CiaoSR: Continuous Implicit Attention-in-Attention Network for Arbitrary-Scale Image Super-Resolution

Jiezhang Cao[1] Qin Wang[1] Yongqin Xian[1*] Yawei Li[1] Bingbing Ni[2] Zhiming Pi[2]
Kai Zhang[1†] Yulun Zhang[1†] Radu Timofte[1,3] Luc Van Gool[1,4]
[1]ETH Zürich [2]Huawei Inc. [3]University of Wurzburg [4]KU Leuven

https://github.com/caojiezhang/CiaoSR

Abstract

*Learning continuous image representations is recently gaining popularity for image super-resolution (SR) because of its ability to reconstruct high-resolution images with arbitrary scales from low-resolution inputs. Existing methods mostly ensemble nearby features to predict the new pixel at any queried coordinate in the SR image. Such a local ensemble suffers from some limitations: i) it has no learnable parameters and it neglects the similarity of the visual features; ii) it has a limited receptive field and cannot ensemble relevant features in a large field which are important in an image. To address these issues, this paper proposes a **c**ontinuous **i**mplicit **a**ttention-in-attenti**o**n network, called **CiaoSR**. We explicitly design an implicit attention network to learn the ensemble weights for the nearby local features. Furthermore, we embed a scale-aware attention in this implicit attention network to exploit additional non-local information. Extensive experiments on benchmark datasets demonstrate CiaoSR significantly outperforms the existing single image SR methods with the same backbone. In addition, CiaoSR also achieves the state-of-the-art performance on the arbitrary-scale SR task. The effectiveness of the method is also demonstrated on the real-world SR setting. More importantly, CiaoSR can be flexibly integrated into any backbone to improve the SR performance.*

1. Introduction

Single image super-resolution (SISR), which aims to reconstruct a high-resolution (HR) image from a low-resolution (LR) one, has been widely employed in many practical applications [24, 61, 91]. However, deep neural networks (DNN)-based SISR methods are facing some limitations in some real-world scenarios with arbitrary scales. For example, camera users may want to enhance the digital zoom quality by super-resolving a photo or a video to continuous arbitrary scales. Most existing DNN-based SISR methods [40, 44, 86] need to train a series of models for all different scales separately. However, it can be impractical to store all these models on the device due to limited storage and computing power. Alternatively, arbitrary-scale image SR methods [13, 28, 39] aim to train a single network for all scales in a continuous manner.

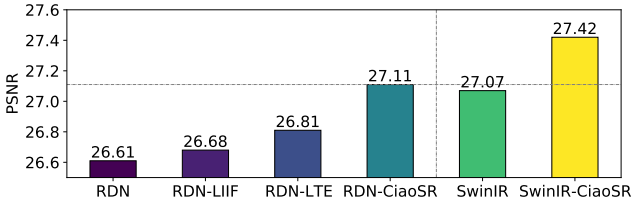


Figure 1. Comparison of different backbones and implicit models. Our proposed implicit neural network on RDN [88] has better performance than SwinIR [40]. Check Section 5 for details.

Most existing SISR methods [40, 44, 86] consist of a DNN and an upsampling module (*e.g.*, pixel shuffling [60]) at a discrete scale. While substantial progresses have been made in the DNN backbones for SR, there is little attempt to study the upsampling module. A natural question to ask is: *Does the pixel shuffling hinder the potential of SR models?* One limitation of the pixel shuffling module is that it cannot synthesize SR images at large unseen and continuous scales. To tackle this, one can treat synthesizing different-scale SR images as a multi-task learning problem, and train a specific upsampling module for each scale [44]. However, these tasks are dependent and highly inter-related. Neglecting the correlation of different-scale SR tasks may lead to discrete representations and limited performance. Under a certain capacity of a network, training a model on multi-tasks may sacrifice the performance or have the comparable performance on each task. These above disadvantages limit its applicability and flexibility in the real-world scenarios.

To address these, most existing arbitrary-scale SR methods [13, 28, 39] replace the upsampling method with an implicit neural function and boost the performance. These methods predict an RGB value at the query point in an image

*Currently with Google. This work was done at ETH Zürich.
†Corresponding Authors: Kai Zhang, cskaizhang@gmail.com; Yulun Zhang, yulun100@gmail.com

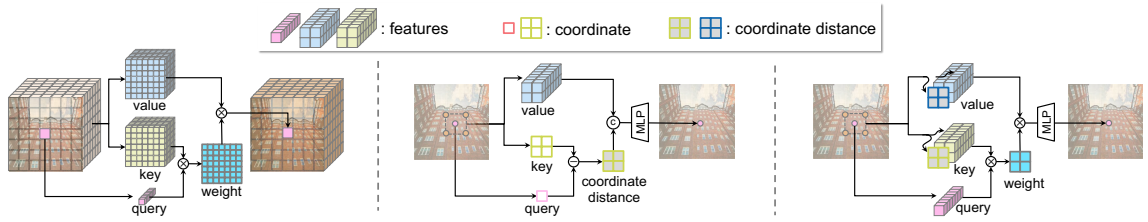


Figure 2. Comparisons of different attention mechanisms. (a) Self-attention can predict pixel features on the grid, but it cannot be directly used in arbitrary-scale SR without considering coordinates. (b) Most existing methods can be treated as coordinate-based implicit attention since they calculate the distance between a key and query coordinate, and then use a function g to aggregate with the value features. However, these methods ignore the distance between the features. (c) Our implicit attention not only considers the coordinate distance, but also the distance among features with visual information.

by ensembling features within a local region. However, the local ensemble methods have limitations in the ensemble weights and insufficient information (*e.g.*, non-local information). The ensemble weights are often calculated by the area of the rectangle between the query point and each nearest point, which is equivalent to the bilinear interpolation. Thus, those methods cannot adaptively ensemble local features since there is no trainable parameter. These weights are only related to the coordinates of the local features, but independent of the local features. Ignoring both the coordinates and the local features lose visual information and result in blurry artifacts. *It is important and necessary to design a new implicit network to predict the weights and exploit more information in the local ensemble.*

In this paper, we propose a novel implicit attention model to enhance arbitrary-scale SR performance. Specifically, we use our attention to predict the ensemble weights by considering both the similarity and coordinate distance of local features, as shown in Figure 2. Based on such learnable weights, the implicit model can adaptively aggregate local features according to different inputs. To enrich more information, we introduce an attention in our implicit attention, which helps discover more features in a larger receptive field.

Our contributions are summarized as follows:

- We propose a novel continuous implicit attention-in-attention network for arbitrary-scale image SR, called CiaoSR. Different from most existing local ensemble methods, our method explicitly learns the ensemble weights and exploits scale-aware non-local information.
- Our CiaoSR can be flexibly integrated into any backbone, allowing the network to super-resolve an image at arbitrary scales and improve the SR performance in Figure 1.
- Extensive experiments demonstrate CiaoSR achieves the state-of-the-art performance in both SISR and arbitrary-scale SR tasks. Besides, our CiaoSR has good generalization on both in-scale and out-of-scale distributions. Last, we extend our method to real-world SR settings to synthesize arbitrary-scale images.

2. Related Work

Single image super-resolution (SISR). SISR aims to synthesize high-resolution (HR) images from low-resolution (LR) images. Compared with DNN-based SR methods [6–8, 23, 27, 50, 66, 67], methods in recent years build on deep convolutional neural network (CNN) to improve the performance, such as SRCNN [19], SRResNet [38], EDSR [44], RDN [88] and RCAN [86]. To further improve SR performance, some methods design CNN with residual block [9, 34, 81], dense block [75, 88, 89] and others [14, 16, 18, 20, 21, 25, 31, 32, 35, 37, 41–43, 59, 65, 70–72, 77, 83, 84, 87, 90]. In addition, some SR methods are built based on attention mechanism [68], such as channel attention [17, 56, 86], self-attention (IPT [11] and SwinIR [40], HAT [12]), non-local attention [45, 48]. However, most methods focus on one specific scale, which limits the applicability and flexibility in arbitrary-scale.

Arbitrary-scale super-resolution. To tackle this problem, very recently, the more practical setup of arbitrary-scale SR is considered, which aims to super-resolute images with arbitrary scales by a single model. MetaSR [28] makes the first attempt to propose an arbitrary-scale meta-upscale module. To improve the performance, many arbitrary-scale methods [64, 73] are proposed. With the help of implicit neural representation [2, 10, 15, 22, 33, 51, 52, 55, 57, 58, 62, 63], LIIF [13] predicts the RGB value at an arbitrary query coordinate by taking an image coordinate and features of backbone around the coordinate. The features are extracted by single image super-resolution methods, *e.g.*, EDSR [44], RDN [88] and SwinIR [40]. To improve the performance, existing methods [39, 78] propose to integrate more features in SR models. For example, LTE [39] proposes a local texture estimator by characterizing image textures in the Fourier space. UltraSR [78] integrates spatial coordinates and periodic encoding in the implicit network. These methods use the bilinear interpolation to ensemble nearby features. However, such an ensemble way has no learnable parameters. Recently, ITSRN [79] learns the weights by taking the coordinate distance and scale token into a mapping. Most methods learn the ensemble without the feature similarity.

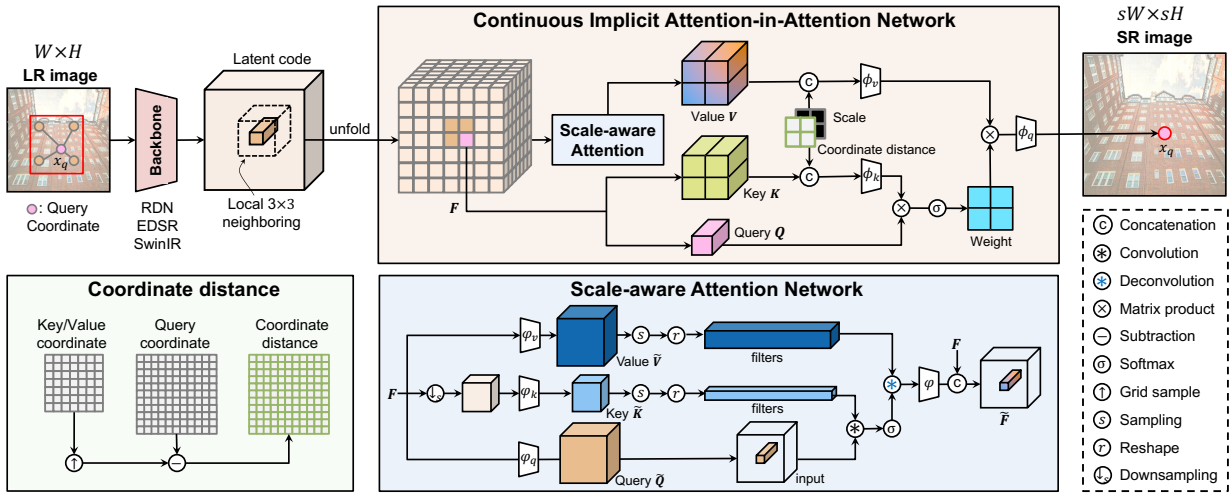


Figure 3. The architecture of our continuous implicit attention-in-attention network. Given an LR image, the encoder extracts features as latent codes. For a query point, we have a query feature and key features close to the query point, and the scale-aware non-local attention module extracts non-local features as value. Last, we use the triple of query, key and value to predict the RGB value at the query point.

3. Preliminary and Motivation

Let $\boldsymbol{I}$ be a continuous image, and $\boldsymbol{x}$ be a 2D coordinate of a pixel in the image $\boldsymbol{I}$. Formally, given a 2D coordinate $\boldsymbol{x}$ in the continuous image domain and latent code $\boldsymbol{Z}$ extracted by deep neural networks, the RGB value can be predicted by an implicit image function which can be defined as follows,

$$\boldsymbol{I}(\boldsymbol{x}) = f(\boldsymbol{Z}, \boldsymbol{x}), \tag{1}$$

where the implicit image function can be parameterized by a multilayer perceptron (MLP). Note that this implicit image function is shared by all images. For recent SISR methods [40, 86], the implicit image function f can be implemented as a PixelShuffle [60] with convolutions with a specific scale. However, these methods are independent on the coordinates, leading to an issue that they only adapt to the specific scale and are inflexible to synthesize arbitrary-scale images.

To predict the RGB value $\boldsymbol{I}_q$ at an arbitrary query coordinate $\boldsymbol{x}_q$, most existing methods [13, 39, 79] propose to compute the RGB value at coordinate $\boldsymbol{x}_q$ by directly ensembling its neighborhood information,

$$\boldsymbol{I}_q := \boldsymbol{I}(\boldsymbol{x}_q) = \sum\nolimits_{(i,j)\in\mathcal{I}} w_{i,j} \cdot f(\boldsymbol{Z}^*_{i,j}, \boldsymbol{x}_q - \boldsymbol{x}^*_{i,j}), \tag{2}$$

where $\mathcal{I}$ is the local region centered at the query coordinate $\boldsymbol{x}_q$, *e.g.*, $\mathcal{I}$ can be top-left, top-right, bottom-left, bottom-right coordinates, and w_i is the weight of the neighboring pixel $\boldsymbol{x}^*_{i,j}$, which is calculated *w.r.t.* the area of the rectangle between $\boldsymbol{x}_q$ and $\boldsymbol{x}^*_{i,j}$, as shown in Figure 4. However, the performance improvement is limited because the ensemble weight $w_{i,j}$ is purely based on the coordinates. The visual similarities are completely ignored in the weight calculation. Besides, these methods only consider the nearest latent codes, leading to a limited receptive field. In this paper, our goal is to *learn* the weights adaptively by leveraging both visual information and coordinate information.

4. Proposed Method

The above local ensemble Eqn. (2) has a similar form to the attention mechanism. Specifically, the weights w_i in Eqn. (2) can be modeled as an attention map and the latter term f can be the value in the attention mechanism. Such an attention map is related to the similarities of the latent code and coordinate, and it can be calculated using both query and key which can integrate the latent code and coordinate information. In this sense, attention models can be used to mitigate the drawbacks of previous methods (*i.e.*, calculated purely based on coordinates without considering any visual information) by learning the soft weights from both visual and coordinate information. However, the use of attention in the implicit function is non-trivial because standard self-attention [68] and neighborhood attention [26] mechanisms are based on visual features, and not conditioned on the coordinate information. To exploit the continuous representation learning from both the coordinate information and visual features, we propose a new attention for arbitrary-scale SR.

Continuous implicit attention-in-attention. The architecture of the proposed continuous implicit attention-in-attention for arbitrary-scale SR (called **CiaoSR**) is shown in Figure 3. It is called attention-in-attention because the value contains an additional embedded attention that additionally captures non-local information of repetitive patterns. Specifically, given an LR image, our attention is provided with both the visual features and coordinate information to *learn* good ensemble weights for the local neighboring latent codes for any query at any arbitrary-scale s.

Given a query coordinate $\boldsymbol{x}_q$ in the upsampled image, the known LR pixels in the local region centered at $\boldsymbol{x}_q$ are considered for the local ensemble. We define $\boldsymbol{x}_k$ and $\boldsymbol{x}_v$ as the key and value coordinates of the local neighbor. Let $\boldsymbol{Q}, \boldsymbol{K}$

and $\boldsymbol{V}$ be the latent code of query, key and value on the corresponding coordinates, respectively. Given pairs of coordinates and latent codes $(\boldsymbol{Q}, \boldsymbol{x}_q)$, $(\boldsymbol{K}, \boldsymbol{x}_k)$ and $(\boldsymbol{V}, \boldsymbol{x}_v)$, we define a new implicit attention function (called i-Attention) to predict RGB values at the given query $\boldsymbol{x}_q$,

$$\boldsymbol{I}_q = \text{i-Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}; \boldsymbol{x}_q, \boldsymbol{x}_k, \boldsymbol{x}_v), \tag{3}$$

where the value feature $\boldsymbol{V}$ can be embedded with another attention to aggregate more non-local information.

As shown in Figure 2, our implicit attention is different from existing attention mechanisms by considering the additional coordinate information. In particular, the self-attention mechanism [68] without considering coordinates cannot predict pixel features that are not on the grid. The coordinate-based implicit attention calculates the distance between a key $\boldsymbol{x}_k$ and query coordinate $\boldsymbol{x}_q$, which ignores the visual information and similarity of the features.

Implicit attention to learn weights for the local ensemble. Unlike previous methods [13, 39] which directly use the normalized area as the ensemble weights, we propose to first use an attention to learn better weights for the ensemble of the local region. Our attention takes both the coordinate distance $\boldsymbol{r}$, the local features $\boldsymbol{F}_l$, as well as non-local features $\boldsymbol{G}$ into account for the learning of the weight. Formally, the attention can be written as

$$\boldsymbol{I}_q = \phi_q\Big(\sum\nolimits_{(i,j)\in\mathcal{I}} \underbrace{\sigma(\boldsymbol{Q}^\top \boldsymbol{K}_{i,j})}_{\text{ensemble weight}} \boldsymbol{V}_{i,j}\Big), \tag{4}$$

where ϕ_q is the query network, σ is the Softmax function, and $\mathcal{I}$ is the local region centered at $\boldsymbol{x}_q$. The query, key, and value for this local ensemble attention are defined as

$$\begin{cases} \boldsymbol{Q} = \boldsymbol{F}^*, \\ \boldsymbol{K}_{i,j} = \phi_k([\boldsymbol{F}_{i,j}, (\boldsymbol{r}_k)_{i,j}, \boldsymbol{s}]), \\ \boldsymbol{V}_{i,j} = \phi_v([\boldsymbol{F}_{i,j}, \tilde{\boldsymbol{F}}_{i,j}], (\boldsymbol{r}_v)_{i,j}, \boldsymbol{s}]), \end{cases} \tag{5}$$

where ϕ_k and ϕ_v are the key and value networks, respectively, $\boldsymbol{F}^*$ are the features of nearest neighbors on the grid of the LR image *w.r.t.* the query coordinate $\boldsymbol{x}_q$, $\boldsymbol{F}_{i,j}$ is a local feature calculated by using the unfolding operator to the latent codes at the (i,j)-th position, which is equivalent to concatenating the latent codes in neighboring 3×3 region. $\tilde{\boldsymbol{F}}_{i,j}$ are non-local features, which will be introduced below in detail. $\boldsymbol{s} = [s_h; s_w]$ is a two-dimensional scale, and $\boldsymbol{r}_k$ and $\boldsymbol{r}_v$ are the relative distance between the coordinates in the local region, *i.e.*,

$$\boldsymbol{r}_k = \boldsymbol{x}_q - (\boldsymbol{x}_k)_{i,j}, \quad \boldsymbol{r}_v = \boldsymbol{x}_q - (\boldsymbol{x}_v)_{i,j}. \tag{6}$$

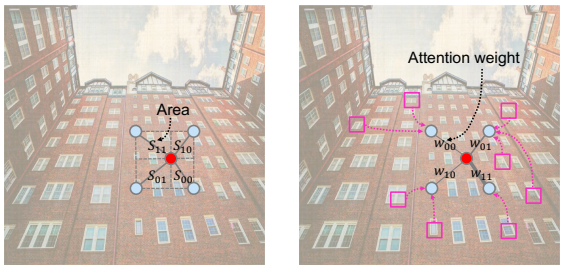
Embedded scale-aware attention for non-local features. Given a query coordinate, the local ensemble attention only considers the keys and values in the small local region, resulting in a limited receptive field. However, there may exist similar textures with different scales within a single image in other positions. For example, repetitive patterns in different scales (*e.g.*, facades, windows, etc. in a building, as shown in Figure 4) can exist in different locations within one image. To exploit the useful non-local information, we propose an embedded scale-aware non-local attention module, which aggregates the information from all coordinates on the grid of the LR image, motivated by [49]. Specifically, we obtain query $\tilde{\boldsymbol{Q}}$ and value features $\tilde{\boldsymbol{V}}$ from the latent code $\boldsymbol{F}$ of the LR image, and downsample $\boldsymbol{F}$ to a scale s' as a key feature $\tilde{\boldsymbol{K}}$. We first calculate the pixel-wise similarity between the query located at (i,j) and the key on other coordinates, and then aggregate the weighted information from all of them. Lastly, we use a convolutional layer φ for downsampling with the scale s' normalized similarity matrix to aggregate the features. Formally, the non-local feature $\tilde{\boldsymbol{F}}$ located at (i,j) can be calculated by

$$\tilde{\boldsymbol{F}}_{i,j} = \varphi\left(\sum_{u,v} \frac{\exp\left(\tilde{\boldsymbol{Q}}_{i,j}^\top \tilde{\boldsymbol{K}}_{u,v}\right)}{\sum_{u',v'} \exp\left(\tilde{\boldsymbol{Q}}_{i,j}^\top \tilde{\boldsymbol{K}}_{u',v'}\right)} \tilde{\boldsymbol{V}}_{s'u,s'v}^{s'p\times s'p}\right), \tag{7}$$

where $\tilde{\boldsymbol{V}}_{s'u,s'v}^{s'p\times s'p}$ is the value feature patch of the size $s'p\times s'p$ located at $(s'u, s'v)$, $\tilde{\boldsymbol{Q}}$, $\tilde{\boldsymbol{K}}$ and $\tilde{\boldsymbol{V}}$ are the query, key and value's non-local features, which can be calculated by

$$\begin{cases} \tilde{\boldsymbol{Q}} = \varphi_q(\boldsymbol{F}), \\ \tilde{\boldsymbol{K}} = \varphi_k(\boldsymbol{F} \downarrow_{s'}), \\ \tilde{\boldsymbol{V}} = \varphi_v(\boldsymbol{F}), \end{cases} \tag{8}$$

where $\downarrow_{s'}$ is the downsampling operation with a given scale s' which can be $\{2,3,4\}$, φ_q, φ_k and φ_v are query, key and value networks. Different from [49], our method can integrate multi-scale features. In the experiment, we can search for relevant features in a small window (*e.g.*, 256×256), instead of the entire image. In this way, our method can guarantee a large receptive field, and reduce computational cost.



(a) Bilinear-based ensemble (b) Attention-based ensemble

Figure 4. Comparison of different ensemble methods. (a) Most existing methods calculate the weights related to the area of the rectangle between the query coordinate and the nearest point. Such weights are equivalent to the weight of the bilinear interpolation. (b) Our attention-based ensemble calculates the attention map as the weights for feature aggregation. Besides, the ensemble method has a large receptive field for exploiting more information.

Table 1. Quantitative comparison with state-of-the-art methods for **arbitrary-scale SR** on the DIV2K validation set (PSNR (dB)). **Bold** indicates the best performance. † indicates the implementation of [39].

Backbones	Methods	In-scale			Out-of-scale				
		×2	×3	×4	×6	×12	×18	×24	×30
-	Bicubic	31.01	28.22	26.66	24.82	22.27	21.00	20.19	19.59
EDSR [44]	EDSR-baseline [44]	34.55	30.90	28.94	-	-	-	-	-
	EDSR-baseline-MetaSR [28]	34.64	30.93	28.92	26.61	23.55	22.03	21.06	20.37
	EDSR-baseline-LIIF [13]	34.67	30.96	29.00	26.75	23.71	22.17	21.18	20.48
	EDSR-baseline-ITSRN† [79]	34.71	30.95	29.03	26.77	23.71	22.17	21.18	20.49
	EDSR-baseline-LTE [39]	34.72	31.02	29.04	26.81	23.78	22.23	21.24	20.53
	EDSR-baseline-CiaoSR (ours)	**34.91**	**31.15**	**29.23**	**26.95**	**23.88**	**22.32**	**21.32**	**20.59**
RDN [88]	RDN-baseline [88]	34.94	31.22	29.19	-	-	-	-	-
	RDN-MetaSR [28]	35.00	31.27	29.25	26.88	23.73	22.18	21.17	20.47
	RDN-LIIF [13]	34.99	31.26	29.27	26.99	23.89	22.34	21.31	20.59
	RDN-ITSRN† [79]	35.09	31.36	29.38	27.06	23.93	22.36	21.32	20.61
	RDN-LTE [39]	35.04	31.32	29.33	27.04	23.95	22.40	21.36	20.64
	RDN-CiaoSR (ours)	**35.15**	**31.42**	**29.45**	**27.16**	**24.06**	**22.48**	**21.43**	**20.70**
SwinIR [40]	SwinIR-baseline [40]	34.94	31.22	29.19	-	-	-	-	-
	SwinIR-MetaSR† [28]	35.15	31.40	29.33	26.94	23.80	22.26	21.26	20.54
	SwinIR-LIIF† [13]	35.17	31.46	29.46	27.15	24.02	22.43	21.40	20.67
	SwinIR-ITSRN† [79]	35.19	31.42	29.48	27.13	23.83	22.31	21.31	20.55
	SwinIR-LTE [39]	35.24	31.50	29.51	27.20	24.09	22.50	21.47	20.73
	SwinIR-CiaoSR (ours)	**35.29**	**31.55**	**29.59**	**27.28**	**24.15**	**22.54**	**21.51**	**20.74**

5. Experiments

Datasets. For the arbitrary-scale SR task, we follow [13, 39] and use DIV2K [1] as the training set, which consists of 800 images in 2K resolution. In testing, we evaluate the models on the DIV2K validation set and a wide range of standard benchmark datasets, including Set5 [3], Set14 [80], B100 [46], Urban100 [29], and Manga109 [47].

Evaluation metrics. We use PSNR to evaluate the quality of the synthesized SR images. Following [13, 39], the PSNR value is calculated on all RGBs channels for the DIV2K validation set, and additionally also calculated on the Y channel (*i.e.*, luminance) of the transformed YCbCr space for other benchmark test sets. In addition, other evaluation metrics (such as SSIM [76] and LPIPS [85]) are provided to evaluate the image quality.

Implementation details. We follow the same experimental setup as prior works [13, 39]. To synthesize paired training data, we first crop the $48s \times 48s$ patches as ground-truth (GT) images, and use the Bicubic downsampling in PyTorch to have LR images with the patch size of 48×48, where s is a scale factor sampled in the uniform distribution $\mathcal{U}(1, 4)$. We sample 48^2 pixels from both GT images and the corresponding coordinates. We use the existing SR models (*e.g.*, EDSR [44], RDN [88] and SwinIR [40]) as backbones by removing their upsampling modules. The detailed architecture of our implicit network is provided in the supplementary material. For the training, we use Adam [36] as the optimizer, and use L1 loss to train all models for 1000 epochs with the batch size of 16 per GPU. We set the learning rate as $1e-4$ at the beginning and decay the learning rate by a factor of 0.5 every 200 epochs. These experimental settings are identical to [13, 28, 39].

5.1. Comparisons with State-of-the-Art

Quantitative results. In Tables 1-3, CiaoSR achieves the best performance with the highest PSNR across all datasets with all backbones concerned on both in-scale and out-of-scale distributions. For the in-scales, in particular, our model has a remarkable PSNR gain of 0.3dB on Urban100 ($\times 4$), compared with the previous SOTA method [39] under the same RDN backbone. It is worth noting that CiaoSR with the RDN backbone [88] can surpass the performance of the better backbone SwinIR [40] (with its original upsampling module). With the help of the continuous-scale training, a single SR model trained with our method has better generalization performance than vanilla SR models (including EDSR [44], RDN [88] and SwinIR [40] trained on specific scale. For the out-of-scale experiments, our model also achieves the best generalization performance on unseen scales. The vanilla EDSR-baseline [44], RDN [88] and SwinIR [40] cannot be applied on the out-of-scale experiments because their decoder can only predict the scale that it was trained on. We observe that our method yields the largest improvement on Urban100 over existing arbitrary-scale methods (*e.g.*, LIIF [13] and LTE [39]).

Qualitative results. In Figure 5, we provide qualitative comparisons with other arbitrary-scale SR methods. Our model is able to synthesize the SR images with sharper textures than other methods. Taking the second line as an example, the textures in the LR image are degraded, but CiaoSR is still able to restore the textures of the building. In contrast, other methods can only restore part of the textures possibly due to the limitations of the local ensemble and insufficient features. More visual results on more test sets are put in the supplementary materials.

Table 2. Quantitative comparison with state-of-the-art methods for **in-scale SR** on benchmark datasets (PSNR (dB)). **Bold** indicates the best performance. † indicates our implementation.

Methods	Set5 [3]			Set14 [80]			B100 [46]			Urban100 [29]			Manga109 [47]		
	×2	×3	×4	×2	×3	×4	×2	×3	×4	×2	×3	×4	×2	×3	×4
RDN [88]	38.24	34.71	32.47	34.01	30.57	28.81	32.34	29.26	27.72	32.89	28.80	26.61	39.18	34.13	31.00
RDN-MetaSR [28]	38.22	34.63	32.38	33.98	30.54	28.78	32.33	29.26	27.71	32.92	28.82	26.55	-	-	-
RDN-LIIF [13]	38.17	34.68	32.50	33.97	30.53	28.80	32.32	29.26	27.74	32.87	28.82	26.68	39.26	34.21	31.20
RDN-ITSRN† [79]	38.23	34.76	32.55	34.19	30.59	28.88	32.38	29.32	27.79	33.07	28.96	26.77	39.34	34.39	31.37
RDN-LTE [39]	38.23	34.72	32.61	34.09	30.58	28.88	32.36	29.30	27.77	33.04	28.97	26.81	39.28	34.32	31.30
RDN-CiaoSR (ours)	**38.29**	**34.85**	**32.66**	**34.22**	**30.65**	**28.93**	**32.41**	**29.34**	**27.83**	**33.30**	**29.17**	**27.11**	**39.51**	**34.57**	**31.57**
SwinIR [40]	38.35	34.89	32.72	34.14	30.77	28.94	32.44	29.37	27.83	33.40	29.29	27.07	39.60	34.74	31.67
SwinIR-MetaSR [28]	38.26	34.77	32.47	34.14	30.66	28.85	32.39	29.31	27.75	33.29	29.12	26.76	39.46	34.62	31.37
SwinIR-LIIF [13]	38.28	34.87	32.73	34.14	30.75	28.98	32.39	29.34	27.84	33.36	29.33	27.15	39.57	34.68	31.71
SwinIR-ITSRN† [79]	38.22	34.75	32.63	34.26	30.75	28.97	32.42	29.38	27.85	33.46	29.34	27.12	39.60	34.75	31.74
SwinIR-LTE [39]	38.33	34.89	32.81	34.25	30.80	29.06	32.44	29.39	27.86	33.50	29.41	27.24	39.63	34.79	31.79
SwinIR-CiaoSR (ours)	**38.38**	**34.91**	**32.84**	**34.33**	**30.82**	**29.08**	**32.47**	**29.42**	**27.90**	**33.65**	**29.52**	**27.42**	**39.67**	**34.84**	**31.91**

Table 3. Quantitative comparison with state-of-the-art methods for **out-of-scale SR** on benchmark datasets (PSNR (dB)). **Bold** indicates the best performance. † indicates our implementation.

Methods	Set5 [3]			Set14 [80]			B100 [46]			Urban100 [29]			Manga109 [47]		
	×6	×8	×12	×6	×8	×12	×6	×8	×12	×6	×8	×12	×6	×8	×12
RDN-MetaSR [28]	29.04	29.96	-	26.51	24.97	-	25.90	24.83	-	23.99	22.59	-	-	-	-
RDN-LIIF [13]	29.15	27.14	24.86	26.64	25.15	23.24	25.98	24.91	23.57	24.20	22.79	21.15	27.33	25.04	22.36
RDN-ITSRN† [79]	29.32	27.25	24.86	26.68	25.17	23.28	26.01	24.93	23.58	24.23	22.81	21.16	27.45	25.04	23.35
RDN-LTE [39]	29.32	27.26	24.79	26.71	25.16	23.31	26.01	24.95	23.60	24.28	22.88	21.22	27.49	25.12	22.43
RDN-CiaoSR (ours)	**29.46**	**27.36**	**24.92**	**26.79**	**25.28**	**23.37**	**26.07**	**25.00**	**23.64**	**24.58**	**23.13**	**21.42**	**27.70**	**25.40**	**22.63**
SwinIR-MetaSR [28]	29.09	27.02	24.82	26.58	25.09	23.33	25.94	24.87	23.59	24.16	22.75	21.31	27.29	24.96	22.35
SwinIR-LIIF [13]	29.46	27.36	-	26.82	25.34	-	26.07	25.01	-	24.59	23.14	-	27.69	25.28	-
SwinIR-ITSRN† [79]	29.31	27.24	24.79	26.71	25.32	23.30	26.05	24.96	23.57	24.50	23.06	21.34	27.72	25.23	22.47
SwinIR-LTE [39]	29.50	27.35	-	26.86	**25.42**	-	26.09	25.03	-	24.62	23.17	-	27.83	25.42	-
SwinIR-CiaoSR (ours)	**29.62**	**27.45**	**24.96**	**26.88**	**25.42**	**23.38**	**26.13**	**25.07**	**23.68**	**24.84**	**23.34**	**21.60**	**28.01**	**25.61**	**22.79**

5.2. Ablation Study

In this section, we conduct ablation studies to investigate effect of each component in our architecture. Our CiaoSR consists of the attention-in-attention network and a scale-aware non-local attention module. Table 4 shows the contributions of each component on performance.

Attention-in-attention. The attention-in-attention network is the main branch of our architecture. On the one hand, the network learns to ensemble weights with both the coordinates and the features. On the other hand, the network is able to aggregate the local and non-local information to improve the SR performance. To investigate how the attention mechanism affects the performance, we remove this module and replace it with an MLP to predict the weights of the ensemble. The coordinate information is still fed into the MLP to enable arbitrary scale prediction. As shown in Table 4, the method suffers a significant performance drop without the attention mechanism.

Non-local attention. The embedded non-local attention is proposed to capture the scale-aware information from other locations of the images to improve the performance. We investigate the importance of the non-local attention in an implicit model by comparing the results with and without this module. As shown in Table 4, by introducing the non-local attention, a further performance gain is achieved. This is because our module can exploit texture information from other scales in a large perceptual field.

Training with different types of scales. We evaluate the effectiveness of our implicit model when trained with discrete (including single scale ($\{2\}/\{3\}/\{4\}$) and multiple scales $\{2, 3, 4\}$) and continuous scales $[1, 4]$, and show the results in Table 5. For the discrete type of scales, we observe that training our implicit model with a specific scale can also achieve significant improvement for that specific scale, compared with the original upsampling module in RDN [88]. Taking $\times 2$ scale as an example, our method has an improvement of 0.24dB over RDN. However, training with a specific scale has poor generalization performance on unseen in-scale and out-of-scale. A few existing works [34, 44] consider the training with multiple discrete scales and found the performance of the joint learned model is *comparable* with single scale networks. Based on our implicit attention module, we are able to deal with more scales and exploit the correlation of multiple SR tasks at different scales. Unlike previous works [34, 44] which only achieve comparable performance compared with single scale models, CiaoSR achieves a better performance than training with one scale. We found our model has further improvement when trained on continuous scales. The improvement gain is attributed to our continuous representations design and our architecture, which enables efficient learning from cross-scale information.

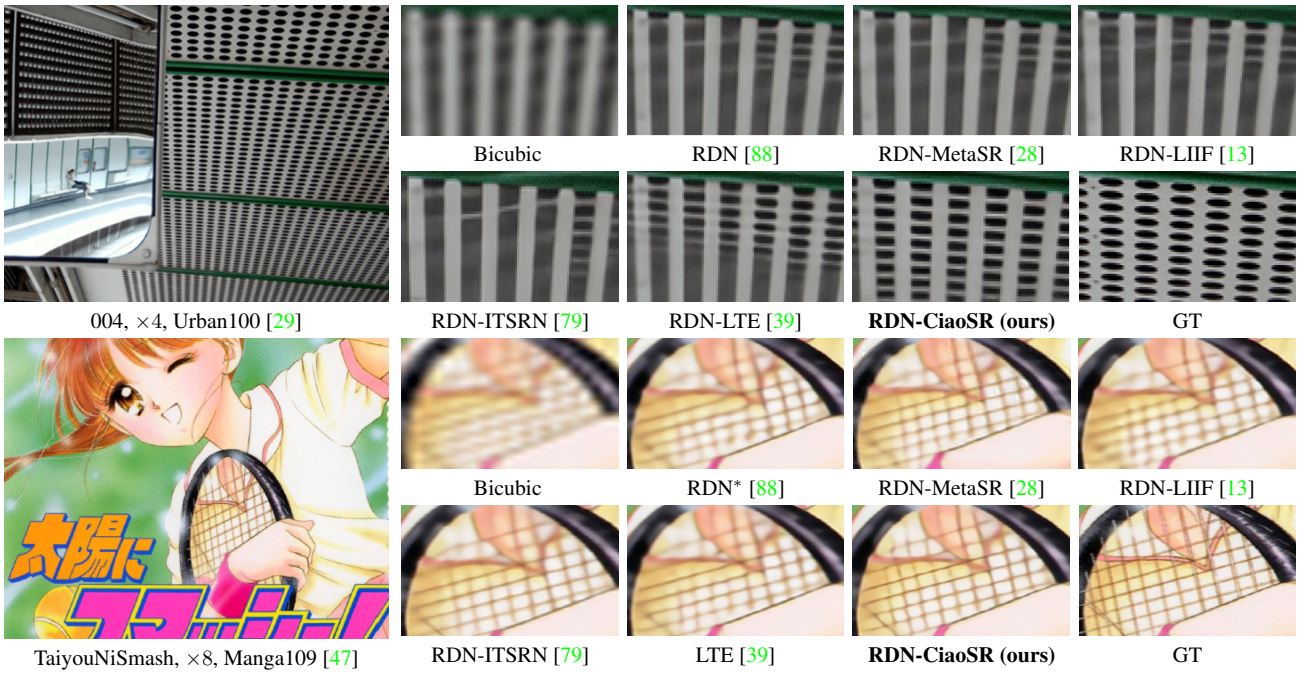


Figure 5. Visual comparison of different methods on benchmarks. "*" means the model first synthesizes twice to ×8 images.

Figure 6. Visual comparison of different methods on the RealSRSet [82] and DPED [30] dataset (×8).

5.3. Further Study

Effect on synthesis steps. Our trained model is able to synthesize SR images with any given arbitrary scale in one step. A natural alternative would be to synthesize a large scale in multiple steps. Taking a large scale (*e.g.*, ×12) as an example, SR images can be generated by two steps (*e.g.*, →×2→×12), or three steps (*e.g.*, →×2→×4→×12). We compare the performance of synthesizing a large scale in one and multiple steps, and the results on the Urban100 and Manga109 datasets are shown in Table 6. We observe that our way of synthesizing SR images in one step has the best performance under the PSNR and SSIM metrics. In contrast, the performance gets worse with more steps because errors can accumulate along the multiple synthesizing steps.

Model size and inference time. To demonstrate the efficiency of our proposed method, we investigate the model size and inference time of different arbitrary-scale SR methods. Results are shown in Table 7. Specifically, we set the input size as 192×192 and calculate the inference time on an NVIDIA Quadro RTX 6000. Our implicit model has much fewer parameters than the previous SOTA methods (*e.g.*, LTE [39]). Our model has the best performance although it requires a little more inference time due to exploiting the non-local features.

More evaluation metrics. To further demonstrate the effectiveness of our proposed method, we apply more evaluation metrics (*e.g.*, SSIM [76] and LPIPS [85]) to evaluate the image quality across different methods on the Urban100 dataset. These evaluation metrics are widely used in the SR community. In general, higher SSIM and lower LPIPS correspond to better performance. As shown in Table 8, our method achieves the highest SSIM and lowest LPIPS, and thus has the best performance on both in-scale and out-of-scale distributions. Compared with other methods, our proposed CiaoSR is able to maintain more structural texture and perceptual information.

Table 4. Ablation study on each component of our networks on Urban100. We use RDN [88] as the backbone.

Attention-in-attention		✗	✓	✓
Scale-aware Attention Network		✗	✗	✓
In-scale	×2	32.87	33.24	**33.30**
	×3	28.82	29.10	**29.17**
	×4	26.69	26.96	**27.11**
Out-of-scale	×6	24.22	24.50	**24.58**
	×8	22.80	22.98	**23.13**

Table 5. Ablation study on training our implicit model with different types of scales on Urban100.

Type	Training scale s	In-scale			Out-of-scale	
		×2	×3	×4	×6	×8
Discrete	$s \in \{2\}$	33.13	27.01	25.60	22.27	22.09
	$s \in \{3\}$	31.39	29.06	25.77	23.44	22.16
	$s \in \{4\}$	31.42	27.87	26.88	24.28	22.85
	$s \in \{2,3,4\}$	33.15	29.14	27.02	24.47	23.03
Continuous	$s \in [1,4]$	**33.30**	**29.17**	**27.11**	**24.58**	**23.13**

Table 6. Comparison of (PSNR and SSIM) for different synthesis steps on Urban100 [29] and Manga109 [47].

Type	Synthesis steps	Urban100 [29]		Manga109 [47]	
		PSNR	SSIM	PSNR	SSIM
Multiple steps	→×2→×4→×12	21.28	0.557	22.46	0.720
	→×2→×12	21.32	0.558	22.53	0.721
One step	→×12	**21.42**	**0.561**	**22.63**	**0.723**

Table 7. Comparisons of model size, inference time and performance gain of different models.

Different models	Meta-SR [28]	LIIF [13]	ITSRN [79]	LTE [39]	CiaoSR
Model size (M)	1.7	1.6	0.7	1.7	1.4
Inference time (ms)	237	171	343	148	528
PSNR (dB)	26.55	26.68	26.77	26.81	27.11
Performance gain (dB)	-0.06	0.07	0.16	0.2	**0.5**

Table 8. Comparison of more evaluation metrics (SSIM↑ [76] and LPIPS↓ [85]) for different methods on Urban100.

RDN-Method	In-scale						Out-of-scale			
	×2		×3		×4		×6		×8	
	SSIM	LPIPS	SSIM	LPIPS	SSIM	LPIPS	SSIM	LPIPS	SSIM	LPIPS
LIIF [13]	0.934	0.104	0.866	0.197	0.804	0.267	0.704	0.360	0.636	0.425
MetaSR [28]	0.935	0.104	0.866	0.195	0.805	0.261	0.698	0.353	0.624	0.423
ITSRN [79]	0.936	0.102	0.868	0.193	0.807	0.259	0.703	0.350	0.633	0.417
LTE [39]	0.936	0.104	0.868	0.195	0.807	0.264	0.707	0.356	0.639	0.421
CiaoSR	**0.938**	**0.100**	**0.871**	**0.188**	**0.814**	**0.255**	**0.718**	**0.346**	**0.651**	**0.408**

Table 9. Quantitative comparisons of different Real-world SR methods on RealSRSet [82].

Methods	Scale	NIQE↓	BRISQUE↓	PIQE↓
RealSR [5]	×8	4.7949	23.3228	29.3976
	×16	3.3946	21.6390	21.4116
BSRGAN [82]	×8	4.7007	37.0638	36.1889
	×16	4.1408	39.6252	37.9176
Real-ESRGAN [74]	×8	4.5280	40.6521	59.8915
	×16	5.6442	53.9364	84.3611
Ours-real	×8	3.5894	30.9219	29.8554
	×16	3.9461	41.8145	58.8244

5.4. Real-World Arbitrary-Scale Image SR

We extend our proposed method to real-world applications with the ultimate goal of synthesizing real images with arbitrary scales. Specifically, we train our model on the DIV2K dataset [1], and test on the RealSRSet [82] and DPED [30] datasets. We apply the practical degradation models (including blur, noise, compression, etc.) of BSRGAN [82] and Real-ESRGAN [74] to synthesize paired training data on DIV2K [1]. We set the HR patch size of training data as 256, and the batch size as 48. We use Adam [36] as the optimizer, and adopt the exponential moving average (EMA). We first train our model with L1 loss for 1000K iterations with the learning rate of 2×10^{-4} and train with L1 loss, perceptual loss and GAN loss together for 400K iterations with the learning rate of 1×10^{-4}. More training details can be found in the supplementary materials.

Since there are no arbitrary-scale SR methods applied to real-world images, we mainly compare with the state-of-the-art real-world image SR models, including RealSR [5], BSRGAN [82] and Real-ESRGAN [74]. Because there are no ground-truth (GT) real images, we apply non-reference image quality assessment (*e.g.*, NIQE [54], BRISQUE [53] and PIQE [69]) to evaluate the image quality. As shown in Table 9, our model does not optimize for all metrics since these metrics cannot well match the actual human perceptual [4]. To address this, we further compare the visual results on the RealSRSet [82] and DPED [30] datasets in Figure 6. Our proposed method is able to synthesize real SR images with realistic textures. In particular, our method is able to synthesize more details of a dog in the first line and more clear digit '5' in the second line, compared with other methods. In contrast, other methods often introduce unnatural artifacts within images, especially for out-of-scale distributions. To further investigate the effectiveness of our model, we provide more visual results of our real model in the supplementary materials.

6. Conclusion

In this paper, we have proposed a novel continuous implicit attention-in-attention network for arbitrary-scale image super-resolution, called CiaoSR. Specifically, our CiaoSR first introduces a scale-aware non-local attention in our attention to exploit more relevant features, and then predicts RGB values by locally ensembling features with the learnable attention weights. More importantly, CiaoSR has good flexibility and applicability since it can be used behind any SR backbone to boost the performance. Extensive experiments on all benchmark datasets demonstrate the superiority of our CiaoSR in both SISR and arbitrary-scale SR tasks. Besides, our CiaoSR has good generalization performance on both in-scale and out-of-scale distributions. Last, we extend our method to the real-world application to synthesize real SR images with arbitrary scale.

Acknowledgements: This work was partly supported by the Huawei Fund, the ETH Zürich General Fund (OK) and the Alexander von Humboldt Foundation.

References

[1] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 126–135, 2017. 5, 8

[2] Matan Atzmon and Yaron Lipman. Sal: Sign agnostic learning of shapes from raw data. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2565–2574, 2020. 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*, pages 135.1–135.10, 2012. 5, 6

[4] Yochai Blau, Roey Mechrez, Radu Timofte, Tomer Michaeli, and Lihi Zelnik-Manor. The 2018 pirm challenge on perceptual image super-resolution. In *European Conference on Computer Vision Workshops*, 2018. 8

[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 *IEEE Conference on International Conference on Computer Vision*, pages 3086–3095, 2019. 8

[6] Jiezhang Cao, Yawei Li, Kai Zhang, and Luc Van Gool. Video super-resolution transformer. *arXiv preprint arXiv:2106.06847*, 2021. 2

[7] Jiezhang Cao, Jingyun Liang, Kai Zhang, Yawei Li, Yulun Zhang, Wenguan Wang, and Luc Van Gool. Reference-based image super-resolution with deformable attention transformer. In *European Conference on Computer Vision*, pages 325–342, 2022. 2

[8] Jiezhang Cao, Jingyun Liang, Kai Zhang, Wenguan Wang, Qin Wang, Yulun Zhang, Hao Tang, and Luc Van Gool. Towards interpretable video super-resolution via alternating optimization. In *European Conference on Computer Vision*, pages 393–411, 2022. 2

[9] Lukas Cavigelli, Pascal Hager, and Luca Benini. Cas-cnn: A deep convolutional neural network for image compression artifact suppression. In *2017 International Joint Conference on Neural Networks*, pages 752–759, 2017. 2

[10] Rohan Chabra, Jan E Lenssen, Eddy Ilg, Tanner Schmidt, Julian Straub, Steven Lovegrove, and Richard Newcombe. Deep local shapes: Learning local sdf priors for detailed 3d reconstruction. In *European Conference on Computer Vision*, pages 608–625, 2020. 2

[11] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 12299–12310, 2021. 2

[12] Xiangyu Chen, Xintao Wang, Jiantao Zhou, and Chao Dong. Activating more pixels in image super-resolution transformer. *arXiv preprint arXiv:2205.04437*, 2022. 2

[13] Yinbo Chen, Sifei Liu, and Xiaolong Wang. Learning continuous image representation with local implicit image function. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8628–8638, 2021. 1, 2, 3, 4, 5, 6, 7, 8

[14] Yunjin Chen and Thomas Pock. Trainable nonlinear reaction diffusion: A flexible framework for fast and effective image restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1256–1272, 2016. 2

[15] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5939–5948, 2019. 2

[16] Wenlong Cheng, Mingbo Zhao, Zhiling Ye, and Shuhang Gu. Mfagan: A compression framework for memory-efficient on-device super-resolution gan. *arXiv preprint arXiv:2107.12679*, 2021. 2

[17] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-order attention network for single image super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 11065–11074, 2019. 2

[18] Xin Deng, Yutong Zhang, Mai Xu, Shuhang Gu, and Yiping Duan. Deep coupled feedback network for joint exposure fusion and image super-resolution. *IEEE Transactions on Image Processing*, 30:3098–3112, 2021. 2

[19] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *European Conference on Computer Vision*, pages 184–199, 2014. 2

[20] Xueyang Fu, Menglu Wang, Xiangyong Cao, Xinghao Ding, and Zheng-Jun Zha. A model-driven deep unfolding method for jpeg artifacts removal. *IEEE Transactions on Neural Networks and Learning Systems*, 2021. 2

[21] Xueyang Fu, Zheng-Jun Zha, Feng Wu, Xinghao Ding, and John Paisley. Jpeg artifacts reduction via deep convolutional sparse coding. In *IEEE International Conference on Computer Vision*, pages 2501–2510, 2019. 2

[22] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. *arXiv preprint arXiv:2002.10099*, 2020. 2

[23] Shuhang Gu, Nong Sang, and Fan Ma. Fast image super resolution via local regression. In *IEEE Conference on International Conference on Pattern Recognition*, pages 3128–3131, 2012. 2

[24] Bahadir K Gunturk, Yucel Altunbasak, and Russell M Mersereau. Super-resolution reconstruction of compressed video using transform-domain statistics. *IEEE Transactions on Image Processing*, 13(1):33–43, 2004. 1

[25] Yong Guo, Jian Chen, Jingdong Wang, Qi Chen, Jiezhang Cao, Zeshuai Deng, Yanwu Xu, and Mingkui Tan. Closed-loop matters: Dual regression networks for single image super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5407–5416, 2020. 2

[26] Ali Hassani, Steven Walton, Jiachen Li, Shen Li, and Humphrey Shi. Neighborhood attention transformer. *arXiv preprint arXiv:2204.07143*, 2022. 3

[27] Kaiming He, Jian Sun, and Xiaoou Tang. Single image haze removal using dark channel prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(12):2341–2353, 2010. 2

[28] Xuecai Hu, Haoyuan Mu, Xiangyu Zhang, Zilei Wang, Tieniu Tan, and Jian Sun. Meta-sr: A magnification-arbitrary network for super-resolution. In *IEEE Conference on Computer*

Vision and Pattern Recognition, pages 1575–1584, 2019. 1, 2, 5, 6, 7, 8

[29] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5197–5206, 2015. 5, 6, 7, 8

[30] Andrey Ignatov, Nikolay Kobyshev, Radu Timofte, Kenneth Vanhoey, and Luc Van Gool. Dslr-quality photos on mobile devices with deep convolutional networks. In *IEEE International Conference on Computer Vision*, pages 3277–3285, 2017. 7, 8

[31] Takashi Isobe, Xu Jia, Shuhang Gu, Songjiang Li, Shengjin Wang, and Qi Tian. Video super-resolution with recurrent structure-detail network. In *European Conference on Computer Vision*, pages 645–660, 2020. 2

[32] Xixi Jia, Sanyang Liu, Xiangchu Feng, and Lei Zhang. Focnet: A fractional optimal control network for image denoising. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6054–6063, 2019. 2

[33] Chiyu Jiang, Avneesh Sud, Ameesh Makadia, Jingwei Huang, Matthias Nießner, Thomas Funkhouser, et al. Local implicit grid representations for 3d scenes. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6001–6010, 2020. 2

[34] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1646–1654, 2016. 2, 6

[35] Yoonsik Kim, Jae Woong Soh, Jaewoo Park, Byeongyong Ahn, Hyun-Seung Lee, Young-Su Moon, and Nam Ik Cho. A pseudo-blind convolutional neural network for the reduction of compression artifacts. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(4):1121–1135, 2019. 2

[36] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. 5, 8

[37] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 624–632, 2017. 2

[38] 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 *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4681–4690, 2017. 2

[39] Jaewon Lee and Kyong Hwan Jin. Local texture estimator for implicit representation function. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1929–1938, 2022. 1, 2, 3, 4, 5, 6, 7, 8

[40] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *IEEE International Conference on Computer Vision Workshops*, pages 1833–1844, 2021. 1, 2, 3, 5, 6

[41] Jingyun Liang, Andreas Lugmayr, Kai Zhang, Martin Danelljan, Luc Van Gool, and Radu Timofte. Hierarchical conditional flow: A unified framework for image super-resolution and image rescaling. In *IEEE Conference on International Conference on Computer Vision*, 2021. 2

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

[43] Jingyun Liang, Kai Zhang, Shuhang Gu, Luc Van Gool, and Radu Timofte. Flow-based kernel prior with application to blind super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 10601–10610, 2021. 2

[44] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 136–144, 2017. 1, 2, 5, 6

[45] Ding Liu, Bihan Wen, Yuchen Fan, Chen Change Loy, and Thomas S Huang. Non-local recurrent network for image restoration. *arXiv preprint arXiv:1806.02919*, 2018. 2

[46] 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 *IEEE Conference on International Conference on Computer Vision*, pages 416–423, 2001. 5, 6

[47] 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*, 76(20):21811–21838, 2017. 5, 6, 7, 8

[48] Yiqun Mei, Yuchen Fan, and Yuqian Zhou. Image super-resolution with non-local sparse attention. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3517–3526, 2021. 2

[49] Yiqun Mei, Yuchen Fan, Yuqian Zhou, Lichao Huang, Thomas S Huang, and Honghui Shi. Image super-resolution with cross-scale non-local attention and exhaustive self-exemplars mining. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5690–5699, 2020. 4

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

[51] Mateusz Michalkiewicz, Jhony K Pontes, Dominic Jack, Mahsa Baktashmotlagh, and Anders Eriksson. Implicit surface representations as layers in neural networks. In *IEEE International Conference on Computer Vision*, pages 4743–4752, 2019. 2

[52] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 2

[53] Anish Mittal, Anush K Moorthy, and Alan C Bovik. Blind/referenceless image spatial quality evaluator. In *Asilomar Conference on Signals, Systems and Computers*, pages 723–727, 2011. 8

[54] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2012. 8

[55] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3504–3515, 2020. 2

[56] 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 *European Conference on Computer Vision*, pages 191–207, 2020. 2

[57] Michael Oechsle, Lars Mescheder, Michael Niemeyer, Thilo Strauss, and Andreas Geiger. Texture fields: Learning texture representations in function space. In *IEEE International Conference on Computer Vision*, pages 4531–4540, 2019. 2

[58] Songyou Peng, Michael Niemeyer, Lars Mescheder, Marc Pollefeys, and Andreas Geiger. Convolutional occupancy networks. In *European Conference on Computer Vision*, pages 523–540, 2020. 2

[59] Yali Peng, Lu Zhang, Shigang Liu, Xiaojun Wu, Yu Zhang, and Xili Wang. Dilated residual networks with symmetric skip connection for image denoising. *Neurocomputing*, 345:67–76, 2019. 2

[60] 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 *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1874–1883, 2016. 1, 3

[61] Wenzhe Shi, Jose Caballero, Christian Ledig, Xiahai Zhuang, Wenjia Bai, Kanwal Bhatia, Antonio M 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*, pages 9–16, 2013. 1

[62] Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. *Advances in Neural Information Processing Systems*, 33:7462–7473, 2020. 2

[63] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[64] Sanghyun Son and Kyoung Mu Lee. Srwarp: Generalized image super-resolution under arbitrary transformation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7782–7791, 2021. 2

[65] Ying Tai, Jian Yang, Xiaoming Liu, and Chunyan Xu. Memnet: A persistent memory network for image restoration. In *IEEE International Conference on Computer Vision*, pages 4539–4547, 2017. 2

[66] Radu Timofte, Vincent De Smet, and Luc Van Gool. Anchored neighborhood regression for fast example-based super-resolution. In *IEEE Conference on International Conference on Computer Vision*, pages 1920–1927, 2013. 2

[67] Radu Timofte, Vincent De Smet, and Luc Van Gool. A+: Adjusted anchored neighborhood regression for fast super-resolution. In *Asian Conference on Computer Vision*, pages 111–126, 2014. 2

[68] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017. 2, 3, 4

[69] N Venkatanath, D Praneeth, Maruthi Chandrasekhar Bh, Sumohana S Channappayya, and Swarup S Medasani. Blind image quality evaluation using perception based features. In *National Conference on Communications*, pages 1–6, 2015. 8

[70] Longguang Wang, Yingqian Wang, Xiaoyu Dong, Qingyu Xu, Jungang Yang, Wei An, and Yulan Guo. Unsupervised degradation representation learning for blind super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 10581–10590, 2021. 2

[71] Longguang Wang, Yingqian Wang, Zhengfa Liang, Zaiping Lin, Jungang Yang, Wei An, and Yulan Guo. Learning parallax attention for stereo image super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 12250–12259, 2019. 2

[72] Longguang Wang, Yingqian Wang, Zaiping Lin, Jungang Yang, Wei An, and Yulan Guo. Learning a single network for scale-arbitrary super-resolution. In *IEEE Conference on International Conference on Computer Vision*, pages 10581–10590, 2021. 2

[73] Longguang Wang, Yingqian Wang, Zaiping Lin, Jungang Yang, Wei An, and Yulan Guo. Learning a single network for scale-arbitrary super-resolution. In *IEEE International Conference on Computer Vision*, pages 4801–4810, 2021. 2

[74] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *International Conference on Computer Vision Workshops*. 8

[75] 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 *European Conference on Computer Vision Workshops*, pages 701–710, 2018. 2

[76] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 5, 7, 8

[77] Yunxuan Wei, Shuhang Gu, Yawei Li, Radu Timofte, Longcun Jin, and Hengjie Song. Unsupervised real-world image super resolution via domain-distance aware training. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 13385–13394, 2021. 2

[78] Xingqian Xu, Zhangyang Wang, and Humphrey Shi. Ultrasr: Spatial encoding is a missing key for implicit image function-based arbitrary-scale super-resolution. *arXiv preprint arXiv:2103.12716*, 2021. 2

[79] Jingyu Yang, Sheng Shen, Huanjing Yue, and Kun Li. Implicit transformer network for screen content image continuous super-resolution. *Advances in Neural Information Processing Systems*, 34:13304–13315, 2021. 2, 3, 5, 6, 7, 8

[80] Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In *International*

Conference on Curves and Surfaces, pages 711–730, 2010. 5, 6

[81] Kai Zhang, Yawei Li, Wangmeng Zuo, Lei Zhang, Luc Van Gool, and Radu Timofte. Plug-and-play image restoration with deep denoiser prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 2

[82] Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *IEEE Conference on International Conference on Computer Vision*, 2021. 7, 8

[83] Kai Zhang, Wangmeng Zuo, and Lei Zhang. Ffdnet: Toward a fast and flexible solution for cnn-based image denoising. *IEEE Transactions on Image Processing*, 27(9):4608–4622, 2018. 2

[84] Kai Zhang, Wangmeng Zuo, and Lei Zhang. Learning a single convolutional super-resolution network for multiple degradations. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3262–3271, 2018. 2

[85] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. 5, 7, 8

[86] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *European Conference on Computer Vision*, pages 286–301, 2018. 1, 2, 3

[87] Yulun Zhang, Kunpeng Li, Kai Li, Bineng Zhong, and Yun Fu. Residual non-local attention networks for image restoration. *arXiv preprint arXiv:1903.10082*, 2019. 2

[88] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2472–2481, 2018. 1, 2, 5, 6, 7, 8

[89] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(7):2480–2495, 2020. 2

[90] Shangchen Zhou, Jiawei Zhang, Wangmeng Zuo, and Chen Change Loy. Cross-scale internal graph neural network for image super-resolution. *arXiv preprint arXiv:2006.16673*, 2020. 2

[91] Wilman WW Zou and Pong C Yuen. Very low resolution face recognition problem. *IEEE Transactions on Image Processing*, 21(1):327–340, 2011. 1