Deep Arbitrary-Scale Image Super-Resolution via Scale-Equivariance Pursuit

Xiaohang Wang[1*] Xuanhong Chen[1*] Bingbing Ni[1†] Hang Wang[2] Zhengyan Tong[1] Yutian Liu[1]
[1]Shanghai Jiao Tong University, Shanghai 200240, China [2]Huawei
{xygz2014010003,chen19910528,nibingbing}@sjtu.edu.cn

Abstract

The ability of scale-equivariance processing blocks plays a central role in arbitrary-scale image super-resolution tasks. Inspired by this crucial observation, this work proposes two novel scale-equivariant modules within a transformer-style framework to enhance arbitrary-scale image super-resolution (ASISR) performance, especially in high upsampling rate image extrapolation. In the feature extraction phase, we design a plug-in module called Adaptive Feature Extractor, which injects explicit scale information in frequency-expanded encoding, thus achieving scale-adaption in representation learning. In the upsampling phase, a learnable Neural Kriging upsampling operator is introduced, which simultaneously encodes both relative distance (i.e., scale-aware) information as well as feature similarity (i.e., with priori learned from training data) in a bilateral manner, providing scale-encoded spatial feature fusion. The above operators are easily plugged into multiple stages of a SR network, and a recent emerging pre-training strategy is also adopted to impulse the model's performance further. Extensive experimental results have demonstrated the outstanding scale-equivariance capability offered by the proposed operators and our learning framework, with much better results than previous SOTA methods at arbitrary scales for SR. Our code is available at https://github.com/neuralchen/EQSR.

1. Introduction

Arbitrary-scale image super-resolution (ASISR), which aims at upsampling the low-resolution (LR) images to high-resolution (HR) counterparts by any proper real-valued magnifications with one single model, has become one of the most interesting topics in low-level computer vision research for its flexibility and practicality. Unfortunately, compared with fixed-scale SISR models [2,3,11,19,20,22], existing methods on ASISR [4,9,18,33] usually offer much lower SR performances (e.g., PSNR), hindering their practical applications. The major causes of the defects of previous ASISR methods are due to lack of scale-equivariance in dealing with scale-varying image features, as explained in detail as follows.

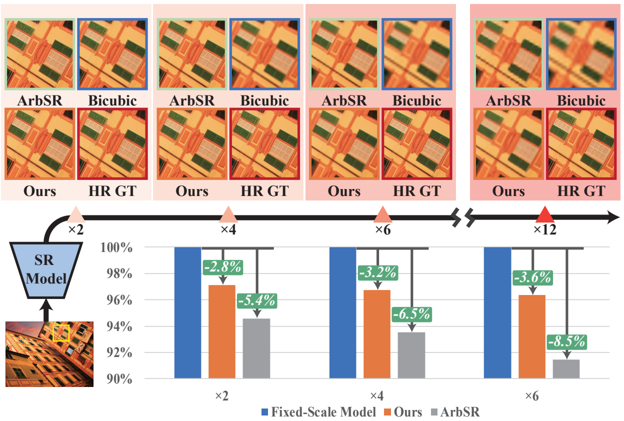

Figure 1. Scale-equivariance of our network. We compare the PSNR degradation rate of our method and ArbSR [33]. Taking the SOTA fixed-scale method HAT [3] as reference, our model presents a more stable degradation as the scale increases, reflecting the equivariance of our method. Please enlarge the pdf for details.

On the one hand, the model's backbone should possess the capability of adaptively processing features according to the sampling scale in order to achieve scale-equivariance in the feature extraction phase. To be concrete, since ASISR models rely on only one single backbone to handle different scales, designing a scale-equivariant feature learning module that extracts, transforms, and pool image information from adaptively adjusted sampling positions (i.e., according to the scaled distance) to obtain scale-equivariant feature maps from the same source image with different scales is essential. As shown in Figure 2, we analyze a series of fixed-scale HAT [3] models and find that the extracted features show apparent divergences from the middle of the network to handle different scale factors, demonstrating that features for different scales should be extracted adaptively.

On the other hand, we also expect the model to possess suitable scale-equivariant properties in the image/feature upsampling stage. Namely, the upsampling module should

[*] Equal Contribution.
[†] Corresponding author: Bingbing Ni.

be designed to perform adaptive interpolation operations according to arbitrary scaling factors. It is worth noting that the ability to handle out-of-distribution scales (i.e., scales that the model has never seen in training) is crucial for ASISR models since the training process is impossible to cover all possible upsampling scales. However, existing methods commonly use 3×3 convolution layers for upsampling, which has been proved to lack scale equivariance by many studies [35–37], leading to insensitivity to the changes of the scale factor. Some implicit-field-based methods, such as LIIF [4], adopt a channel-separated MLP to enhance the scale equivariance; however, additional operations, including feature unfolding and local ensemble, are needed, resulting in a cumbersome upsampler. Alias-Free StyleGAN [12] points out that 1×1 convolution could be regarded as an instance of a continuously E(2)-equivariant model [34] in image generation, but 1×1 receptive field cannot aggregate the crucial local information for SR.

Motivated by the above analysis, this work proposes two novel scale-equivariant modules within a transformer-style framework for enhancing arbitrary-scale image super-resolution performance. In the feature extraction phase, we design a novel module called Adaptive Feature Extractor (AFE), which explicitly injects scale information in the form of frequency-expanded encoding to modulate the weights of subsequent convolution. Combined with the traditional self-attention mechanism, this operator can be plugged into multiple stages of the feature extraction sub-network and achieves a large receptive field as well as good scale-adaption properties in representation learning. When upsampling, instead of monotonically using pixel-independent convolutions (e.g., Alias-Free StyleGAN [12], LIIF [4]), we propose a brand-new paradigm, i.e., Neural Kriging upsampler, which endows vanilla $K \times K$ convolutions with highly competitive equivariance while maintaining excellent spatial aggregation. Specifically, the Neural Kriging upsampler simultaneously encodes geometric information (i.e., relative position) and feature similarity (i.e., prior knowledge learned from training data) in a bilateral manner, providing scale-encoded spatial feature fusion.

Combining the above modules, we construct a model with a certain equivariance named EQSR, which can adaptively handle different sampling rates. We conduct extensive qualitative and quantitative experiments to verify the superiority of our method on ASISR benchmarks. Compared with state-of-the-art methods, average PSNRs of our model have shown significant advantages in both in-distribution and out-of-distribution cases. Under the $\times 2$ and $\times 3$ configurations, we surpass the previous SOTA LTE [18] by 0.33dB (34.83dB v.s. 34.50dB) and 0.35dB (29.76dB v.s. 29.41dB) on the Urban100 dataset. Under the $\times 6$ configuration, we also achieve a large gap of 0.21dB, proving the effectiveness of our scale-equivariant operator.

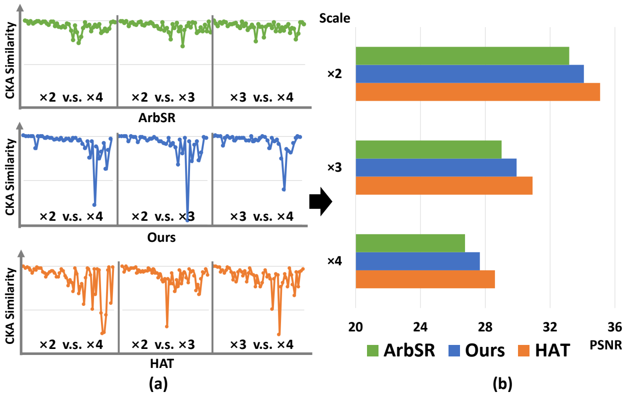

Figure 2. Feature similarity of Different models. We compare the recent SOTA fixed-scale model HAT [3] and arbitrary-scale model ArbSR [33]. (a) shows the CKA similarity of $\times 2/3/4$ features at each layer; (b) compares the performance of these methods on the Urban100 dataset.

2. Related Work

Fixed-scale SR. Since SRCNN [6] proposed the first CNN-based single image super-resolution model, CNN-based deep learning SR methods have outperformed those exemplars or dictionary-based traditional SR methods [30, 32, 38, 39, 42] a large margin. VDSR [13] proposed using deep networks and residual learning for SR model training. EDSR [21] removed batch normalization (BN) layers and used a residual scaling technique to train large SR models. RDN [46] proposed dense feature fusion for image super-resolution. RCAN [45] proposed adding a channel attention mechanism to improve image SR performance. DRCN [14] introduced the first recursive supervision to SR.

Arbitrary-Scale SR. Most of the existing SR techniques train respective models for each specific scale factor (e.g., $\times 2$, $\times 3$, $\times 4$), which limits the deployment on the user side considering the memory and computing resources. In view of this, single model SR for arbitrary-scale factors is convenient and efficient in practical scenarios. EDSR [21] integrates models trained for multiple integer scale factors as a single model MDSR. MetaSR [9] proposed the first single model arbitrary-scale SR method by taking the scale factor as input to predict the weights of the upscale filters. Recently, ArbSR [33] proposed a general plug-in module using conditional convolutions to generate dynamic scale-aware filters. SRWarp [29] proposed a differentiable adaptive warping layer to transform an LR image into any shape deformations in HR representation. However, these methods do not perform well in out-of-distribution cases. LIIF [4] train encoder with implicit neural representation to learn continuous image representation, which can be presented in arbitrary resolution. LTE [18] proposed a dominant-frequency estimator based on LIIF and improve the performance. However, these method extract the same features for all scales, hindering the scale-equivariance of models.

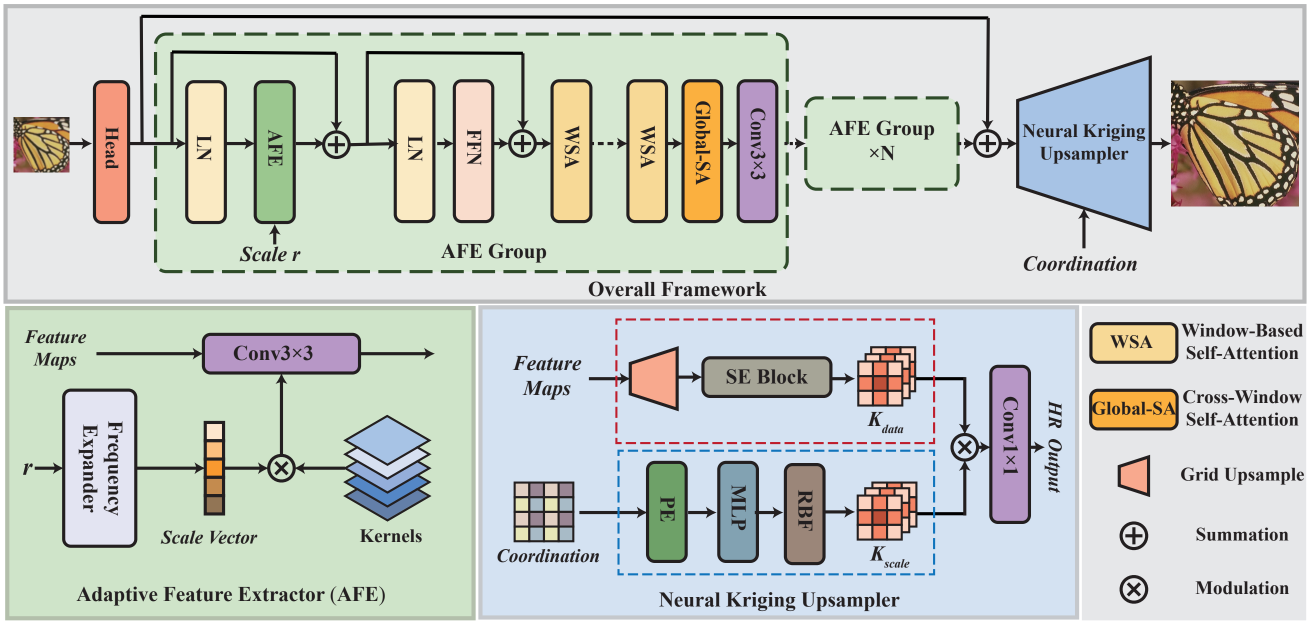

Figure 3. Architectures of proposed modules and networks. The main body of our network contains a series of AFE Groups and a Neural Kriging upsampler. The core of AFE Group is our Adaptive Feature Extractor which extracts the features dynamically according to the scale factor. The Neural Kriging upsampler consists of a scale-insensitive branch for learning prior knowledge and a scale-sensitive branch for perceiving spatial distance information.

3. Methodology

3.1. Problem Formulation

Let $I^{LR} \in \mathbb{R}^{H \times W \times 3}$ denotes a LR image. Arbitrary SR models aim to re-scale it to $I^{SR} \in \mathbb{R}^{rH \times rW \times 3}$ with one single model, where r denotes any proper real-number scale factor. Given a query coordinate x_0, the problem of a typical ASISR method could be formulated as follow:

$$z_o = I^{SR}(x_o) = \Phi(I^{LR}, x_o; r), \tag{1}$$

where Φ denotes the SR model, and z_o denotes the corresponding pixel value prediction.

On the one hand, since image features extracted for the subsequent upsampling stage should match for different scales, a scale-equivariant feature extraction backbone is required to extract features adaptively according to the scales. On the other hand, most existing upsampling algorithms cannot perceive the information of scale transformation, urging us to find a novel sampling theory to enhance the scale-equivariance of the model. To explicitly address the above challenges, we propose a novel framework named Scale-**Eq**uivariant Super-Resolution (EQSR). In this section, we first give an overview of our proposed framework. Then, we describe the working mechanism (in achieving scale-equivariance) and implementation details of our proposed Adaptive Feature Extractor (AFE) and Neural Kriging Upsampler, respectively. Coupling with the proposed above components, EQSR achieves superior performances, especially at out-of-distribution scales.

3.2. Overall Architecture

As shown in Figure 3, our network consists of three parts: the head, the AFE groups, and the Neural Kriging Upsampler. The head is one layer of convolution that extracts shallow features and enriches low-level patterns. The main body of the backbone contains a series of AFE groups to achieve a scale-adaptive feature extraction and gain a large receptive field. The AFE group is composed of an AFE block (i.e., transformer-style block [7]), several window-based self-attention [20] (i.e., WSA for short), a global self-attention [40] (i.e., GSA for short) and a 3 × 3 convolution. AFE block is designed to endow the backbone network with scale adaptability, which will be discussed in Sec. 3.3. In our framework, we employ the naive non-overlap windows-based self-attention with 16×16 window. Global-based self-attention is employed to establish the non-local context interaction, which is highly important in boosting restoration performance. In this paper, we use Overlap Cross-Window Attention Block [3] to act as GSA. The last convolution is responsible for enriching low-level patterns. The last part is our proposed Neural Kriging Upsampler, which is able to resample features at arbitrary scales/coordinates in a scale-equivariant manner.

3.3. Scale-Equivariance in Backbone

Existing fixed-scale SR models rarely possess scale-adaptation ability in their backbone networks. To demonstrate this, we compare the feature similarities between the corresponding layer-wise features extracted based on different target scales but from the same trained SR feature extractor with ArbSR [33], as visualized in Figure 2. We also choose the recent SOTA method HAT [3] for reference, i.e., training different models for different target scales. Note that to facilitate reasonable feature similarity measurement, we use Centered Kernel Alignment (CKA) [16, 26] as the metric with higher values indicating greater similarity.

From Figure 2, we make two important observations. First, ArbSR presents consistently high similarity values between different scaled feature extractors over all examined layers, indicating that the previous method is ineffective in extracting scale-related information in its backbone. In contrast, features obtained from our method show apparent divergences when handling different scale factors, demonstrating that our proposed feature extraction backbone is adaptive to changes in scale. For HAT, since its models are separately trained (i.e., optimized for different scales individually), it presents similar visualization as ours, which further verifies that our scale-adaptive feature extraction scheme could well handle arbitrary scales in SR. Second, we find that the differences are not evident in the early feature extraction stages of the network but are amplified from the middle stages, which indicates that later extraction layers are more flexible for scale encoding in SR. This inspires us to inject scale encoding onto multiple proper layers to ensure scale information can be propagated to the consecutive feature up-sampling module.

Scale-Equivariant Extractor. As shown in Figure 3, to address the above issue, we design a pluggable transformer-style information injector named Adaptive Feature Extractor (AFE), which is able to dynamically adjust the extracted features according to the upsampling rate via injecting scale information into the feature extraction process explicitly. The AFE module works in the following way. First, a set of learned convolutional kernel basis is defined, serving as the basic feature extraction pool that matches/supports different scaling factors. In other words, since enumerating the entire real-valued scale space is impossible, sampling a finite set of operators in this space is necessary. Then, the input scale factor is expanded to a higher dimension by sine-cosine encoding [25], which is further mapped to a latent space (has the same dimensions as convolutional kernels) to carry scale information by a linear layer. Third, this scale vector is modulated with the kernel basis convolution layer, forming an *scale interpolated* adaptive convolutional kernel for feature encoding. The above process can be formulated as follows:

$$X_{out} = \mathcal{F}\left(\begin{bmatrix} sin(2^i \pi r/10) \\ cos(2^i \pi r/10) \\ r \end{bmatrix}\right) \otimes \mathcal{W} * X_{in} + \mathcal{B}, \qquad (2)$$

where $\mathcal{F}$ is the map function, $i = (0, 1, 2, \ldots, n)$, $\mathcal{W} \in \mathbb{R}^{C_{out} \times C_{in} \times k \times k}$ denotes the kernels of convolution, $\otimes$ denotes modulation, $*$ denotes convolution, and X_{in} and X_{out} denote the input and output feature maps, respectively. $\mathcal{B} \in \mathbb{R}^{C_{out}}$ is the bias. As shown in Figure 2, our model shows similar characteristics to the reference high-performance model, demonstrating that AFE helps the backbone extract features adaptively.

3.4. Scale-Equivariance in Upsampler

Commonly used feature upsampling operators such as inverse distance interpolation could be regarded as different interpolation kernels; however, since they are usually not isometric, their interpolation behaviors are not sensitive to scale changes. On the contrary, although the 1×1 convolution operator offers ideal scale-equivariant property, it can not aggregate local contextual information. Indeed, it is required to develop a scale-equivariant upsampler that local context learning ability. Inspired by the excellent scale-equivariant nature of Kriging interpolation [27], which is widely used in geophysics and is capable of integrating both spatial and feature correlation between two visual sites, we design a learnable scale-adaptive upsampler for ASISR named Neural Kriging Upsampler.

Kriging Interpolation Revisit. Without loss of generality, interpolation of an unknown value z_0 at point position/coordinate x_0 can be formulated as a linear combination of the values sampled from reference points: $\hat{z_0} = \sum_i^N \lambda_i z_i$, where λ denotes the interpolation weights to be estimated. Note that the position x_0 of the interpolated point on the target image is unknown, which varies according to scale changes. Kriging interpolation, based on Gaussian process [5], is able to estimate the *optimal* (in minimal squared error sense) interpolation weight vector λ based on feature covariances, derived as follows:

$$\langle z(x_0), z(x_j) \rangle = \sum_{i=1}^{N} \lambda \langle z(x_i), z(x_j) \rangle, \forall j = 0, 1, \ldots, n, \qquad (3)$$

$$\underbrace{\begin{bmatrix} c(0,1) \\ c(0,2) \\ \vdots \\ c(0,N) \end{bmatrix}}_{K_{scale}} = \underbrace{\begin{bmatrix} c(1,1) & c(1,2) & \cdots & c(1,N) \\ \vdots & \ddots & & \vdots \\ c(N,1) & \cdots & \ddots & c(N,N) \end{bmatrix}}_{K_{data}} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_N \end{bmatrix}, \qquad (4)$$

where $c(i, j) = \langle z(x_i), z(x_j) \rangle$ denotes the covariance function between two feature points. Note that in Eq. 4, the first term on the right side has no relationship with the unknown point and is thus irrelevant to upsampling scales, while the left term does since the interpolated position x_0 depends on the scaling factor as well as any scale transformation of the target image. Indeed, from the view of the Gaussian process, Kriging can be simplified as follows:

$$\hat{z_0} = \overbrace{\mathcal{K}_{data}^{-1}(\mathcal{D})}^{scale\,independent} \; \overbrace{\mathcal{K}_{scale}(x_0; X, r)}^{scale\,dependent} \mathbf{z} \qquad (5)$$

where $\mathcal{D}$ is the training set, X and $\mathbf{z}$ denote positions and values of observed data, respectively. $\mathcal{K}_{data}$ denotes the *data kernel* function irrelevant to the upsampling process, which learns the prior distribution from observed data. $\mathcal{K}_{scale}$ denotes the *scale kernel* function related to the spatial location of sampling points, governed by the scale information in the interpolation process.

Neural Kriging. Inspired by the scale-equivalent advantage of Kriging interpolation, we propose an upsampling module named Neural Kriging (NK) with a higher learning ability dedicated to ASISR pipeline. Moreover, this novel network module eliminates two main drawbacks of the Kriging method: first, computationally complex and numerical instability due to large matrix inversion, and second, inflexibility to modify the trained and fixed covariance matrix for adapting to on-the-fly data.

As shown in Figure 3, the proposed Neural Kriging module has two collaborating branches. The first branch is a **scale-insensitive** branch that uses SE Blocks [8] to explore the prior relationship between image feature values and spatial location learned from the training data, yielding *data kernels*. Note that different from the original Kriging formulation, scaled target data point coordinates are also injected into the network, which utilizes on-the-fly data to adjust/enhance/update the accuracy of its data kernels, i.e., to endow it with certain data-adaptive nature. The other branch is **scale-sensitive**, which conducts relative position encoding on sampling points with reference to the target coordinates, calculating the spatial geometric relationship through RBF function to obtain the *distance kernels* according to the target scale. We also cascade the RBF function with a MLP to enhance the learning ability. Then, we fuse both kernels through modulation to generate hybrid features that carry both spatial scale relationships and semantic information, followed by a series of 1×1 convolutions (i.e., inherently scale-equivariant/neighborhood-independent operation), forming the HR output.

3.5. Measure of Scale-Equivariance

Mathematically, operator g is completely equivariant to operation t if the formula $g \circ t = t \circ g$ is satisfied. To quantitatively measure the scale-equivariance of the model, following the measurement of translation-equivariance [44], we propose EQ_S, which is defined as the PSNR between two sets of images, obtained by swapping the order of upsampling and downsampling, as the metric of scale-equivariance. The calculation can be formulated as follows:

$$EQ_S(\phi, r) = 10 \log \left[\left(\phi(f(x;r), \theta; r) - f(\phi(x, \theta; r); r) \right)^2 \right] \tag{6}$$

where $\phi(\cdot)$ denotes the network, θ denotes the weights and $f(\cdot)$ denotes bicubic degradation. We will use this tool to analyze the equivariance of different models in Sec. 4.3.

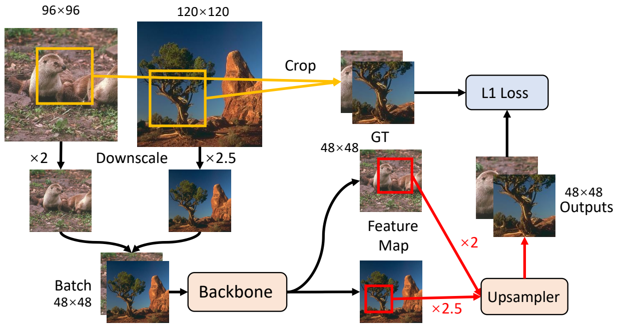

Figure 4. Data processing for arbitrary-scale training.

3.6. Pre-training Strategy

Recent works [3, 19] have demonstrated that pre-training plays an important role in low-level tasks. From the view of Neural Kriging upsampler, pre-training provides the network with more prior knowledge, which has the potential to enhance the representation capability of the data kernel. To this end, we also adopt the pre-training strategy for the training of our EQSR with ImageNet [17], aiming to enhance the generalization and scale-equivariance of the model.

4. Experiments

4.1. Experiment Settings

Datasets and Metrics. Following [3, 31, 45], we employ DF2K dataset [31] as training set. For testing, we adopt five standard benchmarks: Set5 [1], Set14 [41], B100 [23], Urban100 [10] and Manga109 [24]. We conduct our experiments with Bicubic (BI) degradation model [43]. We report peak signal-to-noise ratio (PSNR) results on Y channel (i.e., luminance) of transformed YCbCr space for evaluation.

Data Processing. To enable ASISR training with different training input/output sizes and avoid bias towards fixed scale training samples as in ArbSR [33], we propose a new data processing method to encourage a more general training scale coverage. More concretely, we first downscale HR images to 48×48 pixels, forming stacked input batches. After backbone feature map extraction, only target regions are up-sampled to HR by the learned ASISR model, together with the corresponding regions from the ground-truth HR images, forming $L1$ loss training pairs. In this way, we avoid the inconvenience caused by inputs of different scales. Figure 4 illustrates the above pre-processing pipeline. For data augmentation, we randomly rotate the training images by $90°, 180°, 270°$ and flip them horizontally.

Training Settings. Our model is trained by Adam optimizer [15] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The initial learning rate is set as 2×10^{-4}. Our model is trained for $1000k$ iterations, and we decrease learning rate to half after every $200k$ iterations. If pre-training strategy is enabled, we first train $800k$ iterations on ImageNet [17] and then con-

Table 1. Quantitative comparison (PSNR) for **arbitrary-scale SR** with state-of-the-art methods on benchmark datasets. The best and second-best results are marked in red and blue colors, respectively. "†" indicates that methods adopt the pre-training strategy on ImageNet. "*" indicates the scale is out-of-distribution. Comparison at more scales are available in our supplementary material.

	Set5			Set14			B100			Urban100			Manga109		
	×2	×1.6	×1.55	×2	×1.5	×1.65	×2	×1.4	×1.85	×2	×1.9	×1.95	×2	×1.7	×1.95
Bicubic	33.66	36.10	36.24	30.24	32.87	31.83	29.56	32.95	30.11	26.88	27.25	27.05	30.80	32.91	31.12
MetaSR [9]	38.22	40.66	40.93	34.00	37.51	36.17	32.36	36.95	33.22	33.12	33.62	33.30	39.32	41.30	39.59
ArbSR [33]	38.26	40.69	40.97	34.09	37.53	36.28	32.39	36.93	33.23	33.14	33.55	33.25	39.27	41.32	39.56
LIIF [4]	38.17	40.64	41.00	33.97	37.45	36.25	32.32	36.93	33.14	32.87	33.52	33.20	39.21	41.32	39.52
LTE [18]	38.33	40.75	41.20	34.25	37.79	36.56	32.44	37.05	33.26	33.50	34.11	33.83	39.58	41.69	39.89
Ours	38.35	40.76	41.16	34.45	38.83	36.59	32.46	37.11	33.29	33.62	34.15	33.86	39.44	41.67	39.81
Ours†	38.41	40.83	41.21	34.62	38.05	36.82	32.50	37.18	33.33	33.83	34.45	34.11	39.67	41.87	39.97
	×3	×2.4	×2.75	×3	×2.8	×2.95	×3	×2.2	×2.15	×3	×2.3	×2.35	×3	×2.7	×2.55
Bicubic	30.39	32.41	31.06	27.55	27.84	27.46	27.21	28.88	29.12	24.46	25.91	25.72	26.95	27.77	28.27
MetaSR [9]	34.76	36.58	35.36	30.58	31.00	30.56	29.29	31.44	31.70	28.96	31.43	31.20	34.40	35.55	36.21
ArbSR [33]	34.76	36.59	35.39	30.64	31.01	30.59	29.32	31.48	31.72	28.98	31.48	31.26	34.55	35.64	36.27
LIIF [4]	34.68	36.47	35.38	30.53	30.97	30.56	29.26	31.47	31.67	28.82	31.30	31.09	34.17	35.49	36.18
LTE [18]	34.89	36.66	35.51	30.80	31.28	30.90	29.39	31.59	31.78	29.41	31.95	31.77	34.77	35.94	36.52
Ours	34.83	36.58	35.52	30.82	31.31	30.94	29.42	31.57	31.81	29.53	32.01	31.86	34.89	36.01	36.55
Ours†	34.92	36.65	35.55	30.97	31.50	31.14	29.41	31.62	31.84	29.76	32.29	32.07	34.93	36.04	36.61
	×4	×3.1	×3.25	×4	×3.2	×3.95	×4	×3.2	×3.55	×4	×3.7	×3.85	×4	×3.4	×3.65
Bicubic	28.42	29.89	29.21	26.00	26.98	25.68	25.96	26.91	26.32	23.14	23.38	23.14	24.89	25.97	25.41
MetaSR [9]	32.56	34.46	33.98	28.85	30.08	28.73	27.75	28.86	28.30	26.71	27.25	26.93	31.33	33.00	32.22
ArbSR [33]	32.55	34.50	34.03	28.87	30.08	28.74	27.76	28.93	28.33	26.68	27.22	26.90	31.36	33.12	32.29
LIIF [4]	32.50	34.47	34.12	28.80	30.09	28.88	27.74	28.90	28.34	26.68	27.23	26.94	31.20	32.87	32.11
LTE [18]	32.81	34.69	34.42	29.06	30.35	29.13	27.86	29.03	28.42	27.24	27.86	27.60	31.77	33.42	32.69
Ours	32.71	34.68	34.37	29.12	30.36	29.20	27.86	29.02	28.45	27.30	27.92	27.63	31.86	33.55	32.79
Ours†	32.78	34.74	34.44	29.13	30.48	29.22	27.90	29.05	28.48	27.54	28.09	27.81	32.05	33.69	32.91
	×6*	×5.5*	×6.25*	×6*	×4.25*	×5.25*	×6*	×4.75*	×6.75*	×6*	×5.75*	×6.5*	×6*	×5.25*	×6.75*
Bicubic	24.17	24.46	23.70	23.15	24.17	23.23	23.69	24.25	23.11	20.82	20.73	20.55	21.53	21.83	20.95
MetaSR [9]	29.09	29.96	28.55	26.55	28.47	27.10	25.91	26.92	25.24	24.04	24.37	23.60	27.02	28.19	25.99
ArbSR [33]	28.45	29.24	27.66	26.22	28.46	26.89	25.74	26.89	25.16	23.70	23.81	23.23	26.18	27.59	24.93
LIIF [4]	29.15	30.04	28.73	26.64	28.50	27.38	25.98	26.96	25.53	24.20	24.44	23.79	27.34	28.54	26.37
LTE [18]	29.50	30.20	29.03	26.86	28.55	27.46	26.09	26.98	25.61	24.62	24.79	23.95	27.84	28.96	26.41
Ours	29.41	30.24	28.97	26.79	28.72	27.49	26.07	27.03	25.63	24.66	24.86	24.15	27.97	29.14	26.69
Ours†	29.51	30.38	29.12	26.90	28.78	27.63	26.11	27.10	25.66	24.83	25.10	24.36	28.04	29.36	26.85

duct $200k$ iterations on DF2K dataset. In each iteration, we stack 16 LR patches in a batch with size 48×48 as inputs using the above data processing method. Our proposed model is implemented in PyTorch [28] framework and trained on two Nvidia 3090 GPUs with 24GB RAM.

4.2. Comparisons with the State-of-the-Art

We first compare our EQSR with other state-of-the-art arbitrary-scale SR methods including MetaSR [9], ArbSR [33], LIIF [4] and LTE [18]. Quantitative comparison results are shown in Table 1. Comparisons are conducted for three aspects to comprehensively evaluate the ASISR performances at various up-sampling scales, including 1) integer scale SR results with $\times 2/3/4/6$ settings; 2) real-valued scale SR results with several sampled scales ($\times 1.6/2.4/3.1/3.65$, etc.); and 3) large scale SR results for out-of-distribution ability assessment ($\times 6/6.75$, etc.). Note that all comparison ASISR models are based on one single model protocol. For our method, we also report the results with model pre-training on ImageNet, as denoted by "Ours†". In the meantime, examples of resulting SR images with different up-sampling scales on Urban100 are also illustrated in Figure 5. We also upsample the same image at different scales in Figure 6 to compare with LTE.

From Table 1, Figure 5 and Figure 6 , we make the following observations. First, our model outperforms SOTA approaches at almost all scales (i.e., $\times 2$, $\times 3$, and $\times 4$) on all datasets, showing the superiority of our proposed scale-equivariant model. Second, our method achieves a large gap of $0.21dB$ compared with LTE on the urban100 dataset at out-of-distribution scale $\times 6$, proving the effectiveness of our scale-equivariant upsampler. Third, as shown in Figure 6, our method shows clearer results under various cases, which indicates that our method can effectively deal with arbitrary real-valued scales. Fourth, for the same model, pre-training on ImageNet brings extra performance gain. This could be due to large image variations of the pre-training data, providing an extensive training scale distribu-

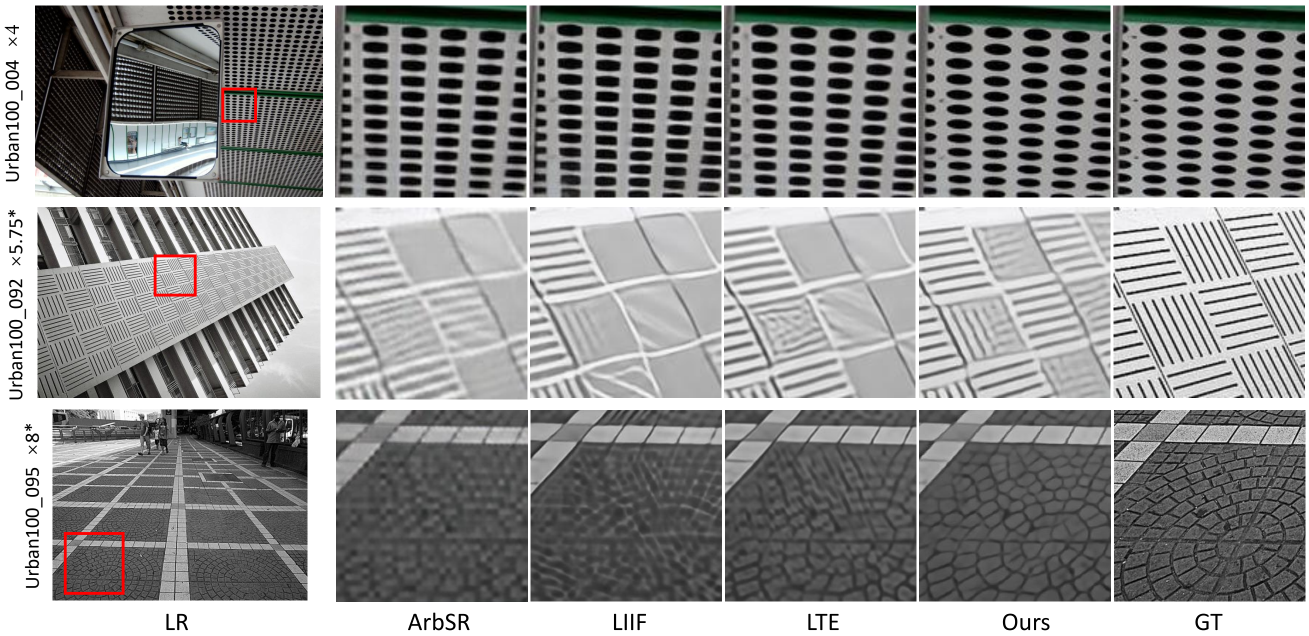

Figure 5. Visual comparison for arbitrary-scale SR models on Urban100 dataset. More results can be found in our Suppl.

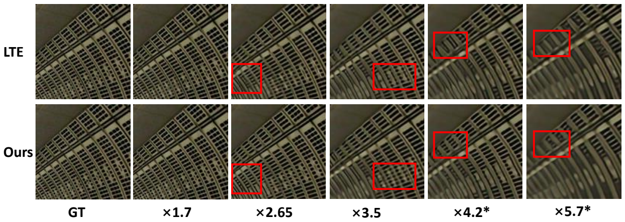

Figure 6. Visual comparison for real-valued SR on the Urban100 dataset. We first downscale the HR image with $\times r$ and then upsample the results with the same scale. "*" indicates out-of-distribution scales. Please enlarge the pdf for details.

tion coverage. From visualized results, we observe that our EQSR has significant advantages in restoring patterns such as repeated textures, edges of cross grids, and other high-frequency dense details which are sensitive to scale changes (for example, in Figure 5, the holes in the first row and the floor tiles in the third row). However, previous methods often damage these repeated structures.

4.3. Analysis

We conduct ablation studies to validate the effects of our proposed two major blocks (i.e., Adaptive Feature Extractor and Neural Kriging Upsampler) and reveal their working mechanisms.

The Effect of Neural Kriging Operator. To support the significance of each component, we first disable the adaptive feature extractor and replace Neural Kriging with Bicubic in the network as our baseline. We conduct two experiments to test Neural Kriging Operator's two branches: Model A uses only data kernels, and Model B uses only scale kernels. As shown in Table 2, Model A outperforms the baseline slightly, which demonstrates that the data kernel is able to acquire prior knowledge for upsampling from the training set. On the contrary, Model B's performance degrades significantly because Model B has scale kernels that perceive scale information only; however, the vital features extracted by the backbone are not effectively used. Then we test Model C, which uses the whole Neural Kriging as the upsampler. The results show that Model C outperforms Model A by a large margin, especially at a higher sampling rate ×6, where a gap of $0.28dB$ is observed. These experiments show that our Neural Kriging effectively perceives scale information and optimizes the results under different sampling conditions.

Table 2. Effects of different modules. We report quantitative results (PSNR) on Urban100 of various network designs for scales $\times 2/4/6$. Note that "NK" is our Neural Kriging module, and "S-A Conv" denotes Scale-Aware Convolution in ArbSR [33].

Model	Config.	Urban100		
		×2	×4	×6
Baseline	EQSR w/o NK/AFE	33.17	26.99	24.23
A	Baseline+K_{data}	33.34	26.93	24.35
B	Baseline+K_{scale}	32.26	25.80	23.79
C	Baseline+NK	33.54	27.26	24.63
D	Model C+S-A Conv.	33.59	27.27	24.52
E	Model C+AFE	33.62	27.30	24.66

The Effect of Adaptive Feature Extractor. To validate the effect of our AFE, we enable Scale-Aware Convo-

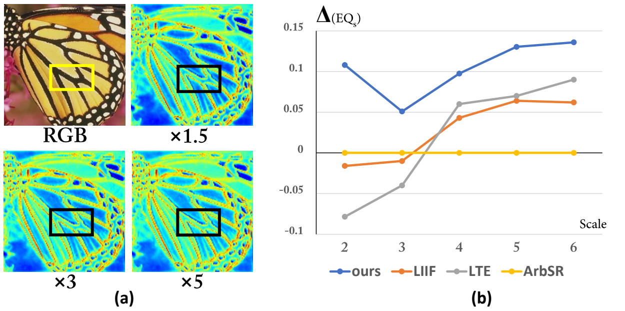

Figure 7. Analysis of equivariance. (a)Visual comparison of feature maps with different upsampling scales. (b) Visualizations of scale-equivariance in EQ_S. We calculate the difference with respect to ArbSR [33] for better comparison.

lution [33] and AFE based on Model C to obtain Model D and Model E, respectively. As shown in Table 2, Model D outperforms Model C at ×2 and ×4 with the help of Scale-Aware Convolution, demonstrating that an adaptive backbone can extract more useful information pertinently. When our AFE is added to Model B's transformer group, the performance continues to increase. Specifically, at the out-of-distribution scale ×6, Model E achieves a gap of $0.14dB$ compared with Model D, demonstrating that our proposed AFE module plays an essential role in our model due to its strong generalization ability. Also note that compared with Scale-Aware Convolution, AFE has a more straightforward structure and significantly fewer parameters (only about 25% of Scale-Aware Convolution). A detailed comparison can be found in our Suppl.

Analysis of Equivariance. To analyze the scale-equivariance of our EQSR, the following two experiments are performed. First, we visualize the extracted features from the end of our model's backbone with different scales in Figure 7(a). Note that we calculate the feature maps in mean and a brighter area indicates a stronger signal. It is observed that feature intensity at edges increases with the increase of scale, proving that our model is capable of extracting different scaled features adaptively.

We also make a quantitative analysis of the different models' equivariance measurement with our proposed EQ_S. To make the comparison more obvious, we calculate $\Delta EQ_S(\phi, r) = EQ_S(\phi, r) - EQ_S(ArbSR, r)$ and the results are shown in Figure 7(b). It can be observed that with the help of AFE and Neural Kriging, our model achieves better equivariance at all the test scales. Specifically, when upsampled with ×2 and ×5, our model outperforms others with a large gap.

Plug-in Ability Demonstration. In addition to our proposed architecture, our key modules AFE and Neural Kriging can also be plugged into most mainstream super-resolution networks to enhance their scale-equivariance ability. Table 3 shows qualitative comparisons with or without our modules for mainstream SR models based on EDSR [21] and SwinIR [20]. We note that models with our operators have consistently achieved performance gain when plugged into different SR models, demonstrating their general ability to encourage scale-equivariance in feature extraction and upsampling. Figure 8 also shows that the performance has a positive correlation with the learning ability of baseline networks, which means that with the development of deep learning, our method can be further improved in the future.

Table 3. Experiments on different backbone networks. "*" indicates combining with our operators.

baseline	Set14			Urban100		
	×2	×3	×4	×2	×3	×4
EDSR	33.95	30.53	28.81	32.95	28.82	26.65
EDSR*	34.06	30.59	28.89	33.10	28.86	26.73
SwinIR	34.14	30.77	28.94	33.40	29.29	27.07
SwinIR*	34.21	30.82	29.02	33.52	29.46	27.28

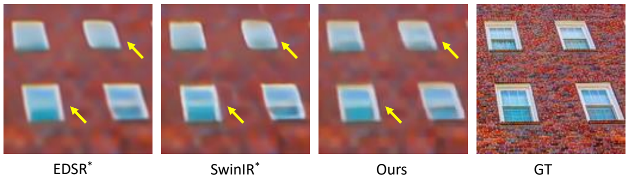

Figure 8. Qualitative comparison on different backbone networks for ×8 SR. "*" indicates combining with AFE and Neural Kriging.

5. Conclusion

In this paper, we revisit the problem of arbitrary-scale super-resolution from the perspective of equivariance. We design an Adaptive Feature Extractor, which can be inserted into most mainstream upsampling networks and enables the model with equivariance. Meanwhile, we explore and design a Neural Kriging upsampling module, which can simultaneously perceive prior information from data and distance information from scale, and integrate both for equivariant upsampling. Combining above modules, we develop a scale-equivariant ASISR model called EQSR which consistently achieves significant improvements over previous methods on integer, decimal, and out-of-distribution scales. Extensive experiments have demonstrated the effectiveness of our method.

Acknowledgments

This work was supported by National Science Foundation of China (U20B2072, 61976137). This work was also partly supported by SJTU Medical Engineering Cross Research Grant YG2021ZD18.

References

[1] Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie Line Alberi-Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. 2012. 5

[2] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12299–12310, 2021. 1

[3] Xiangyu Chen, Xintao Wang, Jiantao Zhou, and Chao Dong. Activating more pixels in image super-resolution transformer. *arXiv preprint arXiv:2205.04437*, 2022. 1, 2, 3, 4, 5

[4] Yinbo Chen, Sifei Liu, and Xiaolong Wang. Learning continuous image representation with local implicit image function. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8628–8638, 2021. 1, 2, 6

[5] Andreas Damianou and Neil D Lawrence. Deep gaussian processes. In *Artificial intelligence and statistics*, pages 207–215. PMLR, 2013. 4

[6] 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. Springer, 2014. 2

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[8] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 5

[9] Xuecai Hu, Haoyuan Mu, Xiangyu Zhang, Zilei Wang, Tieniu Tan, and Jian Sun. Meta-sr: A magnification-arbitrary network for super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1575–1584, 2019. 1, 2, 6

[10] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *CVPR*, pages 5197–5206, 2015. 5

[11] Younghyun Jo and Seon Joo Kim. Practical single-image super-resolution using look-up table. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 691–700, 2021. 1

[12] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, 34:852–863, 2021. 2

[13] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1646–1654, 2016. 2

[14] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Deeply-recursive convolutional network for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1637–1645, 2016. 2

[15] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[16] Simon Kornblith, Jonathon Shlens, and Quoc V Le. Do better imagenet models transfer better? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2661–2671, 2019. 4

[17] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012. 5

[18] Jaewon Lee and Kyong Hwan Jin. Local texture estimator for implicit representation function. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1929–1938, 2022. 1, 2, 6

[19] Wenbo Li, Xin Lu, Jiangbo Lu, Xiangyu Zhang, and Jiaya Jia. On efficient transformer and image pre-training for low-level vision. *arXiv preprint arXiv:2112.10175*, 2021. 1, 5

[20] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1833–1844, 2021. 1, 3, 8

[21] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 136–144, 2017. 2, 8

[22] Jianqi Ma, Zhetong Liang, and Lei Zhang. A text attention network for spatial deformation robust scene text image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5911–5920, 2022. 1

[23] 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 *ICCV*, volume 2, pages 416–423. IEEE, 2001. 5

[24] 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

[25] 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. 4

[26] Thao Nguyen, Maithra Raghu, and Simon Kornblith. Do wide and deep networks learn the same things? uncovering how neural network representations vary with width and depth. *arXiv preprint arXiv:2010.15327*, 2020. 4

[27] Margaret A Oliver and Richard Webster. Kriging: a method of interpolation for geographical information systems. *International Journal of Geographical Information System*, 4(3):313–332, 1990. 4

[28] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 6

[29] Sanghyun Son and Kyoung Mu Lee. Srwarp: Generalized image super-resolution under arbitrary transformation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7782–7791, 2021. 2

[30] Jian Sun, Zongben Xu, and Heung-Yeung Shum. Image super-resolution using gradient profile prior. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008. 2

[31] Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, and Lei Zhang. Ntire 2017 challenge on single image super-resolution: Methods and results. In *CVPR workshops*, pages 114–125, 2017. 5

[32] Radu Timofte, Vincent De Smet, and Luc Van Gool. Anchored neighborhood regression for fast example-based super-resolution. In *Proceedings of the IEEE international conference on computer vision*, pages 1920–1927, 2013. 2

[33] Longguang Wang, Yingqian Wang, Zaiping Lin, Jungang Yang, Wei An, and Yulan Guo. Learning a single network for scale-arbitrary super-resolution. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4801–4810, 2021. 1, 2, 4, 5, 6, 7, 8

[34] Maurice Weiler and Gabriele Cesa. General e (2)-equivariant steerable cnns. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[35] Maurice Weiler, Fred A Hamprecht, and Martin Storath. Learning steerable filters for rotation equivariant cnns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 849–858, 2018. 2

[36] Daniel Worrall and Max Welling. Deep scale-spaces: Equivariance over scale. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[37] Daniel E Worrall, Stephan J Garbin, Daniyar Turmukhambetov, and Gabriel J Brostow. Harmonic networks: Deep translation and rotation equivariance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5028–5037, 2017. 2

[38] Chih-Yuan Yang and Ming-Hsuan Yang. Fast direct super-resolution by simple functions. In *Proceedings of the IEEE international conference on computer vision*, pages 561–568, 2013. 2

[39] Jianchao Yang, Zhaowen Wang, Zhe Lin, Scott Cohen, and Thomas Huang. Coupled dictionary training for image super-resolution. *IEEE transactions on image processing*, 21(8):3467–3478, 2012. 2

[40] Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le. Qanet: Combining local convolution with global self-attention for reading comprehension. *arXiv preprint arXiv:1804.09541*, 2018. 3

[41] 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. Springer, 2010. 5

[42] Kaibing Zhang, Xinbo Gao, Dacheng Tao, and Xuelong Li. Single image super-resolution with non-local means and steering kernel regression. *IEEE Transactions on Image Processing*, 21(11):4544–4556, 2012. 2

[43] Kai Zhang, Wangmeng Zuo, Shuhang Gu, and Lei Zhang. Learning deep cnn denoiser prior for image restoration. In *CVPR*, pages 3929–3938, 2017. 5

[44] Richard Zhang. Making convolutional networks shift-invariant again. In *International conference on machine learning*, pages 7324–7334. PMLR, 2019. 5

[45] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 286–301, 2018. 2, 5

[46] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2472–2481, 2018. 2