Learning Dual-Level Deformable Implicit Representation for Real-World Scale Arbitrary Super-Resolution

Zhiheng Li[1], Muheng Li[1], Jixuan Fan[2], Lei Chen[1⋆], Yansong Tang[2], Jiwen Lu[1], and Jie Zhou[1]

[1] Department of Automation, Tsinghua University, China
[2] Shenzhen International Graduate School, Tsinghua University, China
{lizhihan21, li-mh20, fjx23}@mails.tsinghua.edu.cn,
leichenthu@tsinghua.edu.cn, tang.yansong@sz.tsinghua.edu.cn,
{lujiwen, jzhou}@tsinghua.edu.cn

Abstract. Scale arbitrary super-resolution based on implicit image function gains increasing popularity since it can better represent the visual world in a continuous manner. However, existing scale arbitrary works are trained and evaluated on simulated datasets, where low-resolution images are generated from their ground truths by the simplest bicubic downsampling. These models exhibit limited generalization to real-world scenarios due to the greater complexity of real-world degradations. To address this issue, we build a RealArbiSR dataset, a new real-world super-resolution benchmark with both integer and non-integer scaling factors fo the training and evaluation of real-world scale arbitrary super-resolution. Moreover, we propose a Dual-level Deformable Implicit Representation (DDIR) to solve real-world scale arbitrary super-resolution. Specifically, we design the appearance embedding and deformation field to handle both image-level and pixel-level deformations caused by real-world degradations. The appearance embedding models the characteristics of low-resolution inputs to deal with photometric variations at different scales, and the pixel-based deformation field learns RGB differences which result from the deviations between the real-world and simulated degradations at arbitrary coordinates. Extensive experiments show our trained model achieves state-of-the-art performance on the RealArbiSR and RealSR benchmarks for real-world scale arbitrary super-resolution. The dataset and code are available at https://github.com/nonozhizhiovo/RealArbiSR.

Keywords: Real-World Scale Arbitrary Super-Resolution · Deformable Implicit Neural Representation · Appearance Embedding

1 Introduction

Single image super-resolution (SISR) is a long-standing low-level task that reconstructs high-resolution (HR) images from their low-resolution (LR) inputs [11,21,

⋆ indicates the corresponding author.

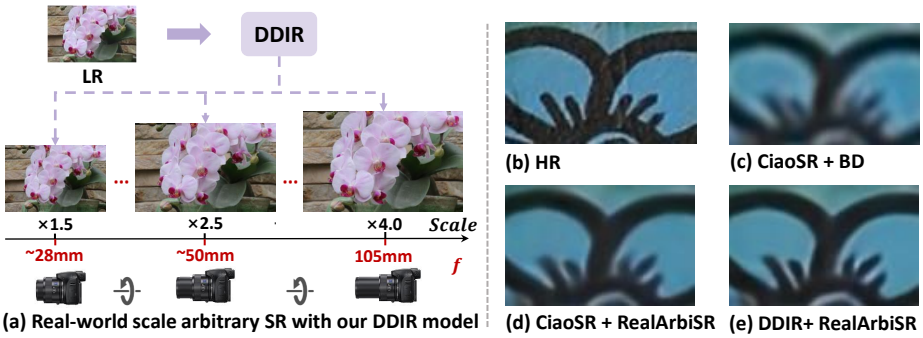

Fig. 1: (a) We propose a Dual-level Deformable Implicit Representation (DDIR) to solve real-world scale arbitrary SR, simulating the continuous optical zoom of a DSLR camera by only one model. We compare (b) the HR image with the SR results (×3.7) of a real-world LR image generated by (c) CiaoSR [5] trained on DIV2K dataset with bicubic degradation (CiaoSR+BD), (d) CiaoSR [5] trained on RealArbiSR dataset with real-world degradation (CiaoSR+RealArbiSR), and (e) our DDIR model trained on RealArbiSR dataset with real-world degradation (DDIR+RealArbiSR).

29,30,32,39,43,49,58]. It has been investigated for decades [6,12–15,18,22,53,54], and various sub-fields of SISR have been proposed [2–4,16,19,27,33,37,45,46,48, 56]. Among them, scale arbitrary super-resolution (SR) has been developed to generate HR images with arbitrary scales (even with non-integer ones) by only one model [9, 17]. It better fulfills practical needs because we demand to zoom in and zoom out continuously in daily life.

Recent works in scale arbitrary SR are either based on convolutional neural networks (CNNs) [17, 47] or implicit neural representations [9, 25, 31]. However, these methods are trained and evaluated on simulated datasets, employing bicubic degradation only. By comparing Figure 1(c) with Figure 1(d), we can see the CiaoSR model [5] trained on DIV2K dataset [44] with bicubic downsampling is ineffective to reconstruct real-world HR images. Such synthetic degradation models cannot be generalized in real-world applications because real-world degradations are much more complex [4, 51]. To handle real-world degradation kernels, real-world SR has been investigated, and current works such as LP-KPN [4] and CDC [51] are based on CNNs and predict RGB values locally. However, existing real-world SR datasets are limited to integer scale factors (*e.g.*, ×2, ×3, ×4) and their models are constrained to work for one fixed scale factor.

To fill the gap between current SR research and the practical needs of zoom functionality, we construct a RealArbiSR dataset, the first real-world SR benchmark with both integer and non-integer scale factors for the training and evaluation of real-world scale arbitrary SR. To get the focal lengths of arbitrary scale factors, we use a checkerboard to calibrate the focal lengths for the desired scales, where the checkerboard is annotated with rectangles with changing areas for different scales. According to the calibrated focal lengths, we capture LR-HR image pairs, which are further aligned by an image registration algorithm [4].

The RealArbiSR dataset provides a good benchmark for real-world scale arbitrary SR. It contains diverse indoor and outdoor scenes with both integer and non-integer scale factors.

We propose a Dual-level Deformable Implicit Representation (DDIR) to solve the real-world scale arbitrary SR, simulating the continuous optical zoom of a high-end DSLR camera, as illustrated in Figure 1(a). According to the analysis of real-world LR and HR images, we notice real-world degradation kernels can lead to image-level and pixel-level deformations on degraded images. We argue the image-level deformation is caused by photometric variations at different scales, and pixel-based deformation results from content-dependent and spatially variant degradation kernels. Therefore, we regard the problem of real-world scale arbitrary SR as a model that is deformed from the synthetic scale arbitrary SR (*e.g.*, bicubic downsampling) along the channel dimension, and thus design a dual-level deformable implicit representation to learn image deformations at the image and pixel levels. For the image level, we model the characteristics of an LR image as the appearance embedding. The appearance embedding grants our model the ability to explain away photometric deformations between different scales, and improves the SR performance by a large margin. In addition, since real-world degradation kernels are content-dependent and spatially variant, we design a deformation branch to simulate the deformation field, which is calculated as the RGB differences that result from the deviations between the real-world and synthetic degradations at arbitrary spatial coordinates. The deformation field focuses on reconstructing image details in a continuous space at the pixel level. Combining appearance embedding and deformation field, our DDIR model is capable of handling complex real-world degradation kernels and reconstructing real-world HR images with high fidelity.

In summary, our contributions are threefold. (a) We build a RealArbiSR dataset, the first real-world SR dataset with both integer and non-integer scale factors in diverse scenes. The RealArbiSR dataset provides a good SR benchmark for the training and evaluation of real-world scale arbitrary SR. (b) We propose a dual-level deformable implicit representation to learn image-level and pixel-level deformations caused by complex real-world degradation kernels. (c) Experiments show our DDIR model achieves state-of-the-art performance on the RealArbiSR and RealSR benchmarks for real-world scale arbitrary SR.

2 Related Work

Scale Arbitrary SR. Scale arbitrary SR is first proposed by Meta-SR [17]. Meta-SR utilizes the meta-upscale module to upscale LR inputs with arbitrary scales. ArbSR [47] proposes the scale-aware adaption blocks and a scale-aware upsampling layer. In addition to the CNN methods above, implicit neural representation [1, 10, 20, 34–36, 41] has been widely used. LIIF [9] makes the first attempt by using the local implicit image function. Following LIIF, LTE [25] designs a local texture estimator to synthesize HR images in the Fourier domain. UltraSR [52] introduces positional encoding with residual connections to

the LIIF model to enhance the SR performance. OPE-SR [42] proposes orthogonal position encoding for scale arbitrary SR. [50] designs super-resolution neural operator to learn the mapping between function spaces. CiaoSR [5] proposes the continuous implicit attention-in-attention network for scale arbitrary SR. ITSRN [55] designs an implicit transformer network to solve screen content SR with arbitrary scales. Different from the previous works, IPF [31] introduces implicit pixel flow to generate perceptual-oriented HR results. However, current works of scale arbitrary SR all generate LR inputs by bicubic downsampling, which cannot simulate SR in real-world situations. [8] solves real-world scale arbitrary SR by reverse modules SASRN and SARDN, but it is only trained and tested at the scale factor of $\times 2$.

Real-World SR Datasets. Different from synthetic datasets, LR-HR image pairs in most real-world SR datasets are captured by adjusting the focal length of the camera. Qu et al. [38] use a beam splitter to collect LR-HR image pairs with two cameras. SupER dataset [24] uses hardware binning to generate corresponding LR versions of ground truths. City100 dataset [7] contains 100 aligned image pairs that are captured from the printed postcards. Just like City100, D2CRealSR dataset [26] takes photos of postcards in the laboratory environment to get image pairs with large scaling factors. SR-RAW dataset [57] is the first real-world SR dataset captured in natural scenes, but their image pairs are not well aligned. RealSR dataset [4] provides a good real-world SR benchmark by using an image registration algorithm to precisely align image pairs. Then, a large-scale DRealSR dataset [51] is constructed. However, all the pixel-aligned real-world SR datasets captured in the indoor and outdoor environments [4, 51] only consist of image pairs with integer scale factors and thus are insufficient for the training and evaluation of real-world scale arbitrary SR.

Real-World SR Methods. In contrast to the bicubic downsampling kernel, real-world degradation kernels are much more complicated because they are spatially variant and content-dependent. Zhang et al. [57] propose a contextual bilateral loss for real-world SR. LP-KPN [4] introduces a Laplacian pyramid network to learn spatially variant kernels and reconstruct HR images. CDC [51] parses an image into three low-level components and proposes a component divide-and-conquer model to reconstruct HR images. DDet [40] introduces a dual-path dynamic enhancement network. STCN [60] designs a spatio-temporal correlation network and proposes a dual restriction to reduce the space of mapping functions in the real world. D2C-SR [26] proposes a novel framework with divergence and convergence stages for real-world SR. These methods are based on CNN models, and only work for one specific integer scale factor.

3 The RealArbiSR Dataset

3.1 Camera Calibration

One significant feature of synthetic scale arbitrary SR is that it can predict HR images even at non-integer scale factors, where LR inputs can be easily generated by setting a bicubic scaling factor. Currently, real-world SR datasets only consist

Table 1: Comparisons of the scale factors in the training set and test set between our RealArbiSR dataset and the existing pixel-aligned real-world SR datasets captured in the indoor and outdoor environments.

Dataset	Scale Factor	
	Train set	Test set
RealSR [4]	×2.0, ×3.0, ×4.0	×2.0, ×3.0, ×4.0
DRealSR [51]	×2.0, ×3.0, ×4.0	×2.0, ×3.0, ×4.0
RealArbiSR (Ours)	×1.5, ×2.0, ×2.5, ×3.0, ×3.5, ×4.0	×1.5, ×1.7, ×2.0, ×2.3, ×2.5, ×2.7, ×3.0, ×3.3, ×3.5, ×3.7, ×4.0

of LR-HR image pairs with multiple integer scale factors such as ×2, ×3, and ×4 [4, 51]. Hence, no real-world dataset can be used to train and evaluate at non-integer scaling factors for real-world scale arbitrary SR. In the real-world situation, arbitrary scale factors especially non-integer ones are hard to indicate, because in general only a sparse set of focal length values are labeled on the zoom lens (*e.g.*, Canon 24∼105mm, $f/4.0$ zoom lens only displays the focal lengths of 105mm, 50mm, 35mm, and 24mm on the zoom ring). Also, the relation between the scaling factor and the focal length can be nonlinear.

To get the desired focal lengths for arbitrary scaling factors, we use a checkerboard to calibrate the focal length of the zoom lens. The checkerboard is annotated with rectangles of various sizes for different scaling factors, as illustrated in the supplementary material. Considering the aberration effect is more severe with a wider angle of view, we take the images captured at the longest focal length as the ground truth of all scales, which corresponds to the smallest rectangle or equivalently the smallest field of view. In this way, the aberration effect can be minimized in all LR-HR image pairs. The widths and heights of other rectangles are enlarged by the desired scaling factors compared to the smallest ground-truth rectangle, so larger rectangles correspond to the LR inputs with larger scaling factors. During calibration, we first match the field of view of the longest focal length with the ground-truth rectangle by adjusting the camera position. After matching, we fix the camera at this steady position on a tripod. Then, we reduce the focal length to increase the field of view of the camera to match the larger rectangle of the desired scale factor. In this way, we can indicate and record the calibrated focal lengths for all scaling factors. For the training set, we collect the LR-HR image pairs of scaling factors from ×1.5 to ×4 with a step of ×0.5 (including ×1.5, ×2.0, ×2.5, ×3.0, ×3.5, and ×4.0). For the test set, in addition to the scale factors that appeared in the training set (including ×1.5, ×2.0, ×2.5, ×3.0, ×3.5, and ×4.0), we further collected image pairs of the scale factors that are not present in the training set (*e.g.*, ×1.7, ×2.3, ×2.7, ×3.3, and ×3.7). We summarize the scale factors of our RealArbiSR dataset and compare with existing pixel-aligned real-world SR datasets captured in the indoor and outdoor environments in Table 1.

3.2 Dataset Collection

We use the DSLR camera (Canon 5D3) to capture LR-HR image pairs for dataset collection. The DSLR camera is equipped with a 24∼105mm, $f/4.0$ zoom lens to cover the range of target scaling factors. We set the focal length of 105mm as the ground truth and the focal lengths of the LR inputs are indicated by the calibration procedure. When collecting images, the camera is fixed on a tripod. We first capture the ground-truth image at the focal length of 105mm, and then gradually zoom out the camera to collect LR inputs according to the calibrated focal lengths. We collect image pairs in diverse indoor and outdoor scenes to ensure our RealArbiSR dataset is generalized. We prefer to photograph objects at a distance of at least 3 meters. A large object distance can alleviate image deformations caused by aberrations. Also, we avoid photographing moving objects, since they are impossible to be aligned between image pairs. After collecting the ground truth and LR versions of all scale factors, we adopt the image registration algorithm [4] to obtain pixel-wise aligned image pairs. For the scale factors from $\times 1.5$ to $\times 4.0$ with a step of $\times 0.5$, we get 250 scenes with 1500 LR-HR image pairs in total (Each scene has six image pairs for six scaling factors). 200 scenes are randomly chosen as the training set and the other 50 scenes are used as the test set. We further collect 83 scenes for the scale factors of $\times 1.7$, $\times 2.3$, $\times 2.7$, $\times 3.3$, and $\times 3.7$ as the test set. More details of the camera setting, image registration process, and the resolutions of the LR and HR versions at different scaling factors are discussed in the supplementary material.

4 Methods

4.1 Analysis of Real-World Scale Arbitrary SR

Before introducing our approach, we conduct a detailed analysis to compare the difference between the real-world scale arbitrary SR and the synthetic scale arbitrary SR. It better explains the motivation of our DDIR model.

Photometric Variation: In real-world photographs, photometric variations on exposure, tone-mapping, white balance, etc., are unavoidable because the camera imaging pipeline is very complex. After changing the focal length of the camera, factors such as lighting conditions inside the field of view, camera sensors, image signal processing (ISP) pipeline, etc., can be all varied, leading to appearance variations over the whole image. To demonstrate such image-level variations, we compare the colour histogram of the ground-truth image with its bicubic-upscaled LR counterpart in Figure 2(d), illustrating the RGB values of LR images are generally shifted away from the ground-truth values. We argue these real-world photometric variations can be regarded as an image-level deformation from the template of synthetic degradation. Most existing real-world SR methods [4, 51] cannot solve this issue because they predict kernels and reconstruct high-resolution RGB targets locally. For the same reason, local implicit neural representation [9,25,31] used in existing scale arbitrary SR works is insufficient to solve this real-world task since its receptive field is limited.

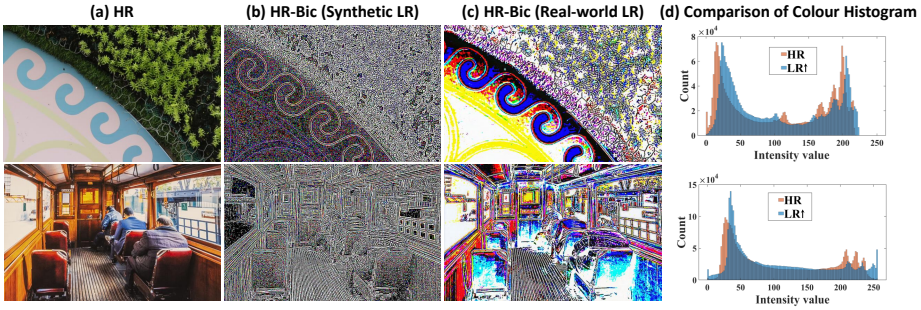

Fig. 2: (a) The ground-truth images; (b) The images computed by subtracting the ground truths with their synthetic low-resolution versions which have been bicubically upscaled to the same resolution as the ground truth; (c) The images computed by subtracting the ground truths with their real-world low-resolution versions (bicubically upscaled); (d) The comparison of colour histograms between the ground truths and their real-world low-resolution versions (bicubically upscaled).

Pixel-based Deformation: One difficulty of real-world scale arbitrary SR is that the model needs to predict real-world degradations at arbitrary spatial coordinates. To visualize and compare the effect of degradation kernels in real-world and synthetic scale arbitrary SR, we compute the RGB difference between the ground truth and bicubic-upscaled LR input in Figure 2. From Figure 2(b), we can see the RGB difference in synthetic scale arbitrary SR is only significant in high-frequency regions, including sharp edges and textures, with minor colour mismatching. In contrast, the RGB difference in real-world scale arbitrary SR is much more complex. In Figure 2(c), we find illumination and colour mismatches occur everywhere regardless of low-frequency or high-frequency regions in the real-world case. These mismatches are caused by content-dependent and spatially variant degradation kernels.

According to the analysis above, we propose dual-level deformable implicit representation to address both image-level and pixel-level deformations in real-world scenarios. In Section 4.3, we introduce an appearance embedding to address the deformation at the image level. In Section 4.4, we design a deformation branch to model the deformation field and reconstruct image details at the pixel level. Figure 3 shows the training pipeline of our DDIR model.

4.2 Overview

The overall architecture of our DDIR model is illustrated in Figure 3. It is composed of two branches, which are the deformation branch and the SR branch. Each branch consists of one encoder and one decoding function, taking the pixel coordinate (x, y) and the LR image as the inputs. Both decoding functions are parameterized by MLPs, and use local implicit neural representation to predict RGB values at query coordinates. Therefore, the prediction of the RGB values I_q at an arbitrary query coordinate x_q by a decoding function f with trainable

weights θ can be formulated as:

$$I_q = \sum_i \frac{S_i}{S} \cdot f_\theta(m_i^*, x_q - x_i^*), \tag{1}$$

where $m_i^* (i \in \{00, 01, 10, 11\})$ are the nearest latent code at the top-left, top-right, bottom-left, and bottom-right coordinates x_i^* respectively, S_i is the rectangle area between x_i^* and x_q, and S is the sum of all four S_i. Cell decoding and feature unfolding are also used.

4.3 Appearance Embedding

The receptive field of local implicit neural representation is limited. To adapt our DDIR model to variable photometric variations, we introduce an appearance embedding to the deformation branch to handle the image-level deformation. Here, the appearance embedding is simply taken as the spatial average pooling of the 2D feature map from the encoder E_ϕ^{sr} of the SR branch. Thus, the appearance embedding l_a can be formulated as:

$$l_a = \frac{1}{WH} \sum_{x=1}^{W} \sum_{y=1}^{H} m_{x,y}^*, \tag{2}$$

where $m_{x,y}^*$ is the latent code at the coordinate of x and y, and W and H are the width and height of the 2D feature map respectively. After getting the appearance embedding of the LR input, we concatenate it with the nearest latent code from the query coordinate x_q. With the appearance embedding, our DDIR model can 'see' the characteristics of the whole LR input and is not purely local anymore. Although the appearance embedding is simply the spatial average pooling of the 2D feature map, we will show it can largely improve the metric results in experiments, especially in the real-world case.

4.4 Deformation Field

There is no way we can model the exact form of degradation kernels in real-world scale arbitrary SR because they are too complex to be known. Instead of directly predicting the form of degradation kernels, we simulate the effects of degradation kernels on RGB values. We regard bicubic downsampling as linear degradation and real-world downsampling as nonlinear degradation. We design the deformation branch to predict RGB differences that result from the derivations between real-world and synthetic degradation at arbitrary coordinates. More specifically, the RGB output of this branch is supervised by the RGB difference between the ground truth and the bicubic-upscaled LR input at this query point. We define the target of this branch as the deformation field $\Delta I(x_q)$ because it models the pixel-level deformation between the nonlinear and linear degradations at the arbitrary coordinate x_q, which can be computed as:

$$\Delta I(x_q) = I^{GT}(x_q) - I^{LR\uparrow}(x_q), \tag{3}$$

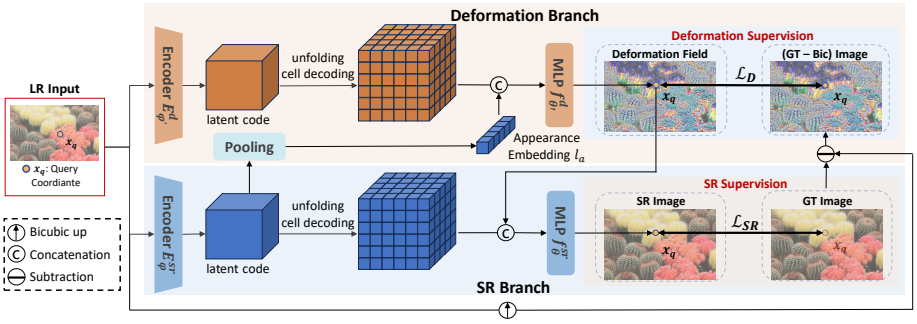

Fig. 3: The training pipeline of our DDIR model. It consists of double branches, which are the deformation branch and the SR branch. Each branch is composed of an encoder and an MLP, taking the LR image and query coordinates as the inputs. The appearance embedding l_a is computed as the spatial average pooling of the 2D feature map from the encoder E_{ϕ}^{sr} of the SR branch, which is fed into the decoding function $f_{\theta'}^{d}$ of the deformation branch by concatenation. The RGB output of the deformation branch is supervised by the deformation field. Then, the predicted deformation field feeds into the decoding function f_{θ}^{sr} of the SR branch by concatenation. Finally, the decoding function f_{θ}^{sr} of the SR branch outputs the target high-resolution RGB values at the query coordinates. Combining the appearance embedding and the deformation field, our DDIR model learns the dual-level deformable implicit representation to address the deformations at the image and pixel levels simultaneously.

where $I^{GT}(x_q)$ is the RGB value of the ground truth at the query coordinate x_q, and $I^{LR\uparrow}(x_q)$ is the RGB value of the upscaled LR (upscaled to the same size as the ground truth by bicubic) at the query coordinate x_q. By taking the residual between the ground truth and bicubic-upscaled LR input, the deformation field can simulate the spatially variant degradation effect and reconstruct texture details at the pixel level. Combining the appearance embedding and deformation field, our DDIR model learns a dual-level deformable implicit representation to address the image-level and pixel-level deformations simultaneously.

4.5 Network Architecture and Training

The training pipeline of our DDIR model is shown in Figure 3. The network components and parameters of the two branches are separated. The deformation branch outputs the RGB values, supervised by the deformation field (Deformation Supervision). The SR branch outputs the target high-resolution RGB values, supervised by the ground truth (SR Supervision). In the deformation branch, the appearance embedding l_a concatenates with the latent code from the deformation encoder $E_{\phi'}^{d}$ before feeding into the decoding function $f_{\theta'}^{d}$ of the deformation branch. In the SR branch, the predicted deformation field concatenates with the latent code from the SR encoder E_{ϕ}^{sr} before feeding into the decoding function f_{θ}^{sr} of the SR branch. The losses of the SR Supervision $\mathcal{L}_{SR}$ and the Deformation Supervision $\mathcal{L}_D$ are both L1 losses. Thus, the final loss $\mathcal{L}$

is formulated as the sum of these two losses:

$$\mathcal{L} = \mathcal{L}_{SR} + \mathcal{L}_{D}. \tag{4}$$

In inference, there is no computation of the bicubic-upscaled LR image and the subtraction between the ground-truth image and the bicubic-upscaled LR image, compared to the training pipeline.

5 Experiment

5.1 Experiment Setup

We use the RealArbiSR dataset and the RealSR dataset [4] for experiments of real-world scale arbitrary SR. The RealArbiSR dataset has 200 image pairs for training and either 50 or 83 image pairs for testing with various integer and non-integer scale factors. The RealSR dataset has around 400 image pairs for training and 100 image pairs for testing with the integer scale factors of $\times 2$, $\times 3$, and $\times 4$. We train the RealArbiSR dataset with the scale factors of $\times 1.5$, $\times 2.0$, $\times 2.5$, $\times 3.0$, $\times 3.5$, and $\times 4.0$, and test at the scale factors of $\times 1.5$, $\times 1.7$, $\times 2.0$, $\times 2.3$, $\times 2.5$, $\times 2.7$, $\times 3.0$, $\times 3.3$, $\times 3.5$, $\times 3.7$, and $\times 4.0$. The RealSR dataset is trained and tested with the scale factors of $\times 2$, $\times 3$, and $\times 4$. In the training time, we crop 48×48 patches as the inputs to the encoder. The corresponding HR patch with a random scale factor is also cropped as the ground-truth counterpart. 2304 pixels are randomly sampled from the ground-truth patch, and converted to coordinate-RGB pairs. We evaluate PSNR on the Y channel (*e.g.*, luminance) of the transformed YCbCr space [4, 51].

The encoders E_{ϕ}^{sr} and $E_{\phi'}^{d}$ of both branches are either EDSR [28] or RDN [59] without the upsampling module. Both decoding functions f_{θ}^{sr} and $f_{\theta'}^{d}$ are 5-layer MLPs with ReLU activations and hidden dimensions of 256. we use an Adam [23] optimizer with an initial learning rate of 2×10^{-4}, which decays by 0.5 at every 200 epochs. The batch size is 16 and the models are trained for 1000 epochs. The last epoch is used for the final results. Experiments are conducted on two GeForce RTX 3090 GPUs.

5.2 Comparisons with State-of-the-Art

Quantitative Results. In Table 2, we compare the quantitative results between LIIF [9], LTE [25], CiaoSR [5], and our DDIR, using EDSR and RDN without the upsampling module as the encoders. Prior work [8] is not included because its results are worse than the LIIF baseline. We can see our DDIR model achieves the best PSNR results at all the scale factors that appeared in the training set in the RealArbiSR dataset (including $\times 1.5$, $\times 2.0$, $\times 2.5$, $\times 3.0$, $\times 3.5$, and $\times 4.0$) and the RealSR dataset (*e.g.*, $\times 2.0$, $\times 3.0$, and $\times 4.0$). In particular, compared with the previous SOTA method [5], our DDIR model achieves remarkable PSNR gains of 0.32dB under the RDN backbone ($\times 2.0$) on the RealArbiSR dataset. Even for the scale factors that are not present in the training set (such as $\times 1.7$,

Table 2: Quantitative comparison on RealArbiSR and RealSR datasets in PSNR(dB). The highest PSNR at each scale factor of each dataset is bolded. One model is trained and tested at the scale factors from ×1.5 to ×4.0 with a step of ×0.5 in RealArbiSR dataset. Another model is trained and tested at the integer scale factors including ×2.0, ×3.0, and ×4.0 in RealSR dataset.

Method	RealArbiSR						RealSR		
	×1.5	×2.0	×2.5	×3.0	×3.5	×4.0	×2.0	×3.0	×4.0
Bicubic [4]	35.46	32.45	30.69	29.42	28.50	27.80	31.67	28.61	27.24
EDSR-baseline [28]	-	34.26	-	31.12	-	29.47	33.88	30.86	29.09
EDSR-LIIF [9]	37.14	34.37	32.54	31.28	30.29	29.63	33.87	30.77	29.18
EDSR-LTE [25]	37.16	34.34	32.53	31.26	30.30	29.68	33.94	30.80	29.21
EDSR-CiaoSR [5]	37.23	34.54	32.80	31.52	30.57	29.88	34.08	30.97	29.37
EDSR-DDIR	**37.51**	**34.85**	**33.02**	**31.78**	**30.80**	**30.05**	**34.19**	**31.02**	**29.39**
RDN-LIIF [9]	37.14	34.41	32.60	31.40	30.34	29.70	33.99	30.90	29.29
RDN-LTE [25]	37.24	34.52	32.76	31.53	30.54	29.84	34.01	30.93	29.29
RDN-CiaoSR [5]	37.38	34.70	32.96	31.68	30.77	30.07	34.26	31.14	29.45
RDN-DDIR	**37.63**	**35.02**	**33.20**	**31.91**	**30.94**	**30.21**	**34.35**	**31.15**	**29.48**

Table 3: Quantitative comparison on RealArbiSR in PSNR(dB). The highest PSNR at each scale factor of each dataset is bolded. One model is trained at the scale factors from ×1.5 to ×4.0 with a step of ×0.5, but tested at the scale factors of ×1.7, ×2.3, ×2.7, ×3.3, and ×3.7 in RealArbiSR dataset.

Method	EDSR Backbone					RDN Backbone				
	×1.7	×2.3	×2.7	×3.3	×3.7	×1.7	×2.3	×2.7	×3.3	×3.7
Bicubic	33.53	31.05	30.12	29.03	28.48	33.53	31.05	30.12	29.03	28.48
LIIF [9]	34.63	32.33	31.39	30.22	29.64	34.66	32.40	31.45	30.28	29.71
LTE [25]	34.65	32.29	31.30	30.10	29.51	34.74	32.44	31.55	30.39	29.81
CiaoSR [5]	34.49	32.44	31.64	30.48	29.87	34.54	32.50	31.67	30.56	29.96
DDIR	**34.90**	**32.73**	**31.80**	**30.61**	**30.01**	**35.07**	**32.88**	**31.96**	**30.75**	**30.15**

×2.3, ×2.7, ×3.3, and ×3.7), our DDIR model also achieves the best PSNR results at all these scale factors, as illustrated in Table 3. These experimental results show the appearance embedding and the deformation field handle real-world degradation kernels from the perspectives of image-level and pixel-level deformations properly, resulting in robust metric gains at all the scale factors under both backbones in the RealArbiSR and RealSR datasets.

We further conduct the cross-dataset testing in Table 4. We train the models on the RealArbiSR dataset and test them on the RealSR dataset. Table 4 shows our DDIR model achieves the best metric results in the cross-dataset experiment.

Qualitative Results. We present a qualitative comparison between LIIF [9], LTE [25], CiaoSR [5] and our DDIR in Figure 4. It shows our DDIR model obtains better visual quality than the competitors, reconstructing sharper edges

Table 4: Quantitative comparison of the cross-dataset testing in PSNR(dB). The highest PSNR at each scale factor is bolded. One model is trained at the scale factors from ×1.5 to ×4.0 with a step of ×0.5 in RealArbiSR dataset, and tested at the scale factors of ×2.0, ×3.0, and ×4.0 in RealSR dataset.

Method	EDSR Backbone			RDN Backbone		
	×2.0	×3.0	×4.0	×2.0	×3.0	×4.0
LIIF [9]	32.47	29.54	28.03	32.39	29.54	28.02
LTE [25]	32.48	29.52	28.06	32.27	29.47	28.06
CiaoSR [5]	32.38	29.56	28.19	32.42	29.58	28.17
DDIR	**32.58**	**29.72**	**28.25**	**32.59**	**29.70**	**28.18**

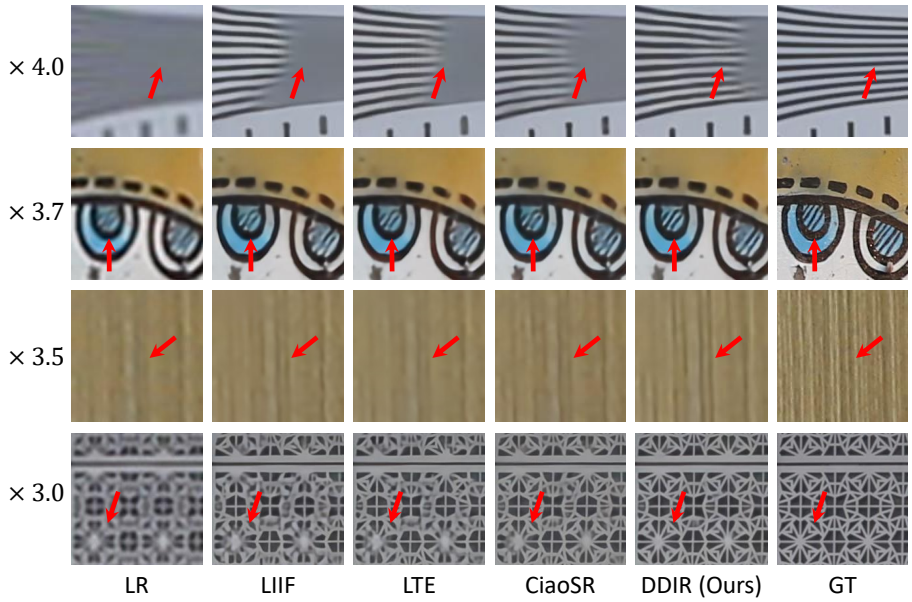

Fig. 4: Qualitative comparisons between different methods on benchmarks. Zoom in to have better views.

and more natural details. In contrast, the results of other methods [5, 9, 25] suffer from unpleasant details, especially blurry edges. Taking the first group of the images as an example, our DDIR can reconstruct more lines with sharp edges. However, more of the lines in other methods are blurred. These qualitative comparisons prove our DDIR reconstructs HR images with better texture details due to the use of appearance embedding and deformation field.

5.3 Analysis of Scale Factors in RealArbiSR Dataset

Compared with the existing real-world SR dataset (*e.g.*, RealSR [4] and DRealSR dataset [51]), our RealArbiSR dataset has three more non-integer scale factors

Table 5: Quantitative Analysis of training scale factors in RealArbiSR dataset. The highest PSNR at each scale factor on each method is bolded. '×2×3×4' represents the models are trained at the scale factors of ×2.0, ×3.0, and ×4.0. 'All' represents the models are trained at the scale factors of ×1.5, ×2.0, ×2.5, ×3.0, ×3.5, and ×4.0.

Method	Training Scale	×1.5	×2.0	×2.5	×3.0	×3.5	×4.0
EDSR-LIIF [9]	×2×3×4	36.70	34.20	32.39	31.19	30.22	29.59
	All	**37.14**	**34.37**	**32.54**	**31.28**	**30.29**	**29.63**
EDSR-LTE [25]	×2×3×4	36.89	34.23	32.40	31.18	30.21	29.59
	All	**37.16**	**34.34**	**32.53**	**31.26**	**30.29**	**29.68**
EDSR-CiaoSR [5]	×2×3×4	36.85	34.45	32.69	31.45	30.49	29.82
	All	**37.23**	**34.54**	**32.80**	**31.52**	**30.57**	**29.88**
EDSR-DDIR	×2×3×4	37.21	34.63	32.80	31.61	30.61	29.90
	All	**37.51**	**34.85**	**33.02**	**31.78**	**30.80**	**30.05**

Table 6: Quantitative ablation study of EDSR-DDIR on RealArbiSR dataset in PSNR(dB). The highest PSNR at each scale factor is bolded.

Deformation Field	Appearance Embedding	Scale ×1.5	×2.0	×2.5	×3.0	×3.5	×4.0
✗	✗	37.09	34.32	32.52	31.29	30.33	29.69
✓	✗	37.26	34.50	32.66	31.41	30.41	29.72
✗	✓	37.32	34.69	32.86	31.61	30.64	29.90
✓	✓	**37.51**	**34.85**	**33.02**	**31.78**	**30.80**	**30.05**

including ×1.5, ×2.5, and ×3.5 in the training set. To demonstrate our RealArbiSR dataset is more suitable for the training of real-world scale arbitrary SR due to the presence of these non-integer scale factors, we compare the metric results of the models trained with only integer scale factors and with all scale factors in Table 5. We can see the models which are trained at all scale factors (including ×1.5, ×2.0, ×2.5, ×3.0, ×3.5, and ×4.0, indicated as 'All' in Table 5) perform better than the ones trained only at integer scale factors (including ×2.0, ×3.0, and ×4.0, indicated as '×2×3×4' in Table 5). Further experimental results are presented in the supplemental material.

5.4 Ablation Study

We show an ablation study in Table 6 to demonstrate the effect of the appearance embedding and deformation field. After removing appearance embedding, the PSNR results are reduced at all scale factors. To illustrate the effect of deformation field, we remove the deformation field and the branch, concatenating the appearance embedding with the latent code of the SR branch. In this case, the PSNR results also drop at all scale factors. Without the appearance embedding and deformation field, the model performs the worst at all these scale factors.

Table 7: Quantitative comparison of appearance embedding in synthetic and real-world scale arbitrary SR in PSNR(dB) on RGB channels. EDSR-LIIF(+a) refers to adding appearance embedding to the EDSR-LIIF baseline. The highest PSNR at each scale factor is bolded. The models are trained and evaluated on RealArbiSR and DIV2K datasets. RealArbiSR dataset is tested at the scale factors from ×1.5 to ×4.0 with a step of ×0.5 and DIV2K dataset is tested at the scale factors of ×2.0, ×3.0, and ×4.0.

Method	RealArbiSR						DIV2K		
	×1.5	×2.0	×2.5	×3.0	×3.5	×4.0	×2.0	×3.0	×4.0
EDSR-LIIF [9]	34.88	32.27	30.49	29.27	28.29	27.62	34.67	30.96	29.00
EDSR-LIIF(+a)	**35.09**	**32.64**	**30.88**	**29.64**	**28.68**	**27.96**	**34.74**	**31.05**	**29.07**

5.5 Analysis of Appearance Embedding on Real-World SR

To demonstrate appearance embedding is particularly useful in real-world scale arbitrary SR, we compare the experimental results with and without the appearance embedding in the synthetic (*e.g.*, bicubic downsampling) and the real-world scale arbitrary SR. Specifically, we use the EDSR-LIIF [9] model as the baseline and choose to concatenate or not concatenate the appearance embedding with the latent code before feeding into the MLP. We train and evaluate these models on DIV2K dataset (bicubic degradation) [44] and our RealArbiSR dataset. In Table 7, we can see the PSNR gains by adding the appearance embedding in the real-world dataset are significantly higher than the gains in the synthetic dataset. This proves the appearance embedding is much more useful in the real-world case. Since bicubic downsampling cannot have image-level deformations, we argue there exist image-level deformations in real-world SR. By adding appearance embedding, our DDIR model can handle image-level degradations caused by photometric variations in real-world photographs, leading to a more significant improvement in the real-world scenario than the synthetic one.

6 Conclusion

In this work, we contribute a RealArbiSR dataset, the first real-world SR benchmark with both integer and non-integer scale factors in diverse scenes for the training and evaluation of real-world scale arbitrary SR. We propose dual-level deformable implicit representation to solve this problem. Specifically, the appearance embedding and deformation field are designed to handle image-level and pixel-level deformations caused by real-world degradation kernels. Extensive experiments show our DDIR model is capable of dealing with complex real-world degradations and reconstructing real-world HR images with high fidelity, achieving state-of-the-art performance on both RealArbiSR and RealSR benchmarks. As for limitations, our RealArbiSR dataset uses one camera to collect image pairs. The differences in imaging pipelines among cameras can lead to more diverse degradations. In the future, we will build a large-scale dataset by using more DSLR cameras to cover diverse real-world degradation among devices.

Acknowledgement. This work was supported in part by the National Natural Science Foundation of China under Grant 62321005, Grant 62336004, Grant 62125603, and Grant 62306031.

References

1. Atzmon, M., Lipman, Y.: Sal: Sign agnostic learning of shapes from raw data. In: CVPR. pp. 2565–2574 (2020)
2. Bell-Kligler, S., Shocher, A., Irani, M.: Blind super-resolution kernel estimation using an internal-gan. NeurIPS **32** (2019)
3. Blau, Y., Mechrez, R., Timofte, R., Michaeli, T., Zelnik-Manor, L.: The 2018 pirm challenge on perceptual image super-resolution. In: ECCVW. pp. 0–0 (2018)
4. Cai, J., Zeng, H., Yong, H., Cao, Z., Zhang, L.: Toward real-world single image super-resolution: A new benchmark and a new model. In: ICCV. pp. 3086–3095 (2019)
5. Cao, J., Wang, Q., Xian, Y., Li, Y., Ni, B., Pi, Z., Zhang, K., Zhang, Y., Timofte, R., Van Gool, L.: Ciaosr: Continuous implicit attention-in-attention network for arbitrary-scale image super-resolution. arXiv preprint arXiv:2212.04362 (2022)
6. Chang, H., Yeung, D.Y., Xiong, Y.: Super-resolution through neighbor embedding. In: CVPR. vol. 1, pp. I–I (2004)
7. Chen, C., Xiong, Z., Tian, X., Zha, Z.J., Wu, F.: Camera lens super-resolution. In: CVPR. pp. 1652–1660 (2019)
8. Chen, H., He, X., Yang, H., Wu, Y., Qing, L., Sheriff, R.E.: Self-supervised cycle-consistent learning for scale-arbitrary real-world single image super-resolution. Expert Syst. Appl. **212**, 118657 (2023)
9. Chen, Y., Liu, S., Wang, X.: Learning continuous image representation with local implicit image function. In: CVPR. pp. 8628–8638 (2021)
10. Chen, Z., Zhang, H.: Learning implicit fields for generative shape modeling. In: CVPR. pp. 5939–5948 (2019)
11. Dong, C., Loy, C.C., He, K., Tang, X.: Learning a deep convolutional network for image super-resolution. In: ECCV. pp. 184–199 (2014)
12. Dong, W., Zhang, L., Shi, G., Wu, X.: Image deblurring and super-resolution by adaptive sparse domain selection and adaptive regularization. TIP **20**(7), 1838–1857 (2011)
13. Freedman, G., Fattal, R.: Image and video upscaling from local self-examples. ACM Trans. Graph. **30**(2), 1–11 (2011)
14. Freeman, W.T., Pasztor, E.C., Carmichael, O.T.: Learning low-level vision. IJCV **40**, 25–47 (2000)
15. Glasner, D., Bagon, S., Irani, M.: Super-resolution from a single image. In: ICCV. pp. 349–356 (2009)
16. Gu, J., Lu, H., Zuo, W., Dong, C.: Blind super-resolution with iterative kernel correction. In: CVPR. pp. 1604–1613 (2019)
17. Hu, X., Mu, H., Zhang, X., Wang, Z., Tan, T., Sun, J.: Meta-sr: A magnification-arbitrary network for super-resolution. In: CVPR. pp. 1575–1584 (2019)
18. Huang, J.B., Singh, A., Ahuja, N.: Single image super-resolution from transformed self-exemplars. In: CVPR. pp. 5197–5206 (2015)
19. Huang, Y., Li, S., Wang, L., Tan, T., et al.: Unfolding the alternating optimization for blind super resolution. NeurIPS **33**, 5632–5643 (2020)

20. Jiang, C., Sud, A., Makadia, A., Huang, J., Nießner, M., Funkhouser, T., et al.: Local implicit grid representations for 3d scenes. In: CVPR. pp. 6001–6010 (2020)
21. Kim, J., Lee, J.K., Lee, K.M.: Accurate image super-resolution using very deep convolutional networks. In: CVPR. pp. 1646–1654 (2016)
22. Kim, K.I., Kwon, Y.: Single-image super-resolution using sparse regression and natural image prior. TPAMI **32**(6), 1127–1133 (2010)
23. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
24. Köhler, T., Bätz, M., Naderi, F., Kaup, A., Maier, A., Riess, C.: Toward bridging the simulated-to-real gap: Benchmarking super-resolution on real data. TPAMI **42**(11), 2944–2959 (2019)
25. Lee, J., Jin, K.H.: Local texture estimator for implicit representation function. In: CVPR. pp. 1929–1938 (2022)
26. Li, Y., Huang, H., Jia, L., Fan, H., Liu, S.: D2c-sr: A divergence to convergence approach for real-world image super-resolution. In: ECCV. pp. 379–394 (2022)
27. Liang, J., Zhang, K., Gu, S., Van Gool, L., Timofte, R.: Flow-based kernel prior with application to blind super-resolution. In: CVPR. pp. 10601–10610 (2021)
28. Lim, B., Son, S., Kim, H., Nah, S., Mu Lee, K.: Enhanced deep residual networks for single image super-resolution. In: CVPRW. pp. 136–144 (2017)
29. Ma, C., Jiang, Z., Rao, Y., Lu, J., Zhou, J.: Deep face super-resolution with iterative collaboration between attentive recovery and landmark estimation. In: CVPR. pp. 5569–5578 (2020)
30. Ma, C., Rao, Y., Cheng, Y., Chen, C., Lu, J., Zhou, J.: Structure-preserving super resolution with gradient guidance. In: CVPR. pp. 7769–7778 (2020)
31. Ma, C., Yu, P., Lu, J., Zhou, J.: Recovering realistic details for magnification-arbitrary image super-resolution. TIP **31**, 3669–3683 (2022)
32. Ma, C., Zhang, J., Zhou, J., Lu, J.: Learning series-parallel lookup tables for efficient image super-resolution. In: ECCV. pp. 305–321 (2022)
33. Michaeli, T., Irani, M.: Nonparametric blind super-resolution. In: ICCV. pp. 945–952 (2013)
34. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Commun. ACM **65**(1), 99–106 (2021)
35. Niemeyer, M., Mescheder, L., Oechsle, M., Geiger, A.: Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In: CVPR. pp. 3504–3515 (2020)
36. Oechsle, M., Mescheder, L., Niemeyer, M., Strauss, T., Geiger, A.: Texture fields: Learning texture representations in function space. In: ICCV. pp. 4531–4540 (2019)
37. Park, S.C., Park, M.K., Kang, M.G.: Super-resolution image reconstruction: a technical overview. IEEE Signal Process. Mag. **20**(3), 21–36 (2003)
38. Qu, C., Luo, D., Monari, E., Schuchert, T., Beyerer, J.: Capturing ground truth super-resolution data. In: ICIP. pp. 2812–2816 (2016)
39. Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A.P., Bishop, R., Rueckert, D., Wang, Z.: Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: CVPR. pp. 1874–1883 (2016)
40. Shi, Y., Zhong, H., Yang, Z., Yang, X., Lin, L.: Ddet: Dual-path dynamic enhancement network for real-world image super-resolution. IEEE Signal Process Lett **27**, 481–485 (2020)
41. Sitzmann, V., Zollhöfer, M., Wetzstein, G.: Scene representation networks: Continuous 3d-structure-aware neural scene representations. NeurIPS **32** (2019)

42. Song, G., Sun, Q., Zhang, L., Su, R., Shi, J., He, Y.: Ope-sr: Orthogonal position encoding for designing a parameter-free upsampling module in arbitrary-scale image super-resolution. In: CVPR. pp. 10009–10020 (2023)
43. Tai, Y., Yang, J., Liu, X.: Image super-resolution via deep recursive residual network. In: CVPR. pp. 3147–3155 (2017)
44. Timofte, R., Agustsson, E., Van Gool, L., Yang, M.H., Zhang, L.: Ntire 2017 challenge on single image super-resolution: Methods and results. In: CVPRW. pp. 114–125 (2017)
45. Timofte, R., Gu, S., Wu, J., Van Gool, L.: Ntire 2018 challenge on single image super-resolution: Methods and results. In: CVPRW. pp. 852–863 (2018)
46. Wang, L., Wang, Y., Dong, X., Xu, Q., Yang, J., An, W., Guo, Y.: Unsupervised degradation representation learning for blind super-resolution. In: CVPR. pp. 10581–10590 (2021)
47. Wang, L., Wang, Y., Lin, Z., Yang, J., An, W., Guo, Y.: Learning a single network for scale-arbitrary super-resolution. In: ICCV. pp. 4801–4810 (2021)
48. Wang, X., Xie, L., Dong, C., Shan, Y.: Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In: ICCV. pp. 1905–1914 (2021)
49. Wang, Z., Liu, D., Yang, J., Han, W., Huang, T.: Deep networks for image super-resolution with sparse prior. In: ICCV. pp. 370–378 (2015)
50. Wei, M., Zhang, X.: Super-resolution neural operator. In: CVPR. pp. 18247–18256 (2023)
51. Wei, P., Xie, Z., Lu, H., Zhan, Z., Ye, Q., Zuo, W., Lin, L.: Component divide-and-conquer for real-world image super-resolution. In: ECCV. pp. 101–117 (2020)
52. Xu, X., Wang, Z., Shi, H.: Ultrasr: Spatial encoding is a missing key for implicit image function-based arbitrary-scale super-resolution. arXiv preprint arXiv:2103.12716 (2021)
53. Yang, J., Lin, Z., Cohen, S.: Fast image super-resolution based on in-place example regression. In: CVPR. pp. 1059–1066 (2013)
54. Yang, J., Wright, J., Huang, T.S., Ma, Y.: Image super-resolution via sparse representation. TIP **19**(11), 2861–2873 (2010)
55. Yang, J., Shen, S., Yue, H., Li, K.: Implicit transformer network for screen content image continuous super-resolution. NeurIPS **34**, 13304–13315 (2021)
56. Yang, W., Zhang, X., Tian, Y., Wang, W., Xue, J.H., Liao, Q.: Deep learning for single image super-resolution: A brief review. IEEE Trans. Multimed. **21**(12), 3106–3121 (2019)
57. Zhang, X., Chen, Q., Ng, R., Koltun, V.: Zoom to learn, learn to zoom. In: CVPR. pp. 3762–3770 (2019)
58. Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., Fu, Y.: Image super-resolution using very deep residual channel attention networks. In: ECCV. pp. 286–301 (2018)
59. Zhang, Y., Tian, Y., Kong, Y., Zhong, B., Fu, Y.: Residual dense network for image super-resolution. In: CVPR. pp. 2472–2481 (2018)
60. Zhou, H., Zhu, X., Han, Z., Yin, X.C.: Real-world image super-resolution via spatio-temporal correlation network. In: ICME. pp. 1–6 (2021)