Cross Transformer Network for Scale-Arbitrary Image Super-Resolution

Dehong He, Song Wu, Jinpeng Liu, and Guoqiang Xiao(✉)

College of Computer and Information Science,
Southwest University, ChongQing, China
{swu20201514,jinpengliu}@email.swu.edu.cn, {songwuswu,gqxiao}@swu.edu.cn

Abstract. Since implicit neural representation methods can be utilized for continuous image representation learning, pixel values can be successfully inferred from a neural network model over a continuous spatial domain. The recent approaches focus on performing super-resolution tasks at arbitrary scales. However, their magnified images are often distorted and their results are inferior compared to single-scale super-resolution methods. This work proposes a novel CrossSR consisting of a base Cross Transformer structure. Benefiting from the global interactions between contexts through a self-attention mechanism of the Cross Transformer, the CrossSR could efficiently exploit cross-scale features. A dynamic position-coding module and a dense MLP operation are employed for continuous image representation to further improve the results. Extensive experimental and ablation studies show that our CrossSR obtained competitive performance compared to state-of-the-art methods, both for lightweight and classical image super-resolution.

Keywords: Super-resolution · Transformer · Arbitrary scale · Computer vision · Deep learning

1 Introduction

With the rapid development of deep learning and computer vision [7,13,20], image super-resolution has shown a wide range of real-world applications, driving further development in this direction. Image super-resolution is a classical computer vision task, which aims to restore high-resolution images from low-resolution images. Generally, according to the manner of feature extractions, image super-resolution methods can be roughly divided into two categories, i.e., traditional interpolation methods, such as bilinear, bicubic, and deep convolutional neural network-based methods, such as SRCNN [6], DRCN [11], CARN [2], etc. Image super-resolution methods based on CNNs have achieved progressive performance. However, these methods cannot solve the problem of continuous image representation, and additional training is required for each super-resolution scale, which greatly limits the application of CNN-based image super-resolution methods.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022
G. Memmi et al. (Eds.): KSEM 2022, LNAI 13369, pp. 633–644, 2022.
https://doi.org/10.1007/978-3-031-10986-7_51

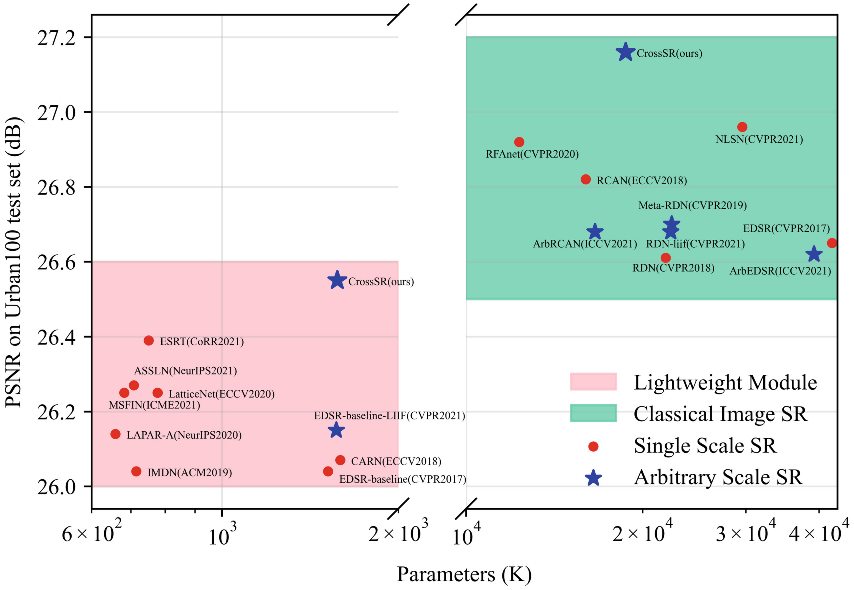

Fig. 1. PSNR results v.s the total number of parameters of different methods for image SR (×4) on Urban100 [9]. Best viewed in color and by zooming in. (Color figure online)

The LIIF [5] was proposed using implicit image representation for arbitrary scale super-resolution, aiming to solve the continuous image representation in super-resolution. However, the implicit function representation of MLP-based LIIF [5] cannot fully utilize the spatial information of the original image, and the EDSR feature encoding module used in LIIF lacks the ability to mine the cross-scale information and long-range dependence of features, so although LIIF has achieved excellent performance in arbitrary-scale super-resolution methods, there is a certain gap compared with the current state-of-the-art single-scale methods.

The currently proposed transformer [21] has attracted a lot of buzz through remarkable performance in multiple visual tasks. The transformer is mainly based on a multi-head self-attention mechanism, which could capture long-range and global interaction among image contexts. The representative methods employing transformers for single-scale image super-resolution are SwinIR [14] and ESRT [17], which obtained superior performance than traditional deep neural networks-based methods.

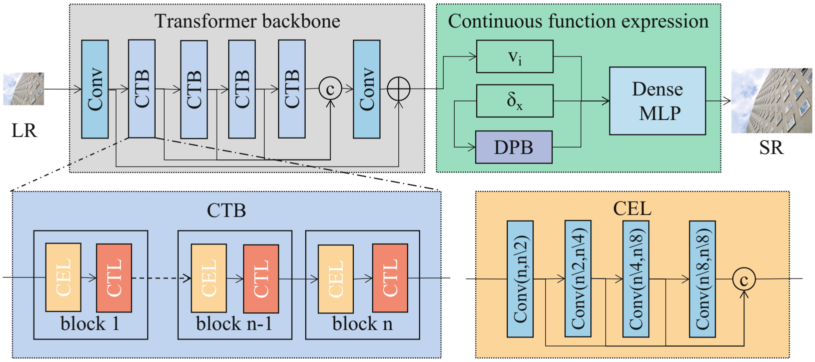

Fig. 2. The architecture of CrossSR is in the top, it consists of two modules: a transformer backbone for feature encoding and an continuous image implicit function representation module. CTB is the inner structure of Cross Transformer Blocks, CEL is the structure of Cross-scale Embedding Layer we proposed, the n and n/2 in Conv(n,n/2) are the number of channels for input and output features.

This paper employs a novel framework of Cross Transformer for effective and efficient arbitrary-scale image super-resolution (CrossSR). Specifically, CrossSR consists of a Cross Transformer-based backbone for feature encoding and an image implicit function representation module. The feature encoding module is mainly based on the framework of Cross Transformer [23] with residual aggregation [16]. The Cross Transformer consists of four stages, and each stage contains multiple cross-scale feature embedding layers and multiple Cross Transformer layers. The image implicit function representation module is based on a modification of LIIF [5]. Specifically, it includes a dynamic position encoding module and a dense MLP module. Extensive experiments on several benchmark datasets and comparisons with several state-of-the-art methods on image super-resolution show that our proposed CrossSR achieves competitive performance with less computing complex. The main contributions of our proposed method include the following three aspects:

- Firstly, a novel image super-resolution network backbone is designed based on a Cross Transformer block [23] combined with a residual feature aggregation [16], and a new cross-scale embedding layer (CEL) is also proposed to reduce the parameters while preserving the performance.
- A series of new network structures are designed for continuous image representation, including dynamic position coding, and dense MLP, which could significantly increase the performance of super-resolution.
- The experiments evaluated on several benchmark datasets show the effectiveness and efficiency of our proposed CrossSR, and it obtained competitive performance compared with state-of-the-art methods on image super-resolution.

2 Proposed Method

2.1 Network Architecture

As shown in Fig. 2, our proposed CrossSR framework consists of two modules: a transformer backbone based on Cross Transformer for feature extraction and a continuous function expression module which utilize a dynamic position encoding block (DPB) [23] and a dense multi-layer perception machine (dense MLP) to reconstruct high quality (HQ) images at arbitrary scales.

Transformer Backbone: Convolutional layers operated at early visual processing could lead to more stable optimization and better performance [25]. Given an input image x, we can obtained the shallow feature through a convolutional layer:

$$F_0 = L_{shallow}(x) \tag{1}$$

where $L_{shallow}$ is the convolutional layer, F_0 is the obtained shallow feature maps.

After that, the deep features F_{LR} are extracted through some Transformer blocks and a convolutional layer. More specifically, the extraction process of the intermediate features and the final deep features can be represented as follows:

$$F_i = L_{CTB_i}(F_{i-1}), i = 1, 2, , , k \tag{2}$$

$$F_{LR} = L_{Conv}(Concat(F_1, F_2, , , F_k)) + F_0 \tag{3}$$

where L_{CTB_i} denotes the i-th Cross Transformer block (CTB) of total k CTBs, L_{Conv} is the last convolutional layer, $Concat(\cdot, \cdot)$ indicates cascading them on the channel direction. Residual Feature Aggregation structure is designed by cascading the output of the transformer block of each layer and passing through a convolutional layer, thus, the residual features of each layer can be fully utilized [16].

Continuous Function Expression: In the continuous function expression module, a continuous image representation function f_θ is parameterized by a dense MLP. The formulation of continuous function expression is as follow:

$$F_{SR} = f_\theta(v_i, \delta_x, dpb(\delta_x)) \tag{4}$$

where F_{SR} is the high-resolution result that to be predicted, v_i is the feature vector for reference, and δ_x is the pixel coordinate information of the F_{SR}, $dpb(\cdot)$ means the dynamic position encoding block [23]. The reference feature vector v_i is extracted from the LR feature map $F_{LR} \in R^{C\times H\times W}$ in its spacial location $\delta_{vi} \in R^{H\times W}$ which is close to δ_x. f_θ is the implicit image function simulated by the dense multi-layer perception machine.

Loss Function: For image super-resolution, the L_1 loss is used to optimize the CrossSR as previous work [5,14,17,19] done,

$$L = |F_{SR} - HR|_1 \tag{5}$$

where F_{SR} is obtained by taking low resolution image as the input of CrossSR, and HR is the corresponding high-quality image of ground-truth.

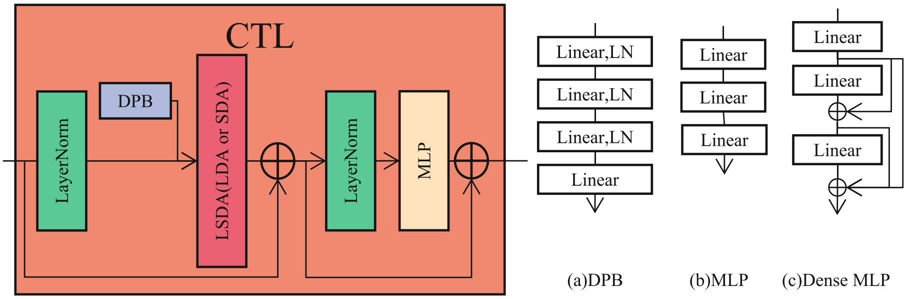

Fig. 3. The architecture of Cross Transformer Layer (CTL) is on the left, SDA and LDA are used alternately in each block. The a diagram on the right shows the architecture of DPB. The middle figure b is the structure of the original MLP, and the rightmost figure c is a schematic of part of the structure of the dense MLP. These three structures contain a ReLU activation layer after each linear layer, which is omitted in the figure.

2.2 Cross Transformer Block

As shown in Fig. 2, CTB is a network structure consisting of multiple groups of small blocks, each of which contains a Cross-scale Embedding Layer (CEL) and a Cross Transformer Layer (CTL). Given the input feature $F_{i,0}$ of the $i-th$ CTB, the intermediate features $F_{i,1}$,$F_{i,2}$,,,$F_{i,n}$ can be extracted by n small blocks as:

$$F_{i,j} = L_{CTL_{i,j}}(L_{CEL_{i,j}}(F_{i,j-1})), j = 1, 2, , , , n \tag{6}$$

where $L_{CEL_{i,j}}(\cdot)$ is the j-th Cross-scale Embedding Layer in the i-th CTB, $L_{CTL_{i,j}}(\cdot)$ is the j-th Cross Transformer Layer in the i-th CTB.

Cross-scale Embedding Layer(CEL): Although Cross Transformer [23] elaborates a Cross-scale Embedding Layer (CEL), it is still too large to be directly applied in image super-resolution. In order to further reduce the number of parameters and the complex operations, a new CEL is designed based on four convolutional layers with different convolutional kernel sizes. Benefiting from each convolution operation with different kernel size is based on the result of the previous convolution, the subsequent convolutions can obtain a substantial perceptual field with only one small convolution kernel, instead of using a 32×32 kernel as large as in Cross Transformer [23]. Moreover, to further reduce the number of complex operations, the dimension of the projection is reduced as the convolution kernel increases. This design could significantly reduce computational effort while achieving excellent image super-resolution results. The output feature of Cross-scale Embedding Layer (CEL) $F_{cel_{i,j}}$ is formulated as:

$$F_{i,j,l} = L_{Conv_l}(F_{i,j,l-1}), l = 1, 2, 3, 4 \tag{7}$$

$$F_{cel_{i,j}} = Concat(F_{i,j,1}, F_{i,j,2}, F_{i,j,3}, F_{i,j,4}) \tag{8}$$

Cross Transformer Layer (CTL): Cross Transformer Layer (CTL) [23] is based on the standard multi-head self-attention of the original Transformer

Table 1. Quantitative comparison (average PSNR/SSIM) with state-of-the-art methods for lightweight image SR on benchmark datasets. The best results are highlighted in red color and the second best is in blue.

Scale	Method	Params	Set5 PSNR/SSIM	Set14 PSNR/SSIM	BSD100 PSNR/SSIM	Urban100 PSNR/SSIM
×2	CARN [2]	1,592 K	37.76/0.9590	33.52/0.9166	32.09/0.8978	31.92/0.9256
	LAPAR-A [12]	548 K	38.01/0.9605	33.62/0.9183	32.19/0.8999	32.10/0.9283
	IMDN [10]	694 K	38.00/0.9605	33.63/0.9177	32.19/0.8996	32.17/0.9283
	LatticeNet [18]	756 K	38.15/0.9610	33.78/0.9193	32.25/0.9005	32.43/0.9302
	ESRT [17]	677 K	38.03/0.9600	33.75/0.9184	32.25/0.9001	32.58/0.9318
	EDSR-baseline [15]	1,370 K	37.99/0.9604	33.57/0.9175	32.16/0.8994	31.98/0.9272
	EDSR-baseline-liif [5]	1,567 K	37.99/0.9602	33.66/0.9182	32.17/0.8990	32.15/0.9285
	CrossSR (ours)	1,574 K	38.13/0.9607	33.99/0.9218	32.27/0.9000	32.63/0.9325
X3	CARN [2]	1,592 K	34.29/0.9255	30.29/0.8407	29.06/0.8034	28.06/0.8493
	LAPAR-A [12]	544K	34.36/0.9267	30.34/0.8421	29.11/0.8054	28.15/0.8523
	IMDN [10]	703 K	34.36/0.9270	30.32/0.8417	29.09/0.8046	28.17/0.8519
	LatticeNet [18]	765 K	34.53/0.9281	30.39/0.8424	29.15/0.8059	28.33/0.8538
	ESRT [17]	770 K	34.42/0.9268	30.43/0.8433	29.15/0.8063	28.46/0.8574
	EDSR-baseline [15]	1,555 K	34.37/0.9270	30.28/0.8417	29.09/0.8052	28.15/0.8527
	EDSR-baseline-liif [5]	1,567 K	34.40/0.9269	30.37/0.8426	29.12/0.8056	28.22/0.8539
	CrossSR (ours)	1,574 K	34.53/0.9283	30.53/0.8460	29.21/0.8082	28.64/0.8616
×4	CARN [2]	1,592 K	32.13/0.8937	28.60/0.7806	27.58/0.7349	26.07/0.7837
	LAPAR-A [12]	659 K	32.15/0.8944	28.61/0.7818	27.61/0.7366	26.14/0.7871
	IMDN [10]	715 K	32.21/0.8948	28.58/0.7811	27.56/0.7353	26.04/0.7838
	MSFIN [24]	682 K	32.28/0.8957	28.66/0.7829	27.61/0.7370	26.25/0.7892
	LatticeNet [18]	777 K	32.30/0.8962	28.68/0.7830	27.62/0.7367	26.25/0.7873
	ESRT [17]	751 K	32.29/0.8964	28.69/0.7844	27.66/0.7384	26.27/0.7907
	ASSLN [28]	708 K	32.19/0.8947	28.69/0.7833	27.69/0.7379	26.39/0.7962
	EDSR-baseline [15]	1,518 K	32.09/0.8938	28.58/0.7813	27.57/0.7357	26.04/0.7849
	EDSR-baseline-liif [5]	1,567 K	32.24/0.8952	28.62/0.7823	27.60/0.7366	26.15/0.7879
	CrossSR (ours)	1,574 K	32.46/0.8975	28.79/0.7856	27.70/0.7405	26.55/0.7995

layer. The main differences lie in short-distance attention (SDA), long-distance attention (LDA), and dynamic position encoding block (DPB). The structure of Cross Transformer is shown in the Fig. 3. For the input image, the embedded features are firstly cropped into small patches to reduce the amount of operations. For short-distance attention, each $G \times G$-adjacent pixel point is cropped into a group. For long-distance attention, pixel points with fixed distance I are grouped together, and then these different grouping features X are used as input for long and short distance attention, respectively. The specific attentions are defined as follows:

$$Attention(Q, K, V) = Softmax(\frac{QK^T}{\sqrt{d}} + B)V \tag{9}$$

where $Q, K, V \in R^{G^2 \times D}$ represent query, key, value in the self-attention module, respectively. And $\sqrt{d}$ is a constant normalizer. $B \in R^{G^2 \times G^2}$ is the position bis matrix. Q, K, V are computed as

$$Q, K, V = X(P_Q, P_K, P_V) \tag{10}$$

where X is the different grouping features for LDA and SDA, P_Q, P_K, P_V are projection matrices implemented through different linear layers.

Next, a multi-layer perception (MLP) is used for further feature transformations. The LayerNorm (LN) layer is added before the LSDA (LDA or SDA) and the MLP, and both modules are connected using residuals. The whole process is formulated as:

$$X = LSDA(LN(X)) + X \tag{11}$$

$$X = MLP(LN(X)) + X \tag{12}$$

2.3 DPB and Dense MLP

Dynamic Position encoding Block (DPB): Image super-resolution aims to recover the high-frequency details of an image. And a well-designed spatial coding operation allows the network to effectively recover the details in visual scenes [26]. With the four linear layers of DPB, we expand the two-dimensional linear spatial input into a 48-dimensional spatial encoding that can more fully exploit the spatial location information, and such design could effectively reduce structural distortions and artifacts in images. The network structure of DPB is shown in Fig. 3.a, and the location information followed the DPB encoding operation is represented as:

$$dpb(\delta_x) = L_4(L_3(L_2(L_1(\delta_x)))) \tag{13}$$

where L_1, L_2, L_3 all consist of three layers: linear layer, layer normalisation, and ReLU. L_4 only consists of one linear layer. Then, the DPB encoded spatial information $dpb(\delta_x)$ and the original location information δ_x are cascaded and input into the dense MLP to predict the high-resolution image as shown in Eq. 4.

Dense MLP: Considering that dense networks have achieved good results in image super-resolution and the fascinating advantages of densenets: they reduce the problem of gradient disappearance, enhance feature propagation, encourage function reuse, and greatly reduce the number of parameters. We design a dense MLP network structure, which connects each layer to each other in a feedforward manner. As shown in Fig. 3.c, for each layer of Dnese MLP, all feature maps of its previous layer are used as input, and its own feature map is used as input of all its subsequent layers.

3 Experiment

3.1 Dataset and Metric

The main dataset we use to train and evaluate our CrossSR is the DIV2K [1] dataset from NTIRE 2017 Challenge. DIV2K consists of 1000 2 K high-resolution images together with the bicubic down-sampled low-resolution images under scale ×2, ×3, and ×4. We maintain its original train validation split, in which we use the 800 images from the train set in training and the 100 images from the

Table 2. Quantitative comparison (average PSNR/SSIM) with state-of-the-art methods for classical image SR on benchmark datasets. NLSA [19] and SWIR [14] train different models for different upsampling scales. The rest methods train one model for all the upsampling scales. The best results are highlighted in red color and the second best is in blue.

Dataset	Method	×2	×	×	×6	×8	×12
Set5	NLSA [19]	38.34/0.9618	34.85/0.9306	32.59/0.9000	–	–	–
	SWIR [14]	38.35/0.9620	34.89/0.9312	32.72/0.9021	–	–	–
	Meta-RDN [8]	38.23/0.9609	34.69/0.9292	32.46/0.8978	28.97/0.8288	26.95/0.7671	24.60/0.6812
	RDN-LIIF [5]	38.17/0.9608	34.68/0.9289	32.50/0.8984	29.15/0.8355	27.14/0.7803	24.86/0.7062
	CrossSR (ours)	38.32/0.9615	34.84/0.9305	32.73/0.9006	29.31/0.8396	27.37/0.7873	24.89/0.7090
Set14	NLSA [19]	34.08/0.9231	30.70/0.8485	28.87/0.7891	–	–	–
	SWIR [14]	34.14/0.9227	30.77/0.8503	28.94/0.7914	–	–	–
	Meta-RDN [8]	33.95/0.9209	30.56/0.8469	28.79/0.7869	26.52/0.6986	25.00/0.6383	23.16/0.5658
	RDN-LIIF [5]	33.97/0.9207	30.53/0.8466	28.80/0.7869	26.64/0.7021	25.15/0.6457	23.24/0.5771
	CrossSR (ours)	34.29/0.9240	30.76/0.8501	28.97/0.7914	26.77/0.7062	25.22/0.6498	23.36/0.5812
BSD100	NLSA [19]	32.43/0.9027	29.34/0.8117	27.78/0.7444	–	–	–
	SWIR [14]	32.44/0.9030	29.37/0.8124	27.83/0.7459	–	–	–
	Meta-RDN [8]	32.34/0.9012	29.26/0.8092	27.72/0.7410	25.91/0.6506	24.83/0.5952	23.47/0.5365
	RDN-LIIF [5]	32.32/0.9007	29.26/0.8094	27.74/0.7414	25.98/0.6540	24.91/0.6010	23.57/0.5445
	CrossSR(ours)	32.41/0.9022	29.37/0.8127	27.84/0.7465	26.06/0.6596	25.00/0.6062	23.62/0.5481
Urban100	NLSA [19]	33.43/0.9394	29.25/0.8726	26.96/0.8109	–	–	–
	SWIR [14]	33.40/0.9393	29.29/0.8744	27.07/0.8164	–	–	–
	Meta-RDN [8]	32.93/0.9356	28.85/0.8662	26.70/0.8017	23.99/0.6927	22.60/0.6182	20.99/0.5281
	RDN-LIIF [5]	32.87/0.9348	28.82/0.8659	26.68/0.8036	24.20/0.7024	22.79/0.6334	21.15/0.5482
	CrossSR (ours)	33.39/0.9393	29.31/0.8745	27.16/0.8164	24.59/0.7191	23.11/0.6496	21.37/0.5604

validation set for testing. Follows many prior works, we also report our model performance on 4 benchmark datasets: Set5 [4], Set14 [27], B100 [3], Urban100 [9]. The SR results are evaluated by PSNR and SSIM metrics on the Y channel of transformed YCbCr space.

3.2 Implementation Details

As with LIIF, we set the input patch size to 48×48. We set the number of channels for the lightweight network and the classic image super-resolution task to 72 and 288, respectively. Our models were trained by ADAM optimizer using $\beta 1 = 0.9$, $\beta 2 = 0.99$ and $\epsilon = 10^{-8}$. The model of the lightweight network was trained for 10^6 iterations with a batch size of 16, and the learning rate was initialized to 1×10^{-4} and then reduced to half at 2×10^5 iterations. In contrast, the classical network has a batch size of 8 and an initial learning rate of 5×10^{-5}. We implemented our models using the PyTorch framework with an RTX3060 GPU.

Table 3. Comparison of PSNR (dB) of non-integer scales of different arbitrary scale super-resolution methods

	Params	Set5			Set14		
		×1.6	×2.4	×3.1	x×1.5	×2.8	×3.2
Bicubic	–	36.10	32.41	29.89	32.87	27.84	26.91
Meta-RDN [8]	21.4M	40.66	36.55	34.42	37.52	30.97	28.90
ArbRCAN [22]	16.6M	40.69	36.59	34.50	37.53	31.01	28.93
ArbRDN [22]	22.6M	40.67	36.55	34.43	37.53	30.98	28.90
RDN-LIIF [5]	21.8M	40.62	36.48	34.49	37.54	31.09	29.97
CrossSR(ours)	18.3M	**40.73**	**36.62**	**34.70**	**37.54**	**31.28**	**30.18**

3.3 Results and Comparison

Table 1 compares the performances of our CrossSR with 8 state-of-the-art light weight SR models. Compared to all given methods, our CrossSR performs best on the four standard benchmark datasets: Set5 [4], Set14 [27], B100 [3], Urban100 [9]. We can find a significant improvement in the results on Urban100. Specifically, a gain of 0.4 dB over EDSR-LIIF [5] on super-resolution for the Urban100 dataset is achieved. This is because Urban100 contains challenging urban scenes that often have cross-scale similarities, as shown in the Fig. 4, and our network can effectively exploit these cross-scale similarities to recover a more realistic image. On other datasets, the gain in PSNR is not as large as the improvement on the Urban100 dataset, but there is still a lot of improvement, all of which is greater than 0.1 dB.

We also compared our method with the state-of-the-art classical image super-resolution methods in Table 2. As can be seen from the data in the table, the results of some current arbitrary-scale methods [5,8] are somewhat worse than those of single-scale super-resolution [14,19]. Our CrossSR is an arbitrary-scale method that simultaneously achieves results competitive with state-of-the-art single-scale methods on different data sets in multiple scenarios, demonstrating the effectiveness of our method.

In Table 3, we also compared the performance of some arbitrary-scale image super-resolution models [8,22] with our CrossSR at different non-integer scales. It can be found that the PSNR results of our model are consistently higher than those of MetaSR and ArbRDN at all scales.

3.4 Ablation Studies

To verify the effectiveness of modified Transformer, DPB and Dense MLP, we conducted ablation experiments in Table 4. Our experiments were performed on the Set5 dataset for x2 lightweight super-resolution.

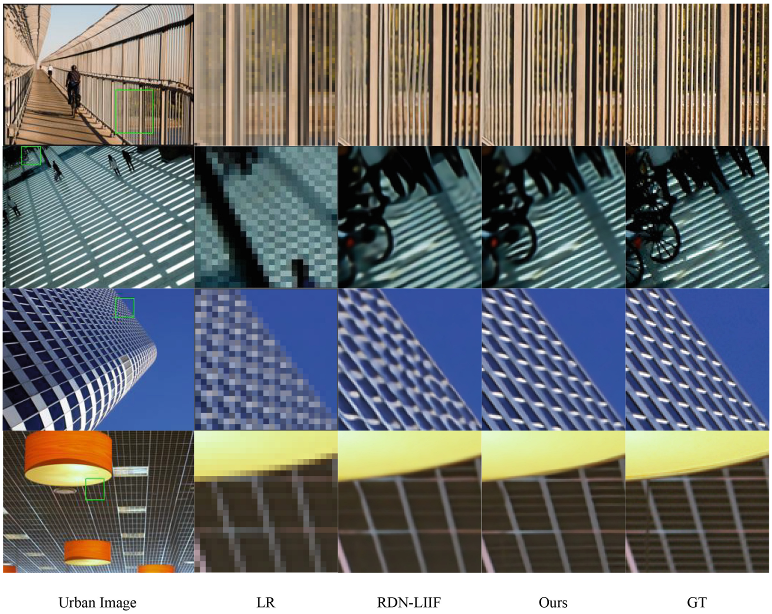

Fig. 4. Visualization comparison with liif on dataset Urban at ×4 SR.

Table 4. Quantitative ablation study. Evaluated on the Set5 validation set for ×2 (PSNR (dB)) after 300 epochs.

Modified transformer	DPB	Dense MLP	PSNR(dB)
×	×	×	37.92
✓	×	×	37.97
×	✓	×	37.93
×	×	✓	37.96
✓	✓	✓	38.01

As can be seen from the data in Table 4, the addition of the modified Transformer improved the test set by 0.05 dB compared to the baseline method, and the addition of DPB and Dense MLP improved it by 0.01 dB and 0.04 dB respectively. This demonstrates the effectiveness of the Transformer backbone, DPB and Dense MLP.

4 Conclusions

In this paper, a novel CrossSR framework has been proposed for image restoration models of arbitrary scale on the basis of Cross Transformer. The model consists of two components: feature extraction and continuous function representation. In particular, a Cross Transformer-based backbone is used for feature extraction, a dynamic position-coding operation is used to incorporate spatial information in continuous image representation fully, and finally, a dense MLP for continuous image fitting. Extensive experiments have shown that CrossSR achieves advanced performance in lightweight and classical image SR tasks, which demonstrated the effectiveness of the proposed CrossSR.

References

1. Agustsson, E., Timofte, R.: NTIRE 2017 challenge on single image super-resolution: dataset and study. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2017, Honolulu, HI, USA, 21–26 July 2017, pp. 1122–1131. IEEE Computer Society (2017)
2. Ahn, N., Kang, B., Sohn, K.-A.: Fast, accurate, and lightweight super-resolution with cascading residual network. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11214, pp. 256–272. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01249-6_16
3. Arbelaez, P., Maire, M., Fowlkes, C., Malik, J.: Contour detection and hierarchical image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **33**(5), 898–916 (2010)
4. Bevilacqua, M., Roumy, A., Guillemot, C., Alberi-Morel, M.: Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In: Bowden, R., Collomosse, J.P., Mikolajczyk, K. (eds.) British Machine Vision Conference, BMVC 2012, Surrey, UK, 3–7 September 2012, pp. 1–10. BMVA Press (2012)
5. Chen, Y., Liu, S., Wang, X.: Learning continuous image representation with local implicit image function. In: CVPR, Computer Vision Foundation/IEEE, pp. 8628–8638 (2021)
6. Dong, C., Loy, C.C., He, K., Tang, X.: Image super-resolution using deep convolutional networks. IEEE Trans. Pattern Anal. Mach. Intell. **38**(2), 295–307 (2015)
7. Hu, F., Lakdawala, S., Hao, Q., Qiu, M.: Low-power, intelligent sensor hardware interface for medical data preprocessing. IEEE Trans. Inf Technol. Biomed. **13**(4), 656–663 (2009)
8. Hu, X., Mu, H., Zhang, X., Wang, Z., Tan, T., Sun, J.: Meta-sr: a magnification-arbitrary network for super-resolution. In: CVPR, Computer Vision Foundation/IEEE, pp. 1575–1584 (2019)
9. Huang, J., Singh, A., Ahuja, N.: Single image super-resolution from transformed self-exemplars. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, 7–12 June 2015, pp. 5197–5206. IEEE Computer Society (2015)
10. Hui, Z., Gao, X., Yang, Y., Wang, X.: Lightweight image super-resolution with information multi-distillation network. In: Amsaleg, L., et al., (eds.) Proceedings of the 27th ACM International Conference on Multimedia, MM 2019, Nice, France, 21–25 October 2019, pp. 2024–2032. ACM (2019)
11. Kim, J., Lee, J.K., Lee, K.M.: Deeply-recursive convolutional network for image super-resolution. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, 27–30 June 2016, pp. 1637–1645. IEEE Computer Society (2016)
12. Li, W., Zhou, K., Qi, L., Jiang, N., Lu, J., Jia, J.: LAPAR: linearly-assembled pixel-adaptive regression network for single image super-resolution and beyond (2021). CoRR abs/2105.10422

13. Li, Y., Song, Y., Jia, L., Gao, S., Li, Q., Qiu, M.: Intelligent fault diagnosis by fusing domain adversarial training and maximum mean discrepancy via ensemble learning. IEEE Trans. Ind. Informatics **17**(4), 2833–2841 (2021)
14. Liang, J., Cao, J., Sun, G., Zhang, K., Gool, L.V., Timofte, R.: Swinir: image restoration using swin transformer. In: IEEE/CVF International Conference on Computer Vision Workshops, ICCVW 2021, Montreal, BC, Canada, 11–17 October 2021, pp. 1833–1844. IEEE (2021)
15. Lim, B., Son, S., Kim, H., Nah, S., Lee, K.M.: Enhanced deep residual networks for single image super-resolution. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2017, Honolulu, HI, USA, 21–26 July 2017, pp. 1132–1140. IEEE Computer Society (2017)
16. Liu, J., Zhang, W., Tang, Y., Tang, J., Wu, G.: Residual feature aggregation network for image super-resolution. In: CVPR, Computer Vision Foundation/IEEE, pp. 2356–2365 (2020)
17. Lu, Z., Liu, H., Li, J., Zhang, L.: Efficient transformer for single image super-resolution (2021). CoRR abs/2108.11084
18. Luo, X., Xie, Y., Zhang, Y., Qu, Y., Li, C., Fu, Y.: LatticeNet: towards lightweight image super-resolution with lattice block. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12367, pp. 272–289. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58542-6_17
19. Mei, Y., Fan, Y., Zhou, Y.: Image super-resolution with non-local sparse attention. In: CVPR, Computer Vision Foundation/IEEE, pp. 3517–3526 (2021)
20. Qiu, H., Zheng, Q., Msahli, M., Memmi, G., Qiu, M., Lu, J.: Topological graph convolutional network-based urban traffic flow and density prediction. IEEE Trans. Intell. Transp. Syst. **22**(7), 4560–4569 (2021)
21. Vaswani, A., et al.: Attention is all you need. In: NIPS, pp. 5998–6008 (2017)
22. Wang, L., Wang, Y., Lin, Z., Yang, J., An, W., Guo, Y.: Learning A single network for scale-arbitrary super-resolution. In: ICCV, pp. 4781–4790. IEEE (2021)
23. Wang, W., Yao, L., Chen, L., Cai, D., He, X., Liu, W.: Crossformer: a versatile vision transformer based on cross-scale attention (2021). CoRR abs/2108.00154
24. Wang, Z., Gao, G., Li, J., Yu, Y., Lu, H.: Lightweight image super-resolution with multi-scale feature interaction network. In: 2021 IEEE International Conference on Multimedia and Expo, ICME 2021, Shenzhen, China, 5–9 July 2021, pp. 1–6. IEEE (2021)
25. Xiao, T., Singh, M., Mintun, E., Darrell, T., Dollár, P., Girshick, R.B.: Early convolutions help transformers see better (2021). CoRR abs/2106.14881
26. Xu, X., Wang, Z., Shi, H.: Ultrasr: spatial encoding is a missing key for implicit image function-based arbitrary-scale super-resolution (2021). arXiv preprint arXiv:2103.12716
27. Zeyde, R., Elad, M., Protter, M.: On single image scale-up using sparse-representations. In: Boissonnat, D., et al. (eds.) Curves and Surfaces 2010. LNCS, vol. 6920, pp. 711–730. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-27413-8_47
28. Zhang, Y., Wang, H., Qin, C., Fu, Y.: Aligned structured sparsity learning for efficient image super-resolution. In: Advances in Neural Information Processing Systems, vol. 34 (2021)