Cascaded Local Implicit Transformer for Arbitrary-Scale Super-Resolution

Hao-Wei Chen[*,1,2] Yu-Syuan Xu[*,2] Min-Fong Hong[2] Yi-Min Tsai[2] Hsien-Kai Kuo[2] Chun-Yi Lee[1]

[1]Elsa Lab, National Tsing Hua University   [2]MediaTek Inc.

{jaroslaw1007, cylee}@gapp.nthu.edu.tw

{Yu-Syuan.Xu, romulus.hong, Yi-Min.Tsai, Hsienkai.Kuo }@mediatek.com

Abstract

Implicit neural representation has recently shown a promising ability in representing images with arbitrary resolutions. In this paper, we present a Local Implicit Transformer (LIT), which integrates the attention mechanism and frequency encoding technique into a local implicit image function. We design a cross-scale local attention block to effectively aggregate local features and a local frequency encoding block to combine positional encoding with Fourier domain information for constructing high-resolution images. To further improve representative power, we propose a Cascaded LIT (CLIT) that exploits multi-scale features, along with a cumulative training strategy that gradually increases the upsampling scales during training. We have conducted extensive experiments to validate the effectiveness of these components and analyze various training strategies. The qualitative and quantitative results demonstrate that LIT and CLIT achieve favorable results and outperform the prior works in arbitrary super-resolution tasks.

1. Introduction

Single Image Super-Resolution (SISR) is the process of reconstructing high-resolution (HR) images from their corresponding low-resolution (LR) counterparts. SISR has long been recognized as a challenging task in the low-level vision domain due to its ill-posed nature, and has attracted a number of researchers dedicated to this field of study over the past decade [1–21]. A line of SISR research referred to as '*fixed-scale SR*' [1–15] focuses on extracting feature embeddings from LR images and leveraging these embeddings to upsample images with a predefined factor through learnable deconvolutions [3] or sub-pixel convolutions [4]. Despite their success, many of the proposed approaches necessitate a distinct deep neural network model for each upsampling scale, which is usually restricted to a limited selection of integers (*e.g.*, $2\times$, $3\times$, $4\times$). Such a limitation constrains the potential applications and deployment options of SISR models. To overcome this limitation, approaches for upsampling LR images in a continuous manner via a single model emerge and attracted considerable attention recently.

Over the past few years, arbitrary-scale SR has emerged and attracted considerable attention from researchers [16–21]. Apart from the pioneering work Meta-SR [16], recent
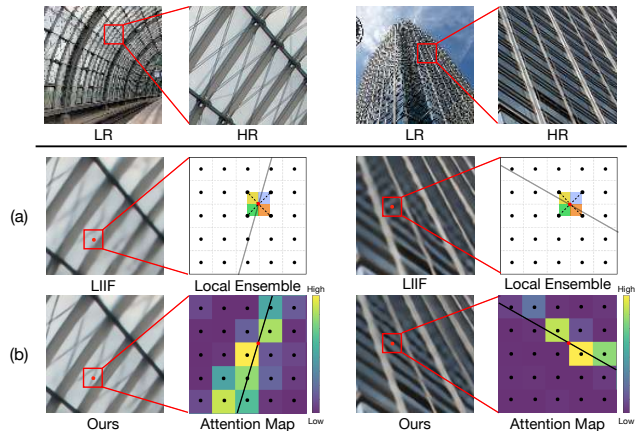
---

*Equal contribution

Figure 1. An illustration and comparison of different approaches that take into account nearby pixels for continuous upsampling: (a) the local ensemble method used in [17], and (b) our proposed local attention mechanism.

endeavors [17–21] have achieved arbitrary-scale SR by replacing the upsampling layers commonly adopted by previous approaches with local implicit image functions, and have demonstrated favorable performance. These local implicit functions employ multi-layer perceptrons (MLPs) to map 2D coordinates and corresponding latent representations to RGB values. Fig. 1 illustrates how different approaches sample latent representations based on the queried coordinates (depicted as the red dots). Fig. 1 (a) illustrates the local ensemble technique adopted by contemporary mainstream methods [17–21]. It calculates the RGB value of the queried coordinate by taking the weighted average of those of the surrounding four pixels based on their relative distances to the queried coordinate. This approach, however, does not consider contextual information and relies solely on distance. For instance, in Fig. 1, the queried coordinates are intentionally designed to lie on edges. However, merely calculating the weighted average of pixels fails to reflect the contextual information about the image content, thereby preventing the accurate capture of the necessary features for performing SR. As a result, although pixel distance plays a vital role in SR tasks, it is essential to concentrate more on the contextual information in an image.

In light of the above observations, we propose a Local Implicit Transformer (LIT), which expands the numbers of referenced latent vectors and accounts for the feature correlation in the context by exploiting the attention mechanism [22]. LIT comprises a Cross-Scale Local Attention Block (CSLAB), a Local Frequency Encoding Block

(LFEB), and a decoder. CSLAB generates attention maps based on the bilinearly interpolated latent vectors at the queried coordinates and the key latent vectors sampled from a grid of coordinates with relative positional bias [23, 24]. The first and second columns of Fig. 1 (b) visualize the attention maps generated by LIT, where the attention areas align closely with the edges. By applying attention maps to feature embeddings, the RGB values of the queried coordinates can be contextually predicted. Moreover, inspired by [21, 25, 26], we introduce LFEB, which projects relative coordinates into latent space to address the spectral bias problem [27] of an implicit neural function. Specifically, relative coordinates are encoded into relative positional encoding, which are multiplied with the frequency encoding extracted from the feature embedding in the Fourier domain to generate the frequency embeddings. This design enables the frequency embedding to integrate the relative positional encoding with texture information, thereby augmenting the expressivity of relative coordinates. Finally, a decoder is adopted to produce RGB values by taking advantage of the attention feature embedding and the frequency embedding.

In order to address the issue of diverse scaling factors and achieve arbitrary-scale super-resolution, it is crucial to consider the role of upsampling factors in constructing high-resolution images within the local implicit image function. However, simultaneously training a local implicit image function with a wide range of upsampling factors (e.g., $1\times \sim 30\times$) poses significant challenges. As a result, we propose a cumulative training strategy to incrementally enhance the fuction's its representative power. The strategy initially trains the local implicit image function with small upsampling factors and then finetunes it with alternatively sampled small and large ones. Furthermore, we present Cascaded LIT (CLIT) to harness the advantages of multi-scale feature embeddings, complementing missing details and information during one-step upsampling. The combination of the cumulative training strategy and CLIT enables efficient and effective handling of arbitrary-scale SR tasks.

The main contributions of our work are summarized as follows: (1) We introduce the LIT architecture, which incorporates the local attention mechanism into arbitrary-scale SR (2) We further develop a cumulative training strategy and the cascaded framework CLIT to effectively handle large-scale upsampling. (3) We carry out comprehensive analyses of the performance impacts for LIT and CLIT. The extensive experimental findings demonstrate that the proposed LIT and CLIT are able to yield remarkable or comparable results across a wide range of benchmark datasets.

The paper is organized as follows. Section 2 reviews the related work. Section 3 walks through the proposed LIT and CLIT frameworks and the implementation details. Section 4 presents the experimental results. Section 5 concludes.

2. Related Work

Implicit neural representation. Implicit neural representation is a technique for representing continuous-domain signals via coordinate-based multi-layer perceptrons (MLPs). Its concept has been adopted in various 3D tasks, *e.g.*, 3D object shape modeling [28–32], 3D scene reconstruction [33–36], and 3D structure rendering [25, 37–39]. For example, NeRF [25] employs implicit neural representation to perform novel view synthesis, which maps coordinates to RGB colors for a specific scene. In the past few years, 2D applications of implicit neural representation have been attempted as well, such as image representation [40, 41] and super-resolution [17–21]. Our work is related to a technique called '*local implicit neural representation*' [17, 21], which encodes LR images to feature embeddings such that similar information could be shared within local regions. Such local implicit neural representations are exploited to upscale LR images to HR ones.

Single image super-resolution. In the past several years, various deep neural network (DNN) based architectures [1–15] have been proposed for SISR. Among these works, SR-CNN [1] pioneered the use of convolutional neural networks (CNNs) to achieve SISR in an end-to-end manner. It is later followed by several subsequent works that incorporated more complicated model architectures, such as residual blocks [6, 7], dense connections [8, 9], attention based mechanisms [10–12], or cascaded frameworks [5, 42, 43], to extract more effective feature representations for SISR. Recently, transformer-based methods [13–15] were introduced to SISR and achieved promising performance.

Arbitrary-scale super-resolution. As discussed in Section 1, most of the contemporary SISR works limit their upsampling scales to specific integer values, and are required to train a distinct model for each upsampling scale. To overcome such a limitation, several approaches [16–21] were proposed to train a unified model for arbitrary upsampling scales. Meta-SR [16] proposed a meta-upscale module for predicting the weights of their convolutional filters from coordinates and scales. The predicted weights are then utilized to perform convolutions to generate HR images. In contrast to Meta-SR, LIIF [17] employs an MLP as a local implicit function, which takes a queried coordinate in an HR image, its nearby feature representations extracted from the corresponding LR image, as well as a cell size to predict an RGB value for that coordinate. UltraSR [18] and IPE [19] extended LIIF by replacing coordinates with the embedded ones to deal with the spectral bias issue [25, 27, 41, 44, 45] inherent in MLPs. LTE [21] further introduced a local texture estimator that transforms coordinates into Fourier domain information to enrich the representational capability of its local implicit function. Different from the above ap-
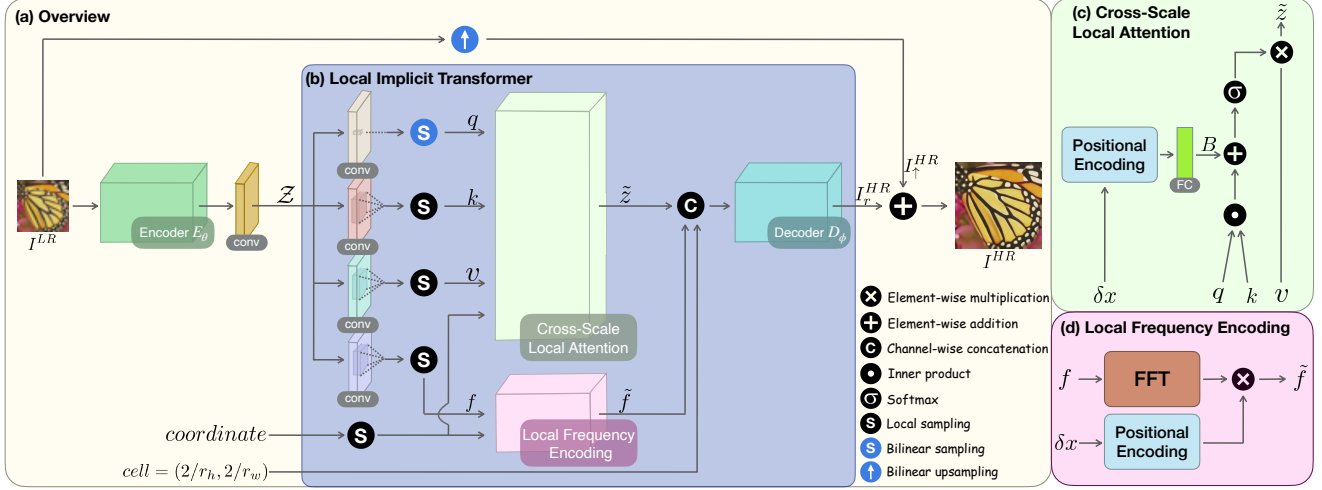
Figure 2. The proposed LIT framework. The local sampling operation samples input embeddings based on a grid of coordinates.

proaches, our proposed methodology exploits a novel local attention mechanism and a cascaded framework to deal with the arbitrary-scale SR. In order to fairly compare with the above approaches, we similarly adopt EDSR [7], RDN [9] and SwinIR [14] as the encoders for our LIT and CLIT.

3. Methodology

In this section, we first provide an overview of the proposed LIT framework, followed by the implementation details of it and its main modules. We then discuss our cumulative training strategy, as well as the framework of CLIT.

3.1. Overview of the LIT Framework

LIT is a framework that employs a novel cross-scaled local attention mechanism and a local frequency encoding technique to perform arbitrary-scale SR tasks. Fig. 2 (a) provides an overview of the proposed framework, which aims at producing an HR image $I^{HR} \in \mathbb{R}^{r_h H \times r_w W \times 3}$ at 2D HR coordinates $\mathbf{x}^{HR} \in \mathcal{X}$ from a given LR image $I^{LR} \in \mathbb{R}^{H \times W \times 3}$ at 2D LR coordinates $\mathbf{x}^{LR} \in \mathcal{X}$ based on an arbitrary upsampling scale $\mathbf{r} = \{r_h, r_w\}$, where $\mathcal{X}$ is the 2D coordinate space that is used to represent an image in the continuous domain. An encoder E_θ first extracts a feature embedding $\mathcal{Z} \in \mathbb{R}^{H \times W \times C}$ from I^{LR}. The extracted $\mathcal{Z}$ is then forwarded into LIT along with the 2D coordinates of I^{HR} and a cell = $(2/s_h, 2/s_w)$ to generate the RGB values of a residual image $I_r^{HR} \in \mathbb{R}^{r_h H \times r_w W \times 3}$ in a pixel-wise fashion. Lastly, the residual image I_r^{HR} is combined with a bilinearly upsampled image $I_\uparrow^{HR} \in \mathbb{R}^{r_h H \times r_w W \times 3}$ via element-wise addition to derive the final HR image I^{HR}.

3.2. Local Implicit Transformer

LIT is developed for mapping any 2D coordinate in the continuous image domain to an RGB color. As highlighted in Fig. 2 (b), it is composed of a cross-scale local attention

block (CSLAB), a local frequency encoding block (LFEB), and a decoder D_ϕ paraterizerized by ϕ. The former two blocks are responsible for estimating a local latent embedding $\tilde{z} \in \mathbb{R}^{G_h G_w \times C}$ and a local frequency embedding $\tilde{f} \in \mathbb{R}^{G_h G_w \times C}$, respectively, where G_h and G_w denote the height and width of local grids employed for performing local coordinate sampling, as depicted in Fig. 3. On the other hand, D_ϕ utilizes these embeddings along with the provided cell to generate I_r^{HR}. More specifically, LIT first projects $\mathcal{Z}$ using four separate convolutional layers to obtain four latent embeddings, corresponding to query q, key k, value v, and frequency f. Based on a queried HR coordinate $x_q \in \mathbf{x}^{HR}$, CSLAB and LFEB estimate $\tilde{z}$ and $\tilde{f}$ as follows:

$$\tilde{z} = CSLAB(\delta x, q, k, v), \tag{1}$$

$$\tilde{f} = LFEB(\delta x, f), \tag{2}$$

$$\delta \mathbf{x} = \left\{ x_q - x^{(i,j)} \right\}_{i \in \{1,2,\dots,G_h\}, j \in \{1,2,\dots,G_w\}}, \tag{3}$$

where $x^{(i,j)} \in \mathbf{x}^{LR}$ denotes an LR coordinate in the local grid indexed by (i,j), and $\delta \mathbf{x}$ represents the set of local relative coordinates defined by Eq. (3). The local grid is sampled in a manner that positions its center $x^{(\lfloor G_h/2 \rfloor + 1, \lfloor G_w/2 \rfloor + 1)}$ at the LR coordinate closest to x_q. The query latent vector $q \in \mathbb{R}^{1 \times C}$ at the HR coordinate x_q is computed by bilinear interpolation, while the remainder of the local latent embeddings $k \in \mathbb{R}^{G_h G_w \times C}$, $v \in \mathbb{R}^{G_h G_w \times C}$, and $f \in \mathbb{R}^{G_h G_w \times C}$ are sampled at the local LR coordinates $\mathbf{x} = \{x^{i,j}\}_{i \in \{1,2,\dots,G_h\}, j \in \{1,2,\dots,G_w\}}$. With the local latent embedding $\tilde{z}$ and the local frequency embedding $\tilde{f}$, the function of D_ϕ is formulated as follows:

$$I^r(x_q) = D_\phi(\tilde{z}, \tilde{f}, c), \tag{4}$$

where $I^r(x_q)$ is the predicted RGB value at the queried coordinate x_q, and $c = \{HR_{\Delta h}, HR_{\Delta w}\}$ denotes the cell that
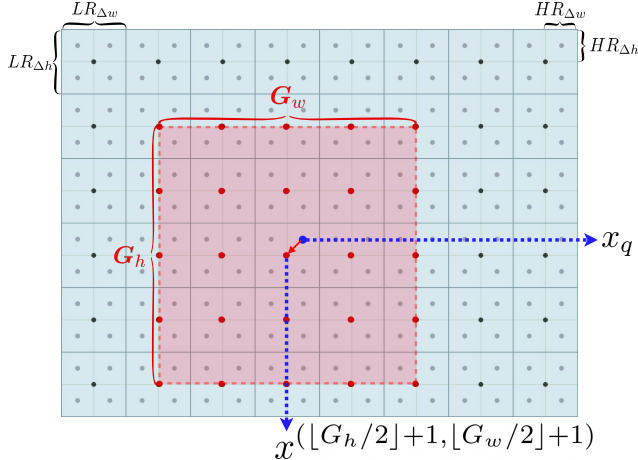
Figure 3. An illustration of the proposed local coordinate sampling scheme. The red rectangular region, outlined by a dashed line, represents the local grid with dimensions $G_h \times G_w$. The red dots indicate the sampled local LR coordinates, while the gray dots correspond to the HR coordinates. The grid dimensions $LR_{\Delta_{h,w}}$ and $HR_{\Delta_{h,w}}$ represent the unit sizes of pixels in LR and HR images, respectively.

represents the height and width of a pixel in an HR image, as illustrated in Fig. 3. D_ϕ is implemented as a five-layer MLP utilizing Gaussian Error Linear Unit (GELU) activation [46], and is employed consistently across all images.

Cross-scale local attention block. LIT exploits CSLAB to perform a local attention mechanism over a local grid to generate a local latent embedding $\tilde{z}$ for each HR coordinate, as illustrated in Fig. 2 (c). CSLAB first calculates the inner product of q and k, adds the relative positional bias B to the result, and obtains an attention matrix. This attention matrix is subsequently normalized by a Softmax operation to produce a local attention map. Finally, CSLAB performs element-wise multiplication of v and the local attention map to derive $\tilde{z}$. The overall procedure is formulated as follows:

$$\tilde{z} = softmax(\frac{qk^\top}{\sqrt{C}} + B) \times v, \qquad (5)$$

$$B = FC(\gamma(\delta\mathbf{x})), \qquad (6)$$

$$\gamma(\delta\mathbf{x}) = [\sin(2^0\delta\mathbf{x}), \cos(2^0\delta\mathbf{x}), ... \\ , \sin(2^{L-1}\delta\mathbf{x}), \cos(2^{L-1}\delta\mathbf{x})], \qquad (7)$$

where C denotes the channel dimension of the local key latent embedding k, FC represents a fully-connected layer, γ is the positional encoding function, and L is a hyperparameter. In this work, L is set to 10, and the multi-head attention mechanism adopted is formulated in Eq. (8) as follows:

$$\tilde{z} = concat(softmax(\frac{q_i k_i^\top}{\sqrt{C/H}} + B_i) \times v_i), \qquad (8)$$

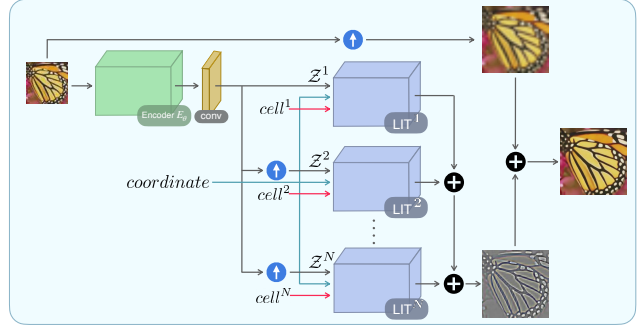where H is the number of attention heads and $i \in [1, ..., H]$.

Figure 4. An overview of the proposed CLIT framework.

Local frequency encoding block. Fig. 2 (d) illustrates the local frequency encoding block, which incorporates dominant-frequency components to generate the local frequency embedding $\tilde{f}$. In order to account for dominant-frequency information, LFEB first encodes local frequency latent embedding f using Fast Fourier Transform (FFT), which is then multiplied by the relative positional encoding to generate $\tilde{f}$. The resultant local frequency embedding $\tilde{f}$ is then concatenated with the flattened local latent embedding $\tilde{z}$ to form a mixing embedding, which can be used by the decoder consisting of MLPs to predict RGB values [47].

3.3. Cumulative Training Strategy

In this section, we discuss our proposed cumulative training strategy, which is developed for enhancing the performance of arbitrary-scale SR. The cumulative training strategy focuses on the schedule of the cell sizes selected during the training phase, as "cell decoding" has been recognized as an essential input to a local implicit image function [17]. Recent studies [17, 19] have observed that the effect of cell decoding on the performance of arbitrary-scale SR is prominent for in-distribution upsampling, but degrades significantly for out-of-distribution large-scale upsampling. Such trends are demonstrated in Table 3 of [17] and Table 4 of [19]. The authors in [21] also constrained their cell sizes to in-distribution ranges to mitigate the negative impact during evaluation. To overcome the degradation issue for out-of-distribution cell sizes, incorporating large cell sizes during training appears to be a promising solution. However, simply training the local implicit image function with a diverse range of cell sizes at once leads to a performance drop. According to our experimental observations presented in Section 4.4, we find that training the local implicit image function by alternatively switching between large and small cells offers positive impacts on the performance. Based on the above observations, we propose a cumulative training strategy which first trains the local implicit image function with large cell sizes, and finetunes it with the alternative training strategy to improve the performance on different upsampling scales. More quantitative results of our training strategies can be found in Section 4.4.

Method	×2	×3	×4	×6	×12	×18	×24	×30
Bicubic [7]	31.01	28.22	26.66	24.82	22.27	21.00	20.19	19.59
EDSR-baseline [7]	34.55	30.90	28.94	-	-	-	-	-
EDSR-baseline-Meta-SR [16, 17]	34.64	30.93	28.92	26.61	23.55	22.03	21.06	20.37
EDSR-baseline-LIIF [17]	34.67	30.96	29.00	26.75	23.71	22.17	21.18	20.48
EDSR-baseline-UltraSR [18]	34.69	31.02	29.05	26.81	23.75	22.21	21.21	20.51
EDSR-baseline-IPE [19]	34.72	31.01	29.04	26.79	23.75	22.21	21.22	20.51
EDSR-baseline-LTE [21]	34.72	31.02	29.04	26.81	23.78	22.23	21.24	20.53
EDSR-baseline-CLIT (Ours)	34.81	31.12	29.15	26.92	23.83	22.29	21.26	20.53
RDN-Meta-SR [16, 17]	35.00	31.27	29.25	26.88	23.73	22.18	21.17	20.47
RDN-LIIF [17]	34.99	31.26	29.27	26.99	23.89	22.34	21.31	20.59
RDN-UltraSR [18]	35.00	31.30	29.32	27.03	23.73	22.36	21.33	20.61
RDN-IPE [19]	35.04	31.32	29.32	27.04	23.93	22.38	21.34	20.63
RDN-LTE [21]	35.04	31.32	29.33	27.04	23.95	22.40	21.36	20.64
RDN-CLIT (Ours)	35.10	31.39	29.39	27.12	24.01	22.45	21.38	20.64
SwinIR-MetaSR [16, 21]	35.15	31.40	29.33	26.94	23.80	22.26	21.26	20.54
SwinIR-LIIF [17, 21]	35.17	31.46	29.46	27.15	24.02	22.43	21.40	20.67
SwinIR-LTE [21]	35.24	31.50	29.51	27.20	24.09	22.50	21.47	20.73
SwinIR-CLIT (Ours)	35.29	31.55	29.55	27.26	24.11	22.51	21.45	20.70

Table 1. The average PSNR (dB) on the DIV2K [48] validation set. The results are obtained from the original manuscripts [17–19, 21]. The best and second-best performing results are highlighted by the red and blue colors, respectively.

3.4. Cascaded Local Implicit Transformer

Previous works [17, 21] typically address arbitrary-scale SR through a single step of upsampling. However, one-step upsampling struggles to reconstruct an HR image when the upsampling scale is large [5, 42, 43]. In light of this, we propose a cascaded upsampling strategy, called "*Cascaded LIT (CLIT)*" to predict residual images from multi-scale feature embeddings. In an N-branched CLIT, the multi-scale feature embeddings $\mathcal{Z}^1, \mathcal{Z}^2, ..., \mathcal{Z}^N$ are derived as follows:

$$\mathcal{Z}^N = \mathcal{Z} \uparrow_{s^1 \times s^2 \times ... \times s^{N-1}},$$
$$\text{where } s^1 = 1 \text{ and } s \in \mathbf{s}, \quad (9)$$

where $\uparrow$ is a bilinear upsampling function and $\mathbf{s}$ is a set of scaling factors, which are configurable hypermeters. For a branch $i, i \in [1, ..., N]$, LIT^i estimates the residual image I_r^i from the feature embedding $\mathcal{Z}^i$ with the coordinate and the corresponding cells. Lastly, the final HR image $I^{HR} \in \mathbb{R}^{r_h H \times r_w W \times 3}$ can be estimated as the following equation:

$$I^{HR} = \lambda^{N-1} I_r^1 + \lambda^{N-2} I_r^2 + ... + \lambda^0 I_r^N + I_\uparrow^{HR}. \quad (10)$$

where λ is a discount factor, with a default value of 0.75. During the training phase, CLIT is trained using the proposed cumulative training strategy. Initially, LIT^1 is trained with the strategy employed by [17]. Subsequently, LIT^1 is finetuned and LIT^2 is initialized by applying the alternative training strategy. By incrementally incorporating LITs into CLIT, the performance is progressively enhanced. The details of the training strategies are specified in Section 4.

4. Experimental Results

In this section, we present the experimental results and discuss their implications. We begin with a brief introduction to our experimental setup in Section 4.1. Following that, we evaluate our CLIT with different datasets in Section 4.2. The, Section 4.3 showcases the learned attention maps. Finally, ablation studies for various configurations of the proposed CLIT and LIT are compared in Section 4.4.

4.1. Experimental Setup

Dataset. We use the DIV2K dataset [48] for network training. It consists of $1,000$ images in 2K resolutions and provides low-resolution counterparts with down-sampling scales, $\times 2, \times 3, \times 4$, which are generated by the bicubic interpolation method. On the other hand, we also evaluate the performance on the validation set of DIV2K [48], Set5 [49], Set14 [50], B100 [51] and Urban100 [52] in terms of peak signal-to-noise (PSNR) values.

Training details During training, we feed batches of size 48×48 low-resolution images into the framework, following the prior works [7]. For each batch, a single upsampling scale is sampled from a uniform distribution $r \sim \mathcal{U}(1, 4)$. With single upsampling scales s, a batch of HR images is cropped into patches of size $48r \times 48r$, while the corresponding LR images are cropped into patches of size 48×48. The patches are augmented by randomly horizontal flipping, vertical flipping, and $90°$ rotating. Then, we sample 48^2 pixels (coordinate-RGB pairs) on each HR patch as the ground-truths. We set the batch size to 32 and use

Method	Set5					Set14					B100					Urban100				
	×2	×3	×4	×6	×8	×2	×3	×4	×6	×8	×2	×3	×4	×6	×8	×2	×3	×4	×6	×8
RDN [9]	38.24	34.71	32.47	-	-	34.01	30.57	28.81	-	-	32.34	29.26	27.72	-	-	32.89	28.80	26.61	-	-
RDN-Meta-SR [16, 17]	38.22	34.63	32.38	29.04	26.96	33.98	30.54	28.78	26.51	24.97	32.33	29.26	27.71	25.90	24.83	32.92	28.82	26.55	23.99	22.59
RDN-LIIF [17]	38.17	34.68	32.50	29.15	27.14	33.97	30.53	28.80	26.64	25.15	32.32	29.26	27.74	25.98	24.91	32.87	28.82	26.68	24.20	22.79
RDN-UltraSR [18]	38.21	34.67	32.49	29.33	27.24	33.97	30.59	28.86	26.69	25.25	32.35	29.29	27.77	26.01	24.96	32.97	28.92	26.78	24.30	22.87
RDN-IPE [19]	38.11	34.68	32.51	29.25	27.22	33.94	30.47	28.75	26.58	25.09	32.31	29.28	27.76	26.00	24.93	32.97	28.82	26.76	24.26	22.87
RDN-LTE [21]	38.23	34.72	32.61	29.32	27.26	34.09	30.58	28.88	26.71	25.16	32.36	29.30	27.77	26.01	24.95	33.04	28.97	26.81	24.28	22.88
RDN-CLIT (Ours)	38.26	34.79	32.69	29.54	27.34	34.09	30.69	28.93	26.83	25.36	32.39	29.33	27.80	26.07	25.00	33.14	29.05	26.93	24.44	23.04
SwinIR [14]	38.35	34.89	32.72	-	-	34.14	30.77	28.94	-	-	32.44	29.37	27.83	-	-	33.40	29.29	27.07	-	-
SwinIR-MetaSR [16, 21]	38.26	34.77	32.47	29.09	27.02	34.14	30.66	28.85	26.58	25.09	32.39	29.31	27.75	25.94	24.86	33.29	29.12	26.76	24.16	22.75
SwinIR-LIIF [17]	38.28	34.87	32.73	29.46	27.36	34.14	30.75	28.98	26.82	25.34	32.39	29.34	27.84	26.07	25.01	33.36	29.33	27.15	24.59	23.14
SwinIR-LTE [21]	38.33	34.89	32.81	29.50	27.35	34.25	30.80	29.06	26.86	25.42	32.44	29.39	27.86	26.09	25.03	33.50	29.41	27.24	24.62	23.17
SwinIR-CLIT (Ours)	38.41	34.97	32.86	29.69	27.62	34.27	30.85	29.08	26.94	25.55	32.46	29.42	27.91	26.15	25.09	33.56	29.43	27.25	24.77	23.33

Table 2. The average PSNR (dB) on Set5 [49], Set14 [50], B100 [51], and Urban100 [52]. The results are obtained from the original manuscripts [17–19, 21]. The best and second-best performing results are highlighted by the red and blue colors, respectively.
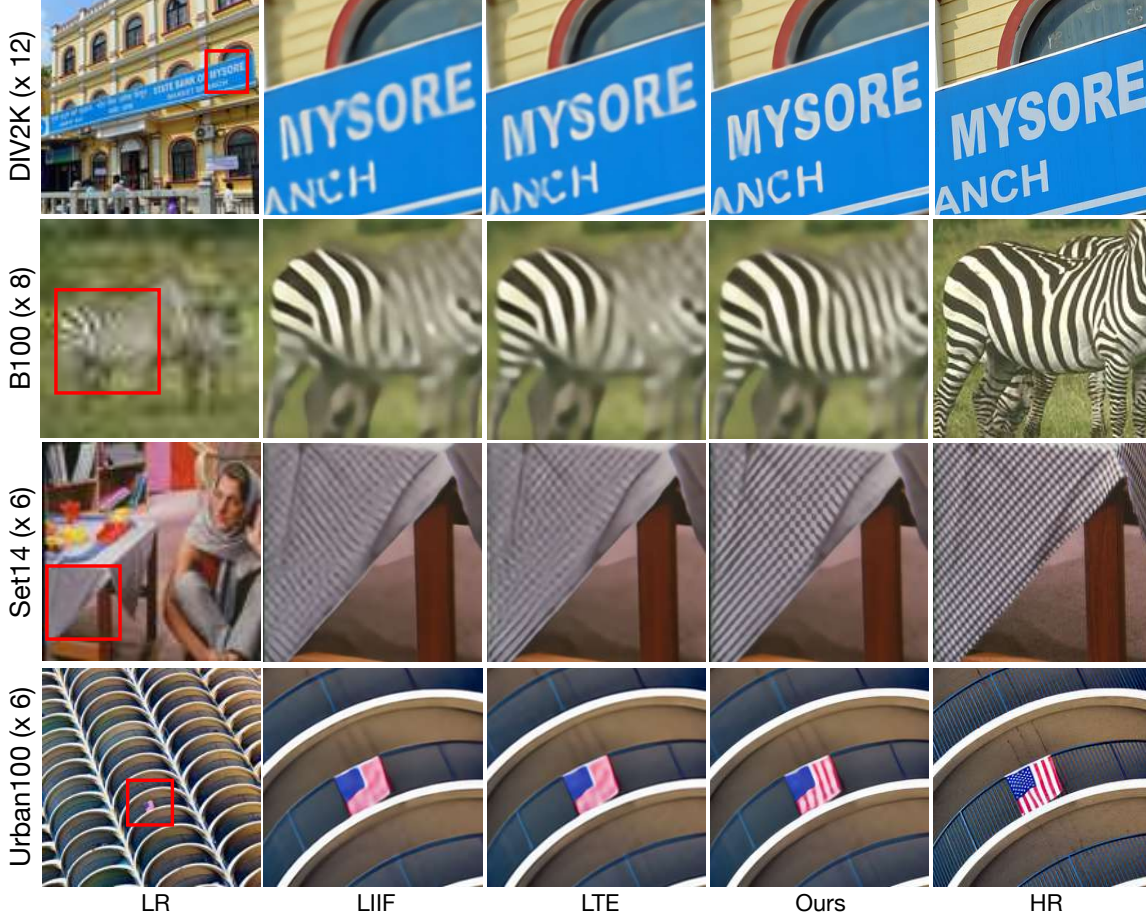

Figure 5. The qualitative results of LIIF [17], LTE [21], and our CLIT with using RDN [9] as the encoder.

the Adam optimizer [53] together with L1 loss for training. We train LIT for $1,000$ epochs, and the learning rate is initialized at $1e^{-4}$ and decayed by a factor 0.5 at [200, 400, 600, 800] epochs. For cumulative training of CLIT, as discussed in Section 3.4, we sample N scale factors $\{s^1, s^2, ..., s^N\}$ from the distribution $\mathcal{U}(1, 4)$ according the number of LITs N in the train step. The total upsampling scale $r = s^1 \times s^2... \times s^N$ is the product of all scale factors. If $48r \times 48r$ is greater than the whole HR image, we clip the scale factor of stage 1. To train N LITs, we fine-tune the model for $500 \times N$ epochs, and the learning rate is initialized at $1e^{-4}$ and decayed by a factor 0.5 at [$100 \times N$, $200 \times N$, $300 \times N$, $400 \times N$] epochs.

4.2. Validation of CLIT

Quantitative results. We first compare our proposed CLIT to existing local implicit neural representation methods for arbitrary-scale SR, including LIIF [17], UltraSR [18], IPE [19], and LTE [21]. Table 1 summarizes the quantitative results in terms of PSNR(dB) on DIV2k [48]. As shown in Table 1, CLIT achieves the best performance when EDSR-Baseline [7], RDN [9] are used as the encoders. These results demonstrate the advantage of CLIT.

Table 2 compares CLIT to prior works [17–19, 21] on widely-used datasets, including Set5 [49], Set14 [50], B100 [51], and Urban100 [52], with RDN and SwinIR.
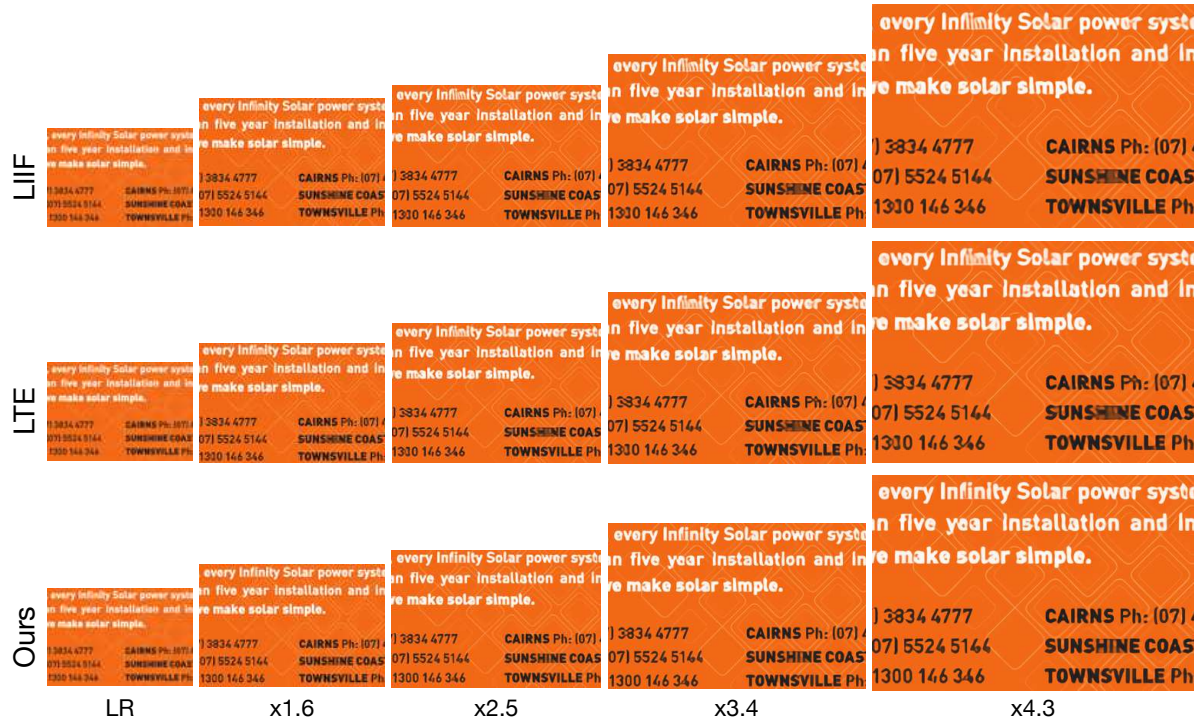
Figure 6. The qualitative results of LIIF [17], LTE [21], and our CLIT using RDN [9] as the encoder and non-integer upsampling scales.

Note that RDN [9] is trained and evaluated for specific upsampling scales. CLIT outperforms those existing methods in most cases across all datasets and scales, even achieving a 0.26dB PSNR improvement on Set5 for the ×8 scale.

Qualitative results. Fig. 5 compares the qualitative results of CLIT with the baseline methods LIIF [17] and LTE [21], on various dataset DIV2k [48], Set14 [50], B100 [51], and Urban100 [52] with variant upsampling scales. The official codes provided by them were used to produce the results. In the first row, the SR visualization results with a ×12 upsampling scale are depicted. It can be observed that both LIIF and LTE struggle to reconstruct the letters continuously, whereas the CLIT result demonstrates continuity for these alphabets, particularly for 'M' and 'S'. In the second row, the stripes on the zebra appear blurry in the LIIF and LTE results, whereas they are more distinct in the CLIT result. In the third row, the cross pattern on the tablecloth is not clearly rendered by LIIF and LTE. In contrast, CLIT is able to successfully generate clean crosses and sharp lines. In the fourth row, despite the blurriness of the original LR image, CLIT captures the texture and produces straight lines on the flag, showcasing its effectiveness.

Fig. 6 presents the results of text image enhancement employing progressively increasing non-integer upsampling factors, alongside a comparison with earlier methods, LIIF [17] and LTE [21]. The input image first undergoes a downscaling process by a factor of 2.2 to generate an LR image. Subsequently, the LR image is then upscaled using predetermined non-integer factors ×1.6, ×2.5, ×3.4, ×4.3.
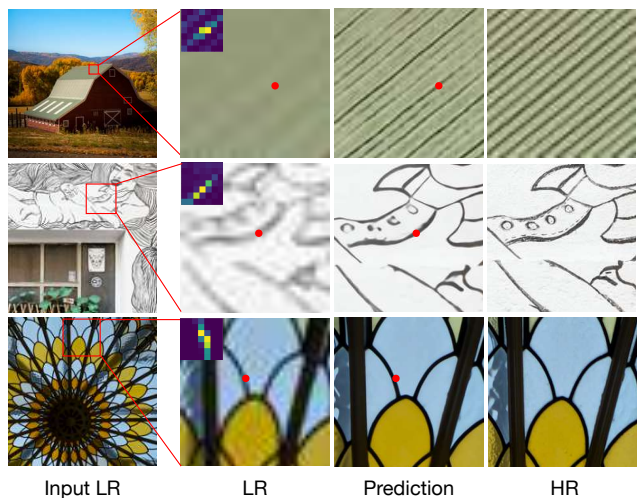

Figure 7. A visualization of the local attention maps of the coordinates highlight as red dots.

It can be observed that our proposed model effectively captures the patterns of text within the image, and accurately estimates words and numbers with enhanced sharpness and clarity. The enhancement becomes particularly evident in the word 'infinity' in the first row of contexts, 'SUNSHINE' in the fifth row, and the number '0' in the final row, as each of these elements demonstrates considerable improvements.

4.3. Visualization Local Attention Maps

In Fig. 7, we providevisualizations of local attention maps. The LR images are generated by applying bicubic downsampling to HR images with scale factors

	In-distribution			Out-of-distribution		
	$\times 2$	$\times 3$	$\times 4$	$\times 6$	$\times 8$	$\times 12$
LIT	**34.79**	31.07	29.10	26.84	25.47	23.77
LIT (-a)	34.71	31.00	29.02	26.76	25.40	23.70
LIT (-f)	34.78	31.07	29.09	26.83	25.46	23.77
LIT (-c)	34.65	31.00	29.04	26.80	25.45	23.77
LIT (+e)	**34.79**	**31.08**	**29.11**	**26.85**	**25.48**	**23.78**

Table 3. The average PSNR (dB) for different design choices. The notations -a/f/c denote the removal of the cross-scale local attention block, frequency encoding block, and cell size, respectively, and +e denotes the addition of local ensemble. The best performing one is indicated in **bolded**.

$\{\times 4.5, \times 6, \times 6\}$. Subsequently, the proposed CLIT with RDN [9] encoder is utilized to produce HR predictions. It can be observed that the attention maps closely align with the edges, indicating that the cross-scale local attention block effectively captures similar latent codes within a local area. This substantiates that our proposed design enables the generation of SR images from LR counterparts, ultimately producing clean and sharp edges in the output.

4.4. Ablation Studies

In this section, we present a series of ablation analyses to substantiate the design decisions proposed in this paper. All of the ablation experiments are conducted on the DIV2K [48] validation set, utilizing the EDSR-baseline [7] as the encoder and employing the PSNR metric for evaluation.

Validation of the design choices. Table 3 presents a summary of the quantitative contributions associated with each LIT component. A significant improvement is observed when adopting the cross-scale local attention block, while a relatively minor gain with the frequency encoding block by comparing LTE with LTE (-a) and LTE (-f), respectively. In addition, LTE (-c) reveals that removing cell decoding leads to more severe performance degradation for in-distribution upsampling scales compared to out-of-distribution counterparts, highlighting the importance of cell decoding for in-distribution upsampling scales. Finally, incorporating the local ensemble technique in LIT (+e) results in a further, albeit modest, enhancement of the overall performance.

The effectiveness of the local grid. Table 4 provides a performance comparison of various local grid sizes employed within the local coordinate sampling scheme, as discussed in Section 3. The results presented in Table 4 reveal that increasing the local grid size contributes to performance enhancements, albeit at the cost of extended training times. As a result, this study adopts a 7×7 local grid size to strike an ideal balance between effectiveness and efficiency.

Analysis on the training strategy. Table 5 presents a quantitative comparison of the proposed cumulative training strategy with the other training strategies for training an LIT. The baseline strategy trains LIT using upsampling scales sampled from a uniform distribution, $r \sim \mathcal{U}(1, 4)$. Expanding the sampling scale distribution to $r \sim \mathcal{U}(1, 12)$ improves the performance of large-scale upsampling while

Local gird $G_h \times G_w$	In-distribution			Out-of-distribution			Training time
	$\times 2$	$\times 3$	$\times 4$	$\times 6$	$\times 8$	$\times 12$	
1×1	34.71	31.00	29.02	26.76	25.40	23.70	20.9
3×3	34.76	31.04	29.07	26.81	25.46	23.76	23.2
5×5	34.78	31.07	29.08	26.83	25.47	23.78	28.4
7×7	34.79	31.07	29.10	26.84	25.47	23.77	33.6
9×9	34.79	31.06	29.10	26.84	25.47	23.77	53.0
15×15	34.83	31.10	29.11	26.86	25.49	23.78	76.1
25×25	**34.86**	**31.13**	**29.15**	**26.88**	**25.50**	**23.80**	190.9

Table 4. The average PSNR (dB) and the training time (hour) of multiple local grid sizes. The **bolded numbers** correspond to the best performance.

Training strategy	$\times 2$	$\times 3$	$\times 4$	$\times 6$	$\times 8$	$\times 12$
Training with $r \sim \mathcal{U}(1, 4)$	**34.79**	31.07	29.10	26.84	25.47	23.77
Training with $r \sim \mathcal{U}(1, 12)$	34.69	31.06	29.11	26.87	25.52	23.82
Alternative training strategy	34.75	31.09	29.12	26.88	25.52	23.81
Cumulative training strategy	34.78	**31.10**	**29.14**	**26.89**	**25.54**	**23.83**

Table 5. The average PSNR (dB) of different training strategies. The best performing results are indicated in **bold**.

compromising the performance of small-scale upsampling. To achieve high-quality results across all upsampling scale ranges, the alternative training strategy trains LIT alternatively by switching the sampling scales between $r \sim \mathcal{U}(1, 4)$ and $r \sim \mathcal{U}(4, 12)$. On the other hand, the proposed cumulative training strategy first trains LIT with $r \sim \mathcal{U}(1, 4)$, followed by fine-tuning using the alternative training strategy. This progressive approach improves the performance of LIT by gradually broadening the distribution of the upsampling scales. Comparing the results of the cumulatively trained LIT with the results of EDSR-baseline-CLIT in Table 1 reveals that the proposed cumulative training strategy works for both LIT and CLIT. Moreover, the results from these tables suggest that the cascaded framework is more effective in addressing arbitrary-scale SR tasks.

5. Conclusion

In this paper, we introduced the attention mechanism and frequency encoding technique to address arbitrary-scale SR tasks. To achieve this objective, we proposed the LIT framework, which consists of a cross-scale local attention block and a local frequency encoding block. The former is designed to find the local latent embedding to reconstruct the corresponding RGB value, while the latter focuses on encoding coordinates with the frequency information derived from the feature embeddings. To enhance the capability of capturing fine details, we proposed CLIT and the corresponding cumulative training strategy that trains the model with progressively increasing upsampling scales. Based on the experimental results, both quantitative and qualitative assessments showcase the superior performance of LIT and CLIT in comparison to existing methods. Furthermore, our comprehensive analyses validated the effectiveness of the training strategies and the components employed by them.

References

[1] C. Dong, C. C. Loy, K. He, and X. Tang. Image super-resolution using deep convolutional networks. *IEEE Trans. Pattern Analysis and Machine Intelligence (TPAMI)*, 38(2): 295–307, 2016. 1, 2

[2] J. Kim, J. K. Lee, and K. M. Lee. Accurate image super-resolution using very deep convolutional networks. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1646–1654, 2016.

[3] C. Dong, C. C. Loy, and X. Tang. Accelerating the super-resolution convolutional neural network. In *Proc. European Conf. on Computer Vision (ECCV)*, pages 391–407, 2016. 1

[4] W. Shi, J. Caballero, F. Huszar, J. Totz, A. P. Aitken, et al. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1874–1883, 2016. 1

[5] W. Lai, J. Huang, N. Ahuja, and M. Yang. Fast and accurate image super-resolution with deep laplacian pyramid networks. *IEEE Trans. Pattern Analysis and Machine Intelligence (TPAMI)*, 41(11):2599–2613, 2019. 2, 5

[6] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 105–114, 2017. 2

[7] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee. Enhanced deep residual networks for single image super-resolution. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition Workshop (CVPRW)*, pages 1132–1140, 2017. 2, 3, 5, 6, 8

[8] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, et al. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proc. European Conf. on Computer Vision Workshop (EC-CVW)*, pages 63–79, 2018. 2

[9] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu. Residual dense network for image super-resolution. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 2472–2481, 2018. 2, 3, 6, 7, 8

[10] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, et al. Image super-resolution using very deep residual channel attention networks. In *Proc. European Conf. on Computer Vision (ECCV)*, pages 294–310, 2018. 2

[11] T. Dai, J. Cai, Y. Zhang, S. Xia, and L. Zhang. Second-order attention network for single image super-resolution. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 11065–11074, 2019.

[12] Y. Mei, Y. Fan, Y. Zhou, L. Huang, T. S. Huang, et al. Image super-resolution with cross-scale non-local attention and exhaustive self-exemplars mining. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 5689–5698, 2020. 2

[13] H. Chen, Y. Wang, T. Guo, C. Xu, Y. Deng, et al. Pre-trained image processing transformer. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 12294–12305, 2021. 2

[14] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, et al. Swinir: Image restoration using swin transformer. In *Proc. IEEE Int. Conf. on Computer Vision Workshop (ICCVW)*, pages 1833–1844, 2021. 3, 6

[15] X. Chen, X. Wang, J. Zhou, and C. Dong. Activating more pixels in image super-resolution transformer. *CoRR*, abs/2205.04437, 2022. 1, 2

[16] X. Hu, H. Mu, X. Zhang, Z. Wang, T. Tan, et al. Meta-sr: A magnification-arbitrary network for super-resolution. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1575–1584, 2019. 1, 2, 5, 6

[17] Y. Chen, S. Liu, and X. Wang. Learning continuous image representation with local implicit image function. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 8628–8638, 2021. 1, 2, 4, 5, 6, 7

[18] X. Xu, Z. Wang, and H. Shi. Ultrasr: Spatial encoding is a missing key for implicit image function-based arbitrary-scale super-resolution. *CoRR*, abs/2103.12716, 2021. 2, 5, 6

[19] Y. Liu, Y. Guo, and S. Zhang. Enhancing multi-scale implicit learning in image super-resolution with integrated positional encoding. *CoRR*, abs/2112.05756, 2021. 2, 4, 5, 6

[20] J. Yang, S. Shen, H. Yue, and K. Li. Implicit transformer network for screen content image continuous super-resolution. In *Proc. Conf. on Neural Information Processing Systems (NeurIPS)*, pages 13304–13315, 2021.

[21] J. Lee and K. H. Jin. Local texture estimator for implicit representation function. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1929–1938, 2022. 1, 2, 4, 5, 6, 7

[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, et al. Attention is all you need. In *Proc. Conf. on Neural Information Processing Systems (NeurIPS)*, pages 5998–6008, 2017. 1

[23] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, et al. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proc. IEEE Int. Conf. on Computer Vision (ICCV)*, 2021. 2

[24] W. Wang, L. Yao, L. Chen, B. Lin, D. Cai, et al. Crossformer: A versatile vision transformer hinging on cross-scale attention. In *Proc. Int. Conf. on Learning Representations (ICLR)*, 2022. 2

[25] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, et al. Nerf: Representing scenes as neural radiance fields for view synthesis. In *Proc. European Conf. on Computer Vision (ECCV)*, pages 405–421, 2020. 2

[26] M. Tancik, P. P. Srinivasan, B. Mildenhall, S. Fridovich-Keil, N. Raghavan, et al. Fourier features let networks learn high frequency functions in low dimensional domains. In *Proc. Conf. on Neural Information Processing Systems (NeurIPS)*, pages 7537–7547, 2020. 2

[27] N. Rahaman, A. Baratin, D. Arpit, F. Draxler, M. Lin, et al. On the spectral bias of neural networks. In *Proc. Int. Conf. on Machine Learning (ICML)*, pages 5301–5310, 2019. 2

[28] Z. Chen and H. Zhang. Learning implicit fields for generative shape modeling. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 5939–5948, 2019. 2

[29] J. J. Park, P. Florence, J. Straub, R. A. Newcombe, and S. Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 165–174, 2019.

[30] L. M. Mescheder, M. Oechsle, M. Niemeyer, S. Nowozin, and A. Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 4460–4470, 2019.

[31] K. Genova, F. Cole, A. Sud, A. Sarna, and T. A. Funkhouser. Local deep implicit functions for 3d shape. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 4856–4865, 2020.

[32] M. Atzmon and Y. Lipman. SAL: sign agnostic learning of shapes from raw data. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 2562–2571, 2020. 2

[33] V. Sitzmann, M. Zollhöfer, and G. Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. In *Proc. Conf. on Neural Information Processing Systems (NeurIPS)*, pages 1119–1130, 2019. 2

[34] C. M. Jiang, A. Sud, A. Makadia, J. Huang, M. Nießner, et al. Local implicit grid representations for 3d scenes. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 6000–6009, 2020.

[35] R. Chabra, J. E. Lenssen, E. Ilg, T. Schmidt, J. Straub, et al. Deep local shapes: Learning local SDF priors for detailed 3d reconstruction. In *Proc. European Conf. on Computer Vision (ECCV)*, pages 608–625, 2020.

[36] S. Peng, M. Niemeyer, L. M. Mescheder, M. Pollefeys, and A. Geiger. Convolutional occupancy networks. In *Proc. European Conf. on Computer Vision (ECCV)*, pages 523–540, 2020. 2

[37] M. Niemeyer, L. Mescheder, M. Oechsle, and A. Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 3501–3512, 2020. 2

[38] L. Liu, J. Gu, K. Z. Lin, T. Chua, and C. Theobalt. Neural sparse voxel fields. In *Proc. Conf. on Neural Information Processing Systems (NeurIPS)*, 2020.

[39] J. T. Barron, B. Mildenhall, M. Tancik, P. Hedman, R. Martin-Brualla, et al. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proc. IEEE Int. Conf. on Computer Vision (ICCV)*, pages 5835–5844, 2021. 2

[40] S. Klocek, Ł. Maziarka, M. Wołczyk, J. Tabor, J. Nowak, et al. Hypernetwork functional image representation. In *Artificial Neural Networks and Machine Learning*, pages 496–510, 2019. 2

[41] V. Sitzmann, J. N. P. Martel, A. W. Bergman, D. B. Lindell, and G. Wetzstein. Implicit neural representations with periodic activation functions. In *Proc. Conf. on Neural Information Processing Systems (NeurIPS)*, pages 7462–7473, 2020. 2

[42] K. Grm, W. J. Scheirer, and V. Struc. Face hallucination using cascaded super-resolution and identity priors. *IEEE Trans. Image Processing*, pages 2150–2165, 2020. 2, 5

[43] Z. Fu, Y. Kong, Y. Zheng, H. Ye, W. Hu, et al. Cascaded detail-preserving networks for super-resolution of document images. In *Int. Conf. on Doc. Analysis Recognition, ICDAR*, pages 240–245, 2019. 2, 5

[44] R. Basri, M. Galun, A. Geifman, D. W. Jacobs, Y. Kasten, et al. Frequency bias in neural networks for input of non-uniform density. In *Proc. Int. Conf. on Machine Learning (ICML)*, pages 685–694, 2020. 2

[45] M. Tancik, B. Mildenhall, T. Wang, D. Schmidt, P. P. Srinivasan, et al. Learned initializations for optimizing coordinate-based neural representations. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 2846–2855, 2021. 2

[46] D. Hendrycks and K. Gimpel. Bridging nonlinearities and stochastic regularizers with gaussian error linear units. *CoRR*, abs/1606.08415, 2016. 4

[47] J. Lee-Thorp, J. Ainslie, I. Eckstein, and S. Ontanon. FNet: Mixing tokens with fourier transforms. In *Proc. Conf. of North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pages 4296–4313, 2022. 4

[48] E. Agustsson and R. Timofte. NTIRE 2017 challenge on single image super-resolution: Dataset and study. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition Workshop (CVPRW)*, pages 1122–1131, 2017. 5, 6, 7, 8

[49] M. Bevilacqua, A. Roumy, C. Guillemot, and M. Alberi-Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In *Proc. British Machine Vision Conf. (BMVC)*, pages 1–10, 2012. 5, 6

[50] R. Zeyde, M. Elad, and M. Protter. On single image scale-up using sparse-representations. In *Curves and Surfaces*, volume 6920 of *Lecture Notes in Computer Science*, pages 711–730, 2010. 5, 6, 7

[51] D. R. Martin, C. C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proc. IEEE Int. Conf. on Computer Vision (ICCV)*, pages 416–425, 2001. 5, 6, 7

[52] J. Huang, A. Singh, and N. Ahuja. Single image super-resolution from transformed self-exemplars. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 5197–5206, 2015. 5, 6, 7

[53] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *Proc. Int. Conf. on Learning Representations (ICLR)*, 2015. 6