plug-and-play super-resolution for arbitrary blur kernels. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 1671–1681. Computer Vision Foundation / IEEE, 2019. 2

[55] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 586–595. IEEE Computer Society, 2018. 7

[56] Wenlong Zhang, Yihao Liu, Chao Dong, and Yu Qiao. Ranksrgan: Generative adversarial networks with ranker for image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3096–3105, 2019. 2

[57] Xuaner Zhang, Qifeng Chen, Ren Ng, and Vladlen Koltun. Zoom to learn, learn to zoom. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 3762–3770. Computer Vision Foundation / IEEE, 2019. 2

[58] 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, 2

[59] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 2472–2481. IEEE Computer Society, 2018. 2

[60] Ruofan Zhou, Majed El Helou, Daniel Sage, Thierry Laroche, Arne Seitz, and Sabine Süsstrunk. W2S: microscopy data with joint denoising and super-resolution for widefield to SIM mapping. In *Computer Vision - ECCV 2020 Workshops - Glasgow, UK, August 23-28, 2020, Proceedings, Part I*, volume 12535 of *Lecture Notes in Computer Science*, pages 474–491. Springer, 2020. 2

[61] Yifeng Zhou, Chuming Lin, Donghao Luo, Yong Liu, Ying Tai, Chengjie Wang, and Mingang Chen. Joint learning content and degradation aware feature for blind super-resolution. In João Magalhães, Alberto Del Bimbo, Shin'ichi Satoh, Nicu Sebe, Xavier Alameda-Pineda, Qin Jin, Vincent Oria, and Laura Toni, editors, *MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022*, pages 2606–2616. ACM, 2022. 2