Benchmarking Ultra-High-Definition Image Super-resolution

Kaihao Zhang[1] Dongxu Li[1] Wenhan Luo[2]
Wenqi Ren[3] Björn Stenger[4] Wei Liu[2] Hongdong Li[1] Ming-Hsuan Yang[5,6,7]
[1] Australian National University [2] Tencent [3] IIE, CAS [4] Rakuten Institute of Technology
[5] Google Research [6] University of California, Merced [7] Yonsei University

Abstract

Increasingly, modern mobile devices allow capturing images at Ultra-High-Definition (UHD) resolution, which includes 4K and 8K images. However, current single image super-resolution (SISR) methods focus on super-resolving images to ones with resolution up to high definition (HD) and ignore higher-resolution UHD images. To explore their performance on UHD images, in this paper, we first introduce two large-scale image datasets, UHDSR4K and UHDSR8K, to benchmark existing SISR methods. With 70,000 V100 GPU hours of training, we benchmark these methods on 4K and 8K resolution images under seven different settings to provide a set of baseline models. Moreover, we propose a baseline model, called Mesh Attention Network (MANet) for SISR. The MANet applies the attention mechanism in both different depths (horizontal) and different levels of receptive field (vertical). In this way, correlations among feature maps are learned, enabling the network to focus on more important features.

1. Introduction

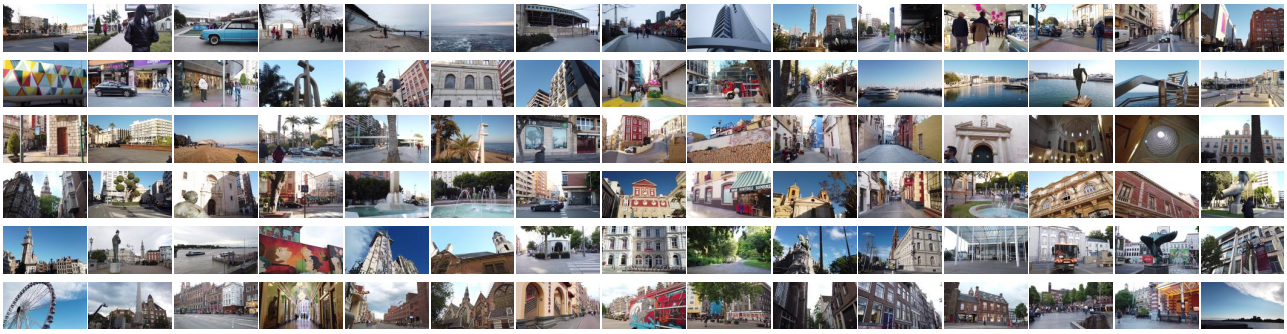
The task of single image super-resolution (SISR) is to produce an image of high resolution (HR) given a low resolution (LR) input. In practice, image super-resolution has a wide range of applications, such as medical image analysis [33], image generation [19], and face recognition at large distances [53]. Super-resolving images is inherently ill-posed, *i.e.*, one LR image can correspond to multiple HR images. To tackle this problem, traditional methods use prior cues from HR images or LR exemplar images [14, 12, 46, 13, 6, 22, 47, 11, 18, 37, 31]. Recent deep learning methods remove the need to explicitly design different types of priors. Networks are trained with pairs of corresponding HR and LR images in an end-to-end manner. With sufficient training data, deep learning models have achieved impressive results [8, 44, 32, 20, 35, 26, 51, 52, 29, 43].

Most of them are trained based on HD images of up to 2K resolution, with the DIV8K [15] dataset being an exception. Thus, it is not clear how they perform in the case of ultra-high definition (UHD) images, including 4K and 8K resolution images. Currently, an increasing number of mobile devices supports capturing images at these resolutions. UHD images provide better visual pleasing effects and they are also better to train SISR approaches, applicable to large upscaling factors like $8\times$ or $16\times$. In this paper, we explore the SR performance of current SISR methods on such UHD images. We collect two large-scale datasets of images with resolutions of 4K and 8K, respectively, from the Internet. The 4K dataset, *UHDSR4K*, includes $5,999$ and $2,100$ images for training and testing, respectively. The 8K dataset, *UHDSR8K*, contains $2,029$ training and 937 test images, respectively. As far as we know, UHDSR4K and UHDSR8K are the largest UHD image datasets for 4K and 8K image super-resolution, respectively. Sample images are shown in Fig. 1.

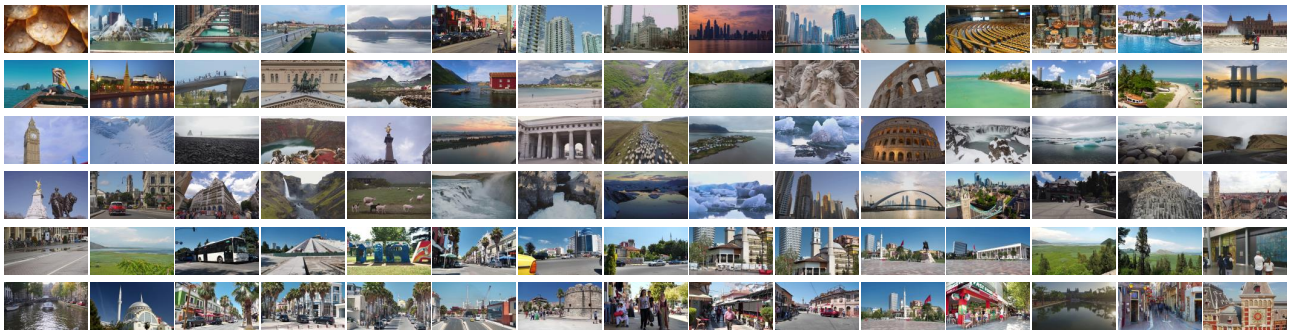
We propose seven settings to assess the performance of existing methods. These include different upsampling factors (from $2\times$ to $16\times$), and two additional settings to evaluate common image degradations, blur plus downsampling and downsampling plus noise. We evaluate ten recent SISR methods on these datasets, and train the respective models on the new datasets. Training one model on a single dataset takes approximately three weeks, and the total training time for all models was over $70,000$ V100 GPU hours.

By conducting this benchmarking study, we thus obtain comprehensive understanding of how the current SISR models work in the specific 4K and 8K settings, both in terms of standard metrics, such as PSNR and SSIM, and perceptual quality.

Further, we propose a Mesh Attention Network (*MANet*) to improve the feature representation ability via learning the inter-dependencies between different feature maps. Specifically, *MANet* is a mesh architecture, whose horizontal and vertical layers represent the feature maps from different depths and different receptive fields, respectively. Within the *MANet*, a novel mesh attention module is introduced to simultaneously learn the relationship between features from different depths and different levels of receptive fields. Fi-



(a) Sample images from the UHDSR4K dataset.



(b) Sample images from the UHDSR8K dataset.

Figure 1. **Sample images from the UHDSR4K and UHDSR8K datasets**. These two datasets consist of a large number of 4K and 8K UHD images, respectively.

nally, the weighted sum of feature maps from horizontal and vertical layers allows the *MANet* to focus on informative depths and receptive fields from input LR features to reconstruct SR images.

In summary, the contributions of this paper are threefold:

- First, we introduce two large-scale UHD image datasets for super resolving. To our knowledge, they are the largest-scale UHD datasets in the field of 4K and 8K image super-resolution. In addition, both datasets provide seven degradation settings to conveniently evaluate SISR methods.
- Second, we extensively evaluate the state-of-the-art SISR methods on the two datasets. By doing so, we are able to understand the potential and limitations of these methods.
- Third, we propose a baseline model, called *MANet* for SISR with a novel mesh attention module. Experiments verify its effectiveness on the UHD SISR task.

2. Related Work

2.1. SISR Datasets

Several datasets for SISR training and evaluation have been introduced in the literature, including T91 [47], Set5 [3], BSDS300 [27], BSDS500 [2], General-100 [10], OutdoorScene [40], PIRM [4], Manga109 [28], Urban100 [17], DIV2K [36], RealSR [5], L20 [38], DIV8K [15], Set14 [48], and Sun-Hays 80 [34]. Among these datasets, the sizes of T91 [47], Set5 [3], BSDS300 [27], BSDS500 [2], General-100 [10], PIRM [4], Manga109 [28], RealSR [5], and Urban100 [17] are relatively small, containing 5-595 images each for training and testing, respectively. Image resolutions range from 264×204 to 826×1169. Wang *et al.* provided a large-scale OutdoorScene dataset, which includes $10,624$ images, but at a mean image resolution of only 553×440. The DIV2K dataset is the current standard dataset for training and testing methods for 2K image super-resolution. It contains 800 and 200 images for training and testing, respectively. Yang *et al.* [45] published an earlier SISR benchmark dataset, evaluating SISR methods on 229 images with resolution lower than 2K.

In order to evaluate the performance on even higher resolution images, Timofte *et al.* [38] introduced the L20 dataset, containing images of 3843×2870 resolution. Although this is within the UHD range, the number of images is too small to train state-of-the-art deep SISR methods. More recently, Gu *et al.* [15] created the DIV8K dataset, which contains $1,504$ images with 8K resolution only. In this paper we focus on benchmarking state-of-the-art deep

learning methods on 4K and 8K resolution images and introduce two new large-scale datasets for this task. See Table 1 for an overview of popular benchmark datasets.

2.2. Deep Learning based SISR Methods

Most state-of-the-art SISR methods are based on deep learning [43]. For classical solutions to SISR, readers can refer to other works [39]. The work by Dong et al. [8, 9] first adopted deep learning for image super-resolution, and many improvements have been proposed since. For example, Kim et al. [21] proposed a deeply-recursive convolutional network (DRCN). Skip connections are introduced to train this network. EDSR [26] is a deep residual network without redundant modules and is combined with multi-scale processing. Efficiently super-resolving images has also attracted attention in recent years [23, 20, 10]. GANs were introduced in [24] to enhance the perceptual quality of the produced HR images. Similarly, GANs are used in [41] to enhance the visual quality using adversarial and perceptual loss functions. Rather than focusing on pixel-wise reconstruction, in [30], Sajjadi et al. proposed a novel network focusing on automated texture synthesis to enhance details. In [16], a Deep Back-Projection Network (DBPN) is developed to study the mutual dependencies between HR and LR images, with a mechanism of error feedback. Hierarchical features are learned in [52] to make full use of cues from various scales. Dense connections are also introduced in this paper to improve the feature representation. Residual channel attention networks (RCAN) were introduced in [51], where a residual-in-residual (RIR) structure and a channel attention module were proposed. To overcome the shortage of channel attention, *i.e.,* ignoring the correlation among different layers, a new holistic attention network (HAN) is proposed in [29], which is composed of a layer attention module (LAM) and a channel-spatial attention module (CSAM). Dai et al. also employed the attention mechanism for the SISR task in [7]. Specifically, they proposed a second-order attention network (SAN) to exploit the correlation of features from the intermediate layers. The feedback mechanism is also employed in [25]. An image super-resolution feedback network (SRFBN) is constructed with RNN structure to refine feature representations with information in difference scales.

3. Benchmark Datasets

We present a benchmark study by evaluating recent state-of-the-art algorithms on UHD image super-resolution. To this end, we first build appropriate datasets. In the following, we introduce the collection process of the UHDSR4K and UHDSR8K datasets, and the settings associated with the two datasets for evaluating the selected methods are represented.

Table 1. **Single image super-resolution datasets**. We introduce two new large-scale UHD (4K and 8K) SR benchmark datasets.

Dataset	Size	Avg. Resolution	Format
T91	91	264×204	PNG
Set5	5	313×336	PNG
BSDS500	500	432×370	JPG
BSDS300	300	435×367	JPG
General-100	100	435×381	BMP
OutdoorScene	10,624	553×440	PNG
PIRM	200	617×482	PNG
Manga109	109	$826 \times 1,169$	PNG
Urban100	100	984×797	PNG
RealSR	595	$1,541 \times 1,302$	PNG
DIV2K	1,000	$1,972 \times 1,437$	PNG
L20	20	$3,843 \times 2,870$	PNG
DIV8K	1,504	$5,557 \times 3,935$	PNG
UHDSR4K	**8,099**	**3,840 × 2,160**	**PNG**
UHDSR8K	**2,966**	**7,680 × 4,320**	**PNG**

3.1. The UHDSR Datasets

We collect UHD images of 4K and 8K from the Internet (Google, Youtube, and Instagram), containing diverse scenes such as city scenes, people, animals, buildings, cars, natural landscapes, and sculptures. These images were captured using various cameras in outdoor and indoor scenes, which are shown in Fig. 1.

The first dataset, UHDSR4K, includes images of $3,840 \times 2,160$ resolution. Its training set contains $5,999$ HR images and the test set $2,100$ HR images, respectively. The city scenarios of training and testing sets are different. These two sets also contain the same number of LR images in each degradation setting, as shown in the next section. The second dataset, UHDSR8K, is composed of $2,029$ images for training and 937 images for testing, with different street scenarios. The image resolution is $7,680 \times 4,320$.

We apply seven different degradation settings to each of these two datasets, obtaining over $77,000$ pairs of HR and LR images in total.

3.2. Image Degradation Settings

Real-world image degradation processes are complex and challenging to capture accurately. The strategy employed in most existing datasets is to simulate the degradation process by specific operations such as downsampling. Some datasets contain HR and LR image pairs captured of the same scene. Other methods use pixel-wise registration to adjust image pairs. However, as we have only the UHD images at their original resolution, we follow the strategy of simulating the degradation [36, 15]. We use seven different degradation settings, named $2\times$, $3\times$, $3\times$*_BD*, $3\times$*_DN*, $4\times$, $8\times$, $16\times$. The numbers indicate the downsampling factor, "D" stands for downsampling, "B" indicates a blur opera-

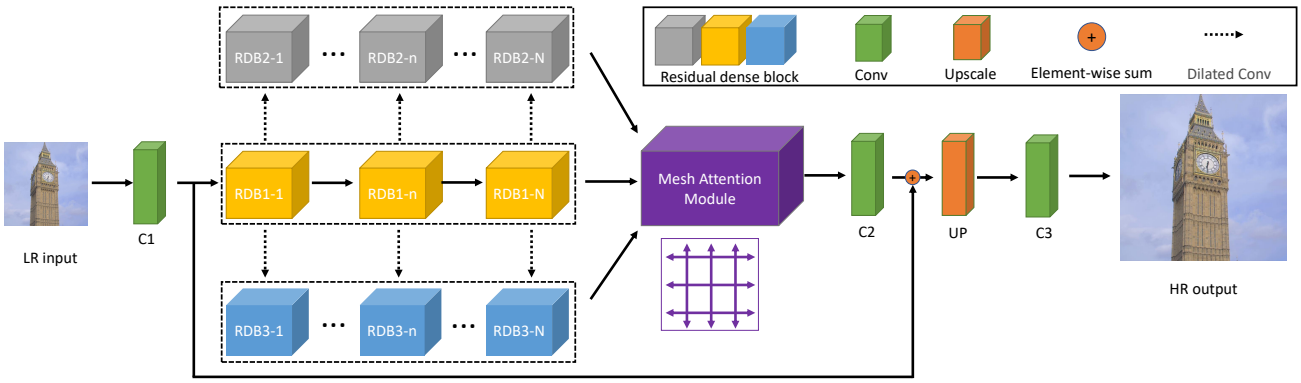


Figure 2. **The architecture of the proposed mesh attention network**. MAN takes a low-resolution image as input, and uses residual dense blocks (RDB) and dilated convolutions to extract feature maps at different levels and from different receptive fields within the same level, respectively. Both are fed into a mesh attention module to learn correlations between different levels and different receptive fields. Finally, a set of layers including upsampling and convolution are used to generate high-resolution images.

tion, and "N " stands for Gaussian noise that is added to the LR images. The order of letters indicates the order of operations, for example, "**BD**" means that the blur artifact is applied prior to the downsampling operation. Similar to [52] and [49], for downsampling, we use bicubic interpolation (**BI**). When blurring images, Gaussian blur is employed with a kernel of size 7×7 and a standard deviation of 1.6. Gaussian noise is added to images to simulate the noise effect. Specifically, the noise level (σ in the Gaussian noise model) is set as 30.

4. Mesh Attention Network for SR

In this section, we introduce the network architecture of the proposed Mesh Attention Network (MAN).

4.1. Network Architecture

As shown in Fig. 2, the proposed MAN is composed of four parts: prepossessing module, dilated convolution module, mesh attention module, and up-sampling module.

Prepossessing module. Given a low-resolution image, the network first extracts features via a convolutional layer.

$$F_{C1} = H_{C1}(I_{LR}), \tag{1}$$

where I_{LR}, H_{C1}, and F_{C1} are the input low-resolution image, the function indicating the first convolutional layer, and features extracted via the first layer, respectively.

Dilated convolution module. F_{C1} is passed to a dilated convolution module to further extract features. The dilated convolutional module consists of several Residual Dense Blocks (RDB) and a dilated convolutional layer. Specially, one RDB first takes the F_{C1} as input to extract features

$$F_{RDB1-1} = H_{RDB1-1}(F_{C1}), \tag{2}$$

where H_{RDB-1} and F_{RDB-1} denote the function representing the RDB and its extracted features, respectively. Then dilated convolution is applied to extract two more features as,

$$F_{RDB2-1} = H_{RDB2-1}(F_{RDB1-1}), \tag{3}$$

$$F_{RDB3-1} = H_{RDB3-1}(F_{RDB1-1}), \tag{4}$$

where H_{RDB2-1} and H_{RDB3-1} are functions of the dilated convolutional layers with dilation parameters set to 2 and 4, respectively, to obtain different levels of the receptive field. F_{RDB2-1} and F_{RDB3-1} are their corresponding features. The proposed dilated convolutional module has N number of RDBs, and the output of the n-th RDB and dilation convolutional layers is denoted as:

$$F_{RDB1-d} = H_{RDB1-n}(F_{RDB1-(n-1)}), \tag{5}$$

where H_{RDB1-n} denotes the n-th RDB operation. $F_{RDB1-(n-1)}$ and F_{RDB1-n} are its input and output.

The operations in the two streams corresponding to the n-th RDB are denoted as:

$$F_{RDB2-n} = H_{RDB2-n}(F_{RDB1-n}), \tag{6}$$

$$F_{RDB3-n} = H_{RDB3-n}(F_{RDB1-n}), \tag{7}$$

where H_{RDB2-n} and H_{RDB3-n} are the dilated convolutional layers. Their input, F_{RDB1-n}, is obtained from the output of $RDB2-n$, and their outputs are F_{RDB2-n} and F_{RDB3-n}. All the F_{RDB1-n}, F_{RDB2-n}, F_{RDB3-n} are of the same size.

Mesh attention module. After obtaining the three hierarchical features by the sets of RDBs and dilation convolutional layers, we introduce a mesh attention module to make full use of the features from all preceding layers, allowing to make use of features from both horizontal and vertical directions. The attention module in the horizontal direction allows the proposed model to address features from different levels, and the attention module in the vertical direction

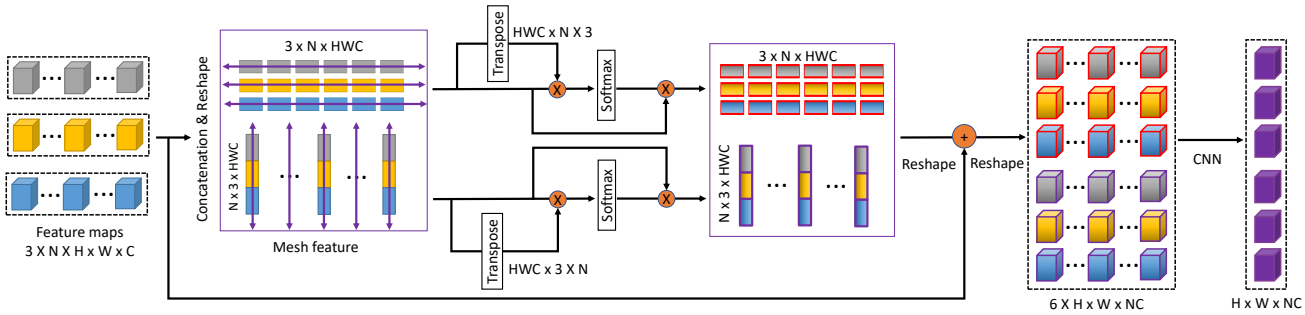


Figure 3. **The architecture of the proposed mesh attention module.** Attention mechanism is applied in both the horizontal and vertical directions to learn the dependency among feature maps from different levels and different receptive fields.

allows the proposed model to process features from different levels of the receptive field.

The above three feature groups are fed into the proposed mesh attention module, which is denoted as:

$$F_{MAM} = H_{MAM}(concat(F_{RDB1-1}, ..., F_{RDB3-N}))\,, \tag{8}$$

where H_{MAM} and F_{MAM} are the functions of the mesh attention module and its output, respectively. We will discuss the mesh attention module in details in Sec. 4.2.

Up-sampling module. After obtaining the mesh attentive features in the LR space, we use an up-sampling module, including a convolutional layer ($C2$), an up-convolutional layer (UP), and another convolutional layer ($C3$) to reconstruct high-resolution images. The process can be described as:

$$I_{SR} = H_{USM}(F_{MAM})\,, \tag{9}$$

where F_{MAM} is the output of the mesh attention module. H_{USM} denotes the operations in the up-sampling module. Its output is the high-resolution image I_{SR}.

4.2. Mesh Attention Module

In order to model inter-dependencies among features at different depths within the network, we propose a mesh attention module to treat the feature maps from each layer differently and learn the relation among them. In the horizontal direction, it learns three groups of dependencies among features of different depths. Similarly, it learns D groups of dependencies among features of different levels of the receptive field in the vertical direction. In this way, the proposed network is capable of learning different attention weights corresponding to features of different depths and levels of the respective field, and thus achieves a better feature representation ability. As shown in Fig. 3, when the feature maps are fed into the MAM, they are reshaped and recombined into two groups. The first group is composed of three matrices, each of shape $N \times HWC$, corresponding to one stream in Fig. 2. This matrix is multiplied with its transpose to derive an N by N correlation matrix, with each element being,

$$w_{i,j} = \phi(\varphi(F_{RDB})_i \cdot \varphi(F_{RDB})_j^\top), i, j = 1, 2, 3, \ldots, N, \tag{10}$$

where ϕ and φ denote the softmax and the reshape operation, respectively. F_{RDB} is the output of the dilated convolution module, and i and j are feature indexes to compute correlations.

Similarly, the second group of features is composed of N matrices, where each is of $3 \times HWC$, corresponding to the depth in Fig. 2. This matrix is multiplied with its transpose to derive a 3 by 3 correlation matrix, with elements being

$$w_{i,j} = \phi(\varphi(F_{RDB})_i \cdot \varphi(F_{RDB})_j^\top), i, j = 1, 2, 3. \tag{11}$$

With this formulation, we obtain $N + 3$ correlation matrices in total. These two groups of feature maps are multiplied with these $N + 3$ correlation matrices to derive two groups of feature maps (the same as the mesh features in terms of size). These two groups of features are reshaped and respectively added with the original feature maps, to derive two sets of feature maps of size $3 \times N \times H \times W \times C$. They are concatenated along the first axis and reshaped to a tensor of size $6 \times H \times W \times NC$, termed as F_{matrix}. The new feature maps F_{matrix} help the proposed MAN focus on different depths and different levels of the respective field. It is further fed into a convolutional layer to create new feature maps of size $H \times W \times NC$ for post-processing. The output of MAM can be represented as:

$$F_{MAM} = H_{one}(F_{matrix})\,, \tag{12}$$

where H_{one} means convolution.

5. Experiments and Analysis

In this section, we benchmark existing SISR methods and our proposed MANet on the proposed UHDSR4K and UHDSR8K datasets.

5.1. Evaluated SISR Methods

We compare ten state-of-the-art SISR methods in a benchmark study, DRLN [1], HAN [29], RDN [52], RCAN

Table 2. **Performance comparison of representative methods for SISR on the UHDSR4K dataset**. Both PSNR and SSIM values are reported.

Scale	Metrics	SRCNN	FSRCNN	VDSR	LapSRN	EDSR	DBPN	RCAN	RDN	HAN	DRLN	MANet
$2\times$	PSNR	42.119	41.535	43.315	43.153	43.614	43.330	43.593	43.642	43.641	43.560	43.742
	SSIM	0.9838	0.9828	0.986	0.9856	0.9863	0.9859	0.9862	0.9862	0.9864	0.9862	0.9865
$3\times$	PSNR	34.082	33.614	35.115	-	35.674	-	35.576	35.769	35.547	35.808	35.842
	SSIM	0.9503	0.9462	0.9575	-	0.9608	-	0.9608	0.9614	0.9601	0.9617	0.9618
$3 \times \_BD$	PSNR	29.681	30.587	32.729	-	35.046	-	35.136	35.199	35.138	34.107	35.240
	SSIM	0.8672	0.8824	0.9187	-	0.9438	-	0.9448	0.9455	0.9449	0.9367	0.9457
$3 \times \_DN$	PSNR	30.026	30.12	30.916	-	31.557	-	31.619	31.703	31.563	31.725	31.589
	SSIM	0.8756	0.8818	0.8959	-	0.9091	-	0.9090	0.9112	0.9085	0.9110	0.9088
$4\times$	PSNR	30.586	30.162	31.540	31.823	32.073	32.157	32.164	32.532	32.177	32.372	32.218
	SSIM	0.9131	0.9058	0.9249	0.9281	0.9310	0.9318	0.9311	0.9353	0.9314	0.9338	0.9315
$8\times$	PSNR	25.421	25.401	25.924	26.563	26.816	26.772	26.816	27.116	26.856	27.009	26.877
	SSIM	0.8126	0.8109	0.8262	0.8411	0.8469	0.8466	0.8483	0.8548	0.8489	0.8520	0.8493
$16\times$	PSNR	22.515	22.464	22.733	23.285	23.479	23.434	23.626	23.639	23.656	23.536	23.523
	SSIM	0.7498	0.7367	0.7569	0.7714	0.7750	0.7762	0.7812	0.7820	0.7821	0.7813	0.7805

[51], DBPN [16], EDSR [26], LapSRN [23], VDSR [20], FSRCNN [10], and SRCNN [8]. All methods are based on deep learning.

5.2. Implementation

Both of the UHDSR4K and UHDSR8K datasets have seven different degradation settings. Each setting corresponds to pairs of LR and HR images, which are used to train an SR model. For each method compared in the benchmark, we use the released code and settings as in the original publication. LapSRN [23] and DBPN [16] do not provide models for the upscaling factor $3\times$. Therefore, we do not evaluate their performance of the settings of $3\times$, $3 \times \_BD$ and $3 \times \_DN$. In addition, almost all of the above original codes do not provide models for the upscaling factor $16\times$. In this paper, we modify them and make them be able to work in the case of $16\times$ super-resolution. We set the number of training epochs for all methods as 1000. All models are trained using V100 GPU for about three weeks, thus the total training hours are $24 \times 7 \times 3 \times 7 \times 10 \times 2 = 70,560$. The best performance is reported in the benchmarking study. Many metrics (like PSNR, SSIM [42] and LPIPS [50]) can be used as quantitative metrics. In this paper, we use PSNR and SSIM since they are most popular for SR. Specifically, we conduct the calculation of PSNR and SSIM in the RGB space. Patch based computation is only applied for $8K$ images, which are cropped to four $4K$-resolution patches

5.3. UHDSR4K SR Dataset

We first evaluate the ten methods and our proposed MANet on the UHDSR4K images to explore their performance on 4K image super-resolution.

BI, BD and DN degradation models are widely used in SISR settings. Table 2 shows a quantitative comparison of $2\times$, $3\times$, $3 \times \_BD$, $3 \times \_DN$, $4\times$, $8\times$ and $16\times$ super-resolution settings. Among the ten state-of-the-art methods, in terms of PSNR, RDN achieves the best performance on the $2\times$, $4\times$, $8\times$, $16\times$ settings. DRLN achieves the best performance on the $3 \times \_BD$ setting, and VDSR achieves the best performance in the case of $3 \times \_DN$. In terms of SSIM, HAN achieves the best performance on the $2\times$ and $16\times$ settings. DRLN achieves the best performance on the $3\times$ and $3 \times \_BD$ settings. RDN achieves the best performance on the $3 \times \_DN$, $4\times$ and $8\times$ settings. Also, based on the results, for all methods it is generally more and more difficult to super-resolve high-quality images with the increasing of upsampling factors. The proposed MANet is based on the residual dense block from RDN, and applies a mesh attention module to capture the correlation of features from the intermediate layers. Therefore, it achieves satisfactory performance on all seven degradation settings. Specially, it outperforms the current state-of-the-art SISR methods on $2\times$, $3\times$ and $3 \times \_BD$ settings.

We also show a visual comparison of different methods on the UHDSR4K dataset for $8\times$, $3 \times \_BD$ and $3 \times \_DN$ SR in Fig. 4. We can find that though the PSNR and SSIM show difference, it is difficult to tell the difference among the qualitative results from the RCAN, RDN, HAN, DRLN and LMNet. At the same time, there still exists a clear gap between the HR images and SR results from current state-of-the-art SISR methods. As Fig. 4(b) and Fig. 4(c) show, the HR images are sharper than the SR versions. Meanwhile, in some cases, even though the SISR methods can generate sharp images, details are still missing like Fig. 4(a).

5.4. UHDSR8K SR Dataset

To evaluate the ten methods on 8K SISR, we provide the quantitative results on the UHDSR8K dataset in Table 3. Based on the PSNR values, HAN achieves the best performance on $2\times$, $3\times$, $3 \times \_BD$ degradation settings. DRLN

(a) Visual results of **BI** models ($\times 8$) on the UHDSR4K dataset.

(b) Visual results of **BD** models ($\times 3$) on the UHDSR4K dataset.

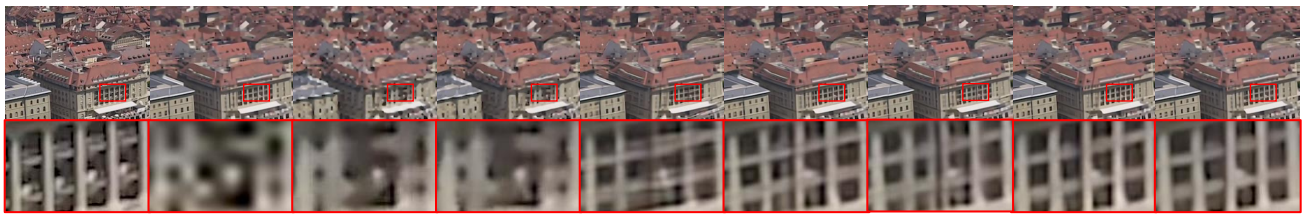
(c) Visual results with **DN** model ($\times 3$) on the UHDSR4K dataset.

Figure 4. **Visual results corresponding to different settings on the UHDSR4K dataset**. From left to right: HR, results of bicubic, SRCNN, VDSR, RCAN, RDN, HAN, DRLN, and ours. Best viewed in color.

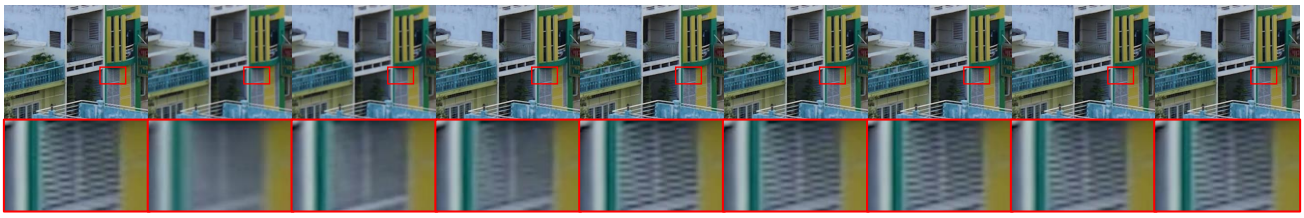
Table 3. **Performance comparison with representative methods for SISR on the UHDSR8K dataset**. Results are reported in terms of both PSNR and SSIM.

Scale	Metrics	SRCNN	FSRCNN	VDSR	LapSRN	EDSR	DBPN	RCAN	RDN	HAN	DRLN	MANet
$2\times$	PSNR	55.591	54.980	55.981	56.128	56.645	55.967	57.323	57.061	57.163	56.740	57.371
	SSIM	0.9965	0.9962	0.9967	0.9968	0.9971	0.9965	0.9972	0.9963	0.9970	0.9970	0.9973
$3\times$	PSNR	51.052	50.644	51.716	-	52.280	-	52.548	52.487	52.562	52.502	52.544
	SSIM	0.9935	0.9932	0.9938	-	0.9941	-	0.9943	0.9943	0.9944	0.9943	0.9943
$3 \times \_BD$	PSNR	44.382	44.988	46.900	-	48.599	-	48.669	48.825	48.835	48.689	48.862
	SSIM	0.9789	0.9817	0.9852		0.9887	-	0.9888	0.9890	0.9891	0.9886	0.9891
$3 \times \_DN$	PSNR	35.296	36.270	36.871	-	37.860	-	37.909	37.962	37.896	37.948	37.893
	SSIM	0.9415	0.9500	0.9546	-	0.9613	-	0.9624	0.9619	0.9623	0.9618	0.9618
$4\times$	PSNR	49.472	48.533	50.030	49.462	50.230	50.299	50.510	50.604	50.563	50.614	50.686
	SSIM	0.9911	0.9904	0.9919	0.9912	0.9919	0.9921	0.9922	0.9923	0.9922	0.9923	0.9924
$8\times$	PSNR	37.814	37.466	38.539	38.928	39.178	39.273	39.326	39.460	39.359	39.497	39.289
	SSIM	0.9486	0.9456	0.9531	0.9555	0.957	0.9577	0.9582	0.9588	0.9583	0.9592	0.9578
$16\times$	PSNR	30.794	30.632	31.388	31.924	32.141	32.206	32.475	32.491	32.514	32.535	32.463
	SSIM	0.8915	0.8912	0.8975	0.9041	0.9064	0.9069	0.9100	0.9101	0.9102	0.9108	0.9095

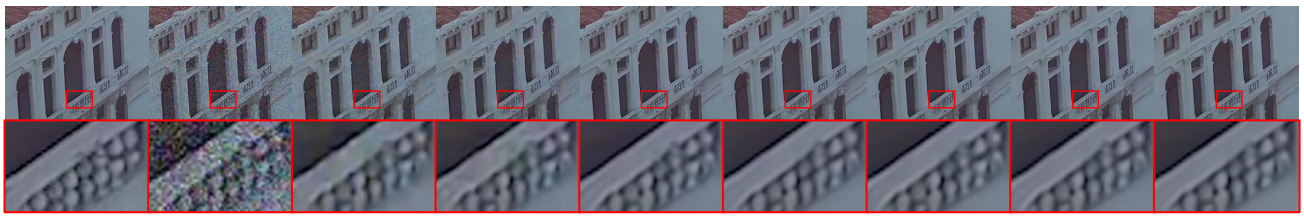
achieves the best performance on $4\times$, $8\times$ and $16\times$ settings, and RDN achieves the best performance on the $3 \times \_BD$ scenario. In terms of SSIM, the best performance on $2\times$, $3\times$, $3 \times \_BD$, $3 \times \_DN$, $4\times$, $8\times$ and $16\times$ are obtained by RCAN, HAN, HAN, RCAN, RDN (and DRLN), DRLN, and DRLN, respectively. We also find that the proposed MANet achieves satisfactory performance for the application of 8K image SR. Specially, it outperforms the current state-of-the-art SISR methods on $2\times$, $3 \times \_BD$ and $4\times$. Fig. 4 shows the visual comparison of different methods on the UHDSR8K dataset. Similar to the 4K image super-resolution, the 8K image SR also faces the problems that the super-resolved images are not sharp enough like Fig. 5(b) and 5(c), and the generated images lose details like Fig. 5(a).



(a) Visual results of **BI** model (16×) on the UHDSR8K dataset.



(b) Visual results of **BD** model (3×) on the UHDSR8K dataset.



(c) Visual results of **DN** model (3×) on the UHDSR8K dataset.

Figure 5. **Visual results corresponding to different settings on the UHDSR8K dataset**. From left to right: HR, results of bicubic, SRCNN, VDSR, RCAN, RDN, HAN, DRLN, and ours. Best viewed in color.

5.5. Discussion

The evaluation results on the UHDSR4K and UHDSR8K datasets, have led to some interesting findings.

First, compared with 2K image SR, noise and blur have greater impact in the case of 4K and 8K SR. In the 2K SR scenario, the results (PSNR) of models with and without blur and noise do not show significant differences, *e.g.*, the [29] in the case of 3× setting. However, for the UHD images SR, the difference is evident. When super-resolving an image to a UHD image (3×), noise and blur are important factors hindering the SR performance. Compared with blur, noise is the more important factor. The results of 3×, $3 \times \_BD$ and $3 \times \_DN$ in Tab. 2 & 3, Fig. 4 & 5, and [29] support this finding.

Second, as shown in Tab. 3 and Fig. 5(a), we can compress images by factors to save space and transmission bandwidth. For instance, images can be downsampled with a bicubic operation for transmission, and it can still be restored with high quality (PSNR $>= 30$). In the case of 8K images, the SR factor can even be as high as 16, while the restored quality is still satisfactory.

Third, in the case of the same SR factor, the downsampled images from 8K images provide more details than those from the 4K images, so it is easier to restore higher-quality images, and the difference is evident. For example, the results of 8K are better than 4K in the case of the 2× setting (Tab. 2 & 3).

6. Conclusion

In this paper, we explored the domain of single image super-resolution for ultra-high-definition (UHD) resolutions. We introduced two large-scale UHD SR datasets, and evaluated the ten state-of-the-art SISR methods. In addition, a baseline model, called Mesh Attention Network for SISR, was proposed to improve the representation ability of extracted features. In the future, we will add more settings, like 32× or 64×, to evaluate the extreme SR performance of current SR methods, and explore new models to super-resolve images to UHD resolution.

Acknowledgments

This work is supported in part by the NSF CAREER Grant #1149783, ARC Centre of Excellence for Robotics Vision (CE140100016), ARC-Discovery (DP 190102261) and ARC-LIEF (190100080) grants.

References

[1] Saeed Anwar and Nick Barnes. Densely residual laplacian super-resolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 5

[2] Pablo Arbelaez, Michael Maire, Charless Fowlkes, and Jitendra Malik. Contour detection and hierarchical image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010. 2

[3] Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie Line Alberi-Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In *British Machine Vision Conference*, 2012. 2

[4] Yochai Blau, Roey Mechrez, Radu Timofte, Tomer Michaeli, and Lihi Zelnik-Manor. The 2018 pirm challenge on perceptual image super-resolution. In *Proceedings of the European Conference on Computer Vision*, 2018. 2

[5] 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 International Conference on Computer Vision*, 2019. 2

[6] Hong Chang, Dit-Yan Yeung, and Yimin Xiong. Super-resolution through neighbor embedding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2004. 1

[7] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-order attention network for single image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 3

[8] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *Proceedings of the European Conference on Computer Vision*, 2014. 1, 3, 6

[9] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015. 3

[10] Chao Dong, Chen Change Loy, and Xiaoou Tang. Accelerating the super-resolution convolutional neural network. In *Proceedings of the European Conference on Computer Vision*, 2016. 2, 3, 6

[11] Weisheng Dong, Lei Zhang, Guangming Shi, and Xiaolin Wu. Image deblurring and super-resolution by adaptive sparse domain selection and adaptive regularization. *IEEE Transactions on Image Processing*, 2011. 1

[12] Gilad Freedman and Raanan Fattal. Image and video upscaling from local self-examples. *ACM Transactions on Graphics*, 2011. 1

[13] William T Freeman, Egon C Pasztor, and Owen T Carmichael. Learning low-level vision. *International Journal of Computer Vision*, 2000. 1

[14] Daniel Glasner, Shai Bagon, and Michal Irani. Super-resolution from a single image. In *Proceedings of the IEEE International Conference on Computer Vision*, 2009. 1

[15] Shuhang Gu, Andreas Lugmayr, Martin Danelljan, Manuel Fritsche, Julien Lamour, and Radu Timofte. Div8k: Diverse 8k resolution image dataset. In *Proceedings of the IEEE International Conference on Computer Vision Workshop*, pages 3512–3516. 1, 2, 3

[16] Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. Deep back-projection networks for super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 3, 6

[17] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 2

[18] Kui Jia, Xiaogang Wang, and Xiaoou Tang. Image transformation based on learning dictionaries across image spaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2012. 1

[19] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 1

[20] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 1, 3, 6

[21] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Deeply-recursive convolutional network for image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 3

[22] Kwang In Kim and Younghee Kwon. Single-image super-resolution using sparse regression and natural image prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010. 1

[23] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 3, 6

[24] 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*, 2017. 3

[25] Zhen Li, Jinglei Yang, Zheng Liu, Xiaomin Yang, Gwanggil Jeon, and Wei Wu. Feedback network for image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 3

[26] 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*, 2017. 1, 3, 6

[27] David Martin, Charless 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 of the IEEE International Conference on Computer Vision*, 2001. 2

[28] Yusuke Matsui, Kota Ito, Yuji Aramaki, Azuma Fujimoto, Toru Ogawa, Toshihiko Yamasaki, and Kiyoharu Aizawa. Sketch-based manga retrieval using manga109 dataset. *Multimedia Tools and Applications*, 2017. 2

[29] Ben Niu, Weilei Wen, Wenqi Ren, Xiangde Zhang, Lianping Yang, Shuzhen Wang, Kaihao Zhang, Xiaochun Cao, and Haifeng Shen. Single image super-resolution via a holistic attention network. In *Proceedings of the European Conference on Computer Vision*, 2020. 1, 3, 5, 8

[30] Mehdi SM Sajjadi, Bernhard Scholkopf, and Michael Hirsch. Enhancenet: Single image super-resolution through automated texture synthesis. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017. 3

[31] Samuel Schulter, Christian Leistner, and Horst Bischof. Fast and accurate image upscaling with super-resolution forests. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 1

[32] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 1

[33] Wenzhe Shi, Jose Caballero, Christian Ledig, Xiahai Zhuang, Wenjia Bai, Kanwal Bhatia, Antonio M Simoes Monteiro de Marvao, Tim Dawes, Declan O'Regan, and Daniel Rueckert. Cardiac image super-resolution with global correspondence using multi-atlas patchmatch. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2013. 1

[34] Libin Sun and James Hays. Super-resolution from internet-scale scene matching. In *IEEE International Conference on Computational Photography*, 2012. 2

[35] Ying Tai, Jian Yang, and Xiaoming Liu. Image super-resolution via deep recursive residual network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 1

[36] Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, and Lei Zhang. Ntire 2017 challenge on single image super-resolution: Methods and results. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017. 2, 3

[37] Radu Timofte, Vincent De Smet, and Luc Van Gool. Anchored neighborhood regression for fast example-based super-resolution. In *Proceedings of the IEEE International Conference on Computer Vision*, 2013. 1

[38] Radu Timofte, Rasmus Rothe, and Luc Van Gool. Seven ways to improve example-based single image super resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 2

[39] JD Van Ouwerkerk. Image super-resolution survey. *Image and Vision Computing*, 2006. 3

[40] Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2

[41] 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*, 2018. 3

[42] 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*, 2004. 6

[43] Zhihao Wang, Jian Chen, and Steven CH Hoi. Deep learning for image super-resolution: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 1, 3

[44] Zhaowen Wang, Ding Liu, Jianchao Yang, Wei Han, and Thomas Huang. Deep networks for image super-resolution with sparse prior. In *Proceedings of the IEEE International Conference on Computer Vision*, 2015. 1

[45] Chih-Yuan Yang, Chao Ma, and Ming-Hsuan Yang. Single-image super-resolution: A benchmark. In *Proceedings of the European Conference on Computer Vision*, 2014. 2

[46] Jianchao Yang, Zhe Lin, and Scott Cohen. Fast image super-resolution based on in-place example regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013. 1

[47] Jianchao Yang, John Wright, Thomas S Huang, and Yi Ma. Image super-resolution via sparse representation. *IEEE Transactions on Image Processing*, 2010. 1, 2

[48] Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In *International Conference on Curves and Surfaces*, 2010. 2

[49] Kai Zhang, Wangmeng Zuo, Shuhang Gu, and Lei Zhang. Learning deep cnn denoiser prior for image restoration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 4

[50] 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*, pages 586–595, 2018. 6

[51] 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*, 2018. 1, 3, 6

[52] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 1, 3, 4, 5

[53] Wilman WW Zou and Pong C Yuen. Very low resolution face recognition problem. *IEEE Transactions on Image Processing*, 2011. 1