A Two-Stage Attentive Network for Single Image Super-Resolution

Jiqing Zhang, Chengjiang Long*, Yuxin Wang*, Haiyin Piao, Haiyang Mei, Xin Yang*, Baocai Yin

***Abstract*—Recently, deep convolutional neural networks (CNNs) have been widely explored in single image super-resolution (SISR) and contribute remarkable progress. However, most of the existing CNNs-based SISR methods do not adequately explore contextual information in the feature extraction stage and pay little attention to the final high-resolution (HR) image reconstruction step, hence hindering the desired SR performance. To address the above two issues, in this paper, we propose a two-stage attentive network (TSAN) for accurate SISR in a coarse-to-fine manner. Specifically, we design a novel multi-context attentive block (MCAB) to make the network focus on more informative contextual features. Moreover, we present an essential refined attention block (RAB) which could explore useful cues in HR space for reconstructing fine-detailed HR image. Extensive evaluations on four benchmark datasets demonstrate the efficacy of our proposed TSAN in terms of quantitative metrics and visual effects. Code is available at https://github.com/Jee-King/TSAN.**

***Index Terms*—single image super-resolution, deep learning, attention mechanism, multi-context block, two-stage, cross-dimension interaction.**

I. Introduction

SINGLE Image Super-Resolution (SISR) refers to reconstructing a visually pleasing high-resolution (HR) image from a low-resolution (LR) one. It is a fundamental topic in the computer vision community and is an intense demand for diverse applications such as medical imaging, security, and surveillance imaging. The key to SISR problem lies in how to effectively extract useful information from the input image and how to leverage extracted features to reconstruct the fine-detailed HR image. Since multiple HR images can be downsampled to the same LR image and it is a one-to-many mapping relation to recover HR images from one LR image, SISR is an ill-posed and still challenging problem in spite that numerous methods have been proposed.

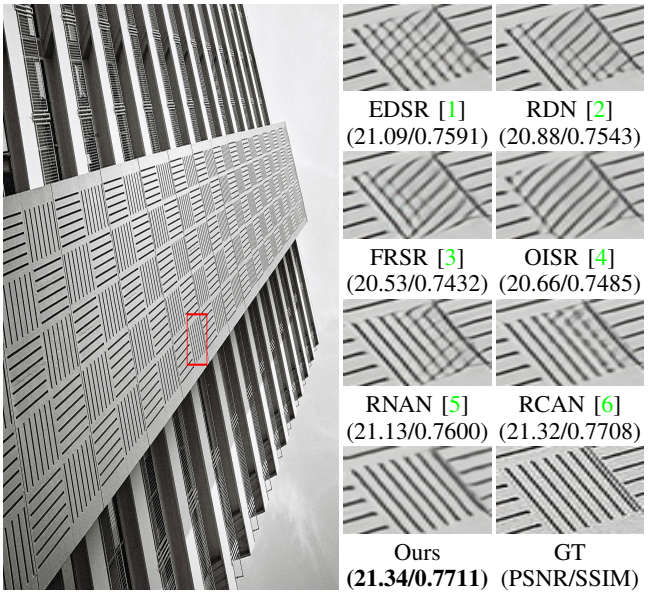


Fig. 1. Visual comparison between different algorithms on *img092* from Urban100 [7] with scale factor $\times 3$. Our TSAN obtains better visual quality and recovers more image details compared with other state-of-the-art SR methods.

As a cutting-edge technique, deep-learning especially convolutional neural networks (CNNs) have been widely used to handle SISR [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [22], [24], [25]. In spite of remarkable progress achieved in SISR, existing CNN-based methods still face with three main limitations: (1) for feature extraction, early methods first apply interpolation strategies (*e.g.*, bicubic) to process the input image to the desired size and then use CNNs to extract features from the upsampled image. As the interpolation often results in visible reconstruction artifacts, some models extract raw features directly from the input LR image and struggle to enhance the ability of feature extraction by simply deepening/widening the network. These methods blindly increase the depth of the network to enhance the performance of the network but ignore taking full use of the contextual information. As the depth of the network increases, the features gradually disappear in the process of transmission; (2) all features are treated equally in these methods, which weakens the discrimination ability of the network to extract more informative features. Although approaches [6], [26] retain some detailed information with channel attention, they struggle in preserving informative textures and restoring

This work was supported in part by the National Natural Science Foundation of China under Grant 91748104, Grant 61972067, Grant 61632006, in part by the National Key Research and Development Program of China under Grant 2018AAA0102003, and the Innovation Technology Funding of Dalian (Project No. 2018J11CY010, 2020JJ26GX036).

Jiqing Zhang, Yuxin Wang, Haiyang Mei, Xin Yang, and Baocai Yin are with the Department of Electronic Information and Electrical Engineering, Dalian University of Technology, Dalian, 116024, China. E-mail: jqz@mail.dlut.edu.cn; wyx@dlut.edu.cn; mhy666@mail.dlut.edu.cn; xinyang@dlut.edu.cn; ybc@dlut.edu.cn.

Chengjiang Long is with JD Finance America Corporation, CA, USA 94043. E-mail: chengjiang.long@jd.com.

Haiyin Piao is with School of Electronics and Information, Northwestern Polytechnical University, Xian, China. E-mail: haiyinpiao@mail.nwpu.edu.cn.

*Xin Yang (xinyang@dlut.edu.cn), Chengjiang Long, and Yuxin Wang are the corresponding authors.

natural details since they ignore to explore the cross-dimension interaction; (3) for HR image reconstruction process, most models reconstruct HR image in one upsampling step at the end of the network, using features learned only in LR space. This setting would increase the difficulties of training for large scaling factors and make the network failed to explore useful cues in HR space for reconstructing visually pleasant HR image.

To address the above limitations, we propose a two-stage attentive network (TSAN) for accurate SISR in this paper. As illustrated in Figure 2, TSAN consists of two stages to solve the SISR problem in a coarse-to-fine manner. At LR-stage, we adopt a dilated residual block (DRB) as a fundamental unit to efficiently extract contextual features and further, based on DRB, propose a multi-context attentive block (MCAB) to make the network focus on more informative contextual features. Multiple MCABs are leveraged to extract attentive contextual features used for reconstructing an initial SR result. At HR-stage, we propose a refined attention block (RAB) to refine the initial SR result to a more fine-detailed one by exploring useful cues in HR space.

Specifically, the DRB pushes the boundaries of conventional cascading and parallel strategies for feature extraction, which could simultaneously distill features with different receptive fields and different context characteristics. Dilated convolutions are adopted in DRB to further explore more contextual information through the larger receptive field. Based on the compact yet powerful DRB, the well-designed MCAB could distill attentive contextual features by introducing the attention mechanism. MCAB contains two branches: a contextual feature extraction branch in which several DRBs are stacked in a dense connection manner to enhance the capability of the network to extract contextual features, and an attention branch that consists of a cutting-splicing block (CSB), a 1^{st}-order attention triplet, and a 2^{nd}-order attention triplet. The CSB is proposed to extract abundant structure cues and self-similarities in local and global regions simultaneously. The purpose of the 1^{st}-order triplet and the 2^{nd}-order triplet is to enhance the discriminative learning ability of the network through the interaction between spatial and channel dimensions. Unlike DRB and MCAB, RAB is designed to explore available cues in HR space to reconstruct fine-detailed HR image. The intuition behind the RAB is that the information in LR space is limited, and we believe features extracted in HR space could benefit a better recovery of images details. As shown in Figure 1, our TSAN obtains better visual quality and recovers more image details than other state-of-the-art SR methods.

To sum up, the main contributions of this paper are three-fold: (1) we propose a two-stage TSAN which could address the SISR problem in a coarse-to-fine manner; (2) we design a novel multi-context attentive block (MCAB) with cross-dimension interaction; (3) our TSAN outperforms the state-of-the-art SISR methods in terms of accuracy and visual effects.

II. RELATED WORK

The related work can be divided into two categories, *i.e.*, *single image super-resolution* and *attention mechanisms*.

A. Single image super-resolution

Single image super-resolution has been extensively studied in the past few decades. Numerous SISR methods have been proposed, ranging from early conventional methods [27], [28], [29], [30] and traditional learning-based methods [7], [31], [32], [33], to recent deep learning-based methods. In particular, deep learning-based methods have led to dramatic improvements in SISR due to the powerful representational capability of deep networks. In this section, we mainly detail the most relevant deep learning-based SISR methods that can be categorized into three types, depending on how the network approaches the SR problem by either pre-upsampling, post-upsampling, or sampling.

For pre-upsampling based methods such as SRCNN [8], VDSR [34], DRCN [35], DRRN [36] and MemNet [37], the upsampling operator, *i.e.*, bicubic interpolation, often results in visible reconstruction artifacts. Moreover, as these methods only learned the mapping in HR space, the raw features cannot be extracted from the original LR image to enhance the representational power of the network.

Post-upsampling based methods like FSRCNN [38], IDN [39], SRResNet [40], EDSR [1], MSRN [41], RCAN [6], RDN [2], CARN [42], RNAN [5], OISR [4], SAN [26], DNCL [43], and IMSSRnet [44], directly extracted features from input LR images and then used the features merely extracted in LR space to construct the HR image by a transposed/sub-pixel convolution layer. However, as these methods focused on extracting features in LR space, a deeper or wider complex network (*e.g.*, EDSR [1] and RDN [2]) was required to obtain sufficient information, for the purpose of reconstructing fine-detailed HR images. Besides, the setting of one-step upsampling at the end of the network would also increase the difficulties of training large scaling factors.

Regarding sampling based methods [45], [46], [47], different sampling strategies were adopted in the network for some specific purposes. For example, LapSRN [45] progressively reconstructed the SR predictions to ease the difficulties of training for large scaling factors. Due to limited available features in the LR space, DBPN [46] proposed an iterative up-and-down sampling approach that could obtain HR features in different depths for SR reconstruction. It is worth mentioning that our proposed TSAN is sampling based. We leverage well-designed MCABs to efficiently extract abundant attentive contextual features with long-range dependencies from the input image in LR space. We also distill HR features from the initial coarse HR image through RAB, aiming at refining more local details.

B. Attention mechanisms

Attention in human perception generally means that human visual systems adaptively process visual information and focus on salient areas. Attention mechanisms have been widely applied in many tasks [48], [49], [50], [51], including image super-resolution [6], [26], [52], [21], [53]. Zhang et al. [6] introduced attention mechanisms into the residual in residual structure to adaptively rescale channel-wise features for image super-resolution. Dai et al. [26] proposed a second-order

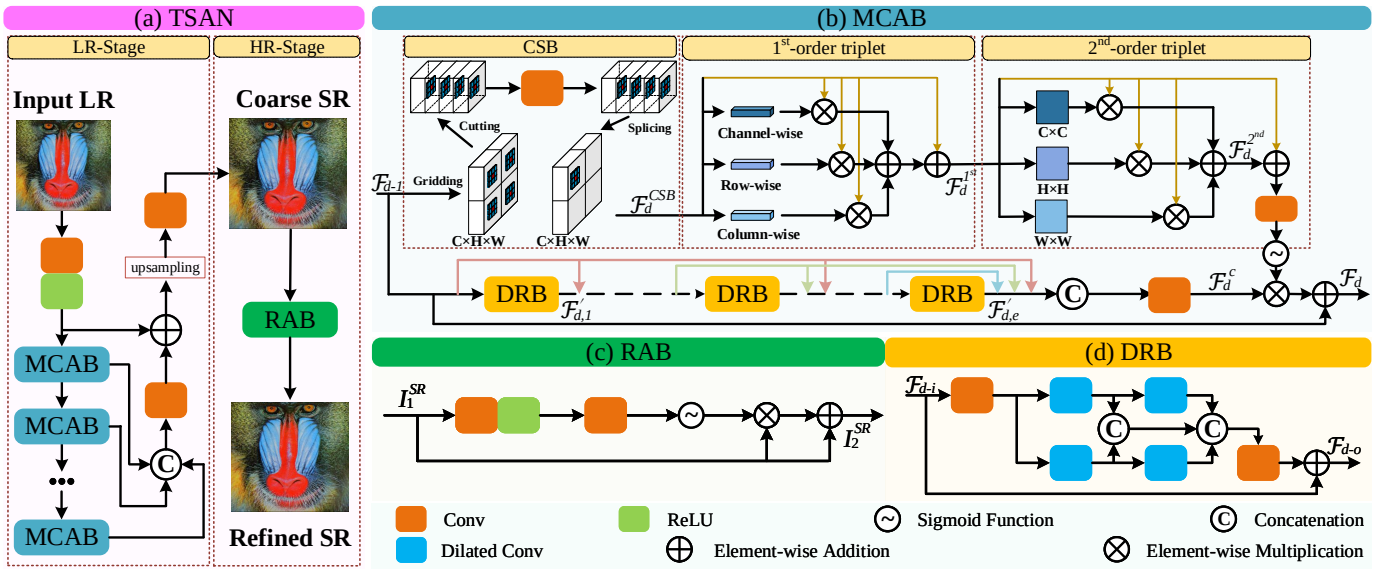


Fig. 2. (a) The overview of our proposed TSAN. TSAN is a two-stage network which reconstructs SR image in a coarse-to-fine manner. In LR-stage, MCABs are leveraged to extract attentive contextual features used for reconstructing an initial SR result. In HR-stage, RAB refines the initial SR result to a more fine-detailed one by exploring useful cues in HR space. (b) Our proposed Multi-Context Attentive Block (MCAB). (c) Our proposed Refined Attention Block (RAB). (d) the Dilated Residual Block (DRB).

channel attention module to learn feature interdependencies by global covariance pooling for more discriminative representations. Hu et al. [21] constructed a set of channel-wise and spatial attention residual blocks and stacked them in a chain structure to dynamically modulate the multi-level features in global and local manners. Du et al. [52] extracted orientation-aware features and combined them by a channel-wise attention mechanism to generate more distinctive features. Wu et al. [53] exploited the advantages of multi-scale and attention mechanisms in SR tasks. However, the inter-dependence between the channel dimension and the spatial dimension is absent in the above-mentioned SISR methods when computing attention on these single pixel channels. Therefore, we propose the 1^{st}-order and 2^{nd}-order triplet attention to focus on inter-dependencies among channel dimension and different spatial dimensions.

III. Methodology

A. Overview

As illustrated in Figure 2, our proposed TSAN consists of two stages to solve the SISR problem in a coarse-to-fine manner. At LR-stage, several multi-context attentive blocks (MCABs) are proposed to efficiently extract sufficient contextual features from the input LR image and construct an initial SR result based on the extracted features. At HR-stage, a simple yet effective refined attention block (RAB) is proposed to further refine the coarse SR result obtained in LR-stage to a more fine-detailed one.

Given an input LR image I^{LR}, we first extract shallow features $\mathcal{F}_s$ by

$$\mathcal{F}_s = \delta(\mathcal{C}_{1\times1}(I^{LR})), \tag{1}$$

where $\mathcal{C}_{k\times k}$ represents convolution operation where kernel size is $k \times k$; δ denotes the rectified linear unit (ReLU) activation function. Then, $\mathcal{F}_s$ is fed to stacked multiple MCABs to distill attentive contextual features of different levels,

$$\mathcal{F}_d = \mathcal{M}_d(\mathcal{F}_{d-1}) = \mathcal{M}_d(\mathcal{M}_{d-1}(\cdots\mathcal{M}_1(\mathcal{F}_s)\cdots)), \tag{2}$$

where $\mathcal{M}_d$ denotes the d-th MCAB and $\mathcal{F}_d$ denotes the output of the d-th MCAB. Here we embed three MCABs, *i.e.*, $d = 3$. Later, all $\mathcal{F}_i, i \in [1, d]$ are fused by applying a convolution layer upon the concatenation, *i.e.*,

$$\mathcal{F}_{fusion} = \mathcal{C}_{1\times1}([\mathcal{F}_1, \mathcal{F}_2, \cdots, \mathcal{F}_d]), \tag{3}$$

where $[\cdot]$ denotes the concatenation operation. After that, we can get an initial SR result by applying a convolution layer upon the upsampled element-wise addition of the aggregated hierarchical features $\mathcal{F}_{fusion}$ and the raw features $\mathcal{F}_s$, *i.e.*,

$$I_1^{SR} = \mathcal{C}_{1\times1}(\kappa(\mathcal{F}_{fusion} + \mathcal{F}_s)), \tag{4}$$

where κ denotes the sub-pixel operation.

Then we design a refined attention block (RAB) (denoted as $\mathcal{R}$) to refine the initial SR result by modeling the local details in HR space in the form of residual:

$$I_2^{SR} = \mathcal{R}(I_1^{SR}). \tag{5}$$

Finally, these two stages are optimized jointly with the loss function defined as

$$\mathcal{L} = w_1\mathcal{L}_m(I_1^{SR}, I^{GT}) + w_2\mathcal{L}_m(I_2^{SR}, I^{GT}), \tag{6}$$

where I^{GT} is the ground truth image, $\mathcal{L}_m$ is the mean absolute error (MAE) loss, and w_1 and w_2 are the balancing parameters.

B. Multi-Context Attentive Block

As not all features contribute a positive effect to the desired SR result, we propose the multi-context attentive block (MCAB) to distill attentive contextual features with long-range

dependencies for high-quality SR reconstruction. Specifically, MCAB contains two branches: a contextual feature extraction branch (the lower part of Figure 2(b)) and an attention branch (the upper part of Figure 2(b)).

1) the contextual feature extraction branch: In the contextual feature extraction branch, the Dilated Residual Block (DRB) is adopted as the fundamental unit to explore more context cues by enlarging receptive field, and simultaneously extract features with different contextual characteristics for reconstructing visually pleasant HR image.

Cascading several convolution layers usually is an effective way to enlarge the receptive fields. In the cascading structure, as shown in Figure 3(a), as a deeper layer accepts the output of a shallower layer, large receptive fields can be produced efficiently. Then the output of each layer would be fused to obtain features covering different scales of receptive fields. Mathematically,

$$\begin{aligned}\mathcal{F}_{co} =&\mathcal{F}_{ci} + \mathcal{C}_{1\times1}([\mathcal{D}^1_{3\times3}(\mathcal{F}'), \mathcal{D}^2_{3\times3}(\mathcal{F}'),\\ &\mathcal{D}^3_{3\times3}(\mathcal{F}'), \mathcal{D}^4_{3\times3}(\mathcal{F}')]),\\ \mathcal{F}' =&\mathcal{C}_{1\times1}(\mathcal{F}_{ci}),\end{aligned} \tag{7}$$

where $\mathcal{F}_{ci}$ and $\mathcal{F}_{co}$ are the input and output of the cascaded network. $\mathcal{D}^n_{k\times k}$ represents dilated convolution operation where kernel size is $k\times k$, and n represents the number of consecutive use of the same dilated convolution on $\mathcal{F}'$. On the other hand, employing a parallel structure can harvest features with different context characteristics. In the parallel structure, as shown in Figure 3(b), as multiple different convolution layers accept the same input and their outputs are concatenated together, the obtained output is indeed a sampling of the input using different contexts. Mathematically,

$$\begin{aligned}\mathcal{F}_{po} =&\mathcal{F}_{pi} + \mathcal{C}_{1\times1}([\mathcal{D}^1_{3\times3}(\mathcal{F}'), \mathcal{D}^1_{3\times3}(\mathcal{F}'),\\ &\mathcal{D}^1_{3\times3}(\mathcal{F}'), \mathcal{D}^1_{3\times3}(\mathcal{F}')]),\\ \mathcal{F}' =&\mathcal{C}_{1\times1}(\mathcal{F}_{pi}),\end{aligned} \tag{8}$$

where $\mathcal{F}_{pi}$ and $\mathcal{F}_{po}$ are the input and output of the parallel network.

To simultaneously distill features with different receptive fields and different context characteristics, we incorporate the advantages of both the cascading and parallel strategies and propose a novel compact structure shown in Figure 2(d). The proposed DRB consists of two branches used for extracting features with different contextual characteristics. Each branch contains two cascaded convolution layers used for distilling features with different receptive fields. In order to obtain a larger receptive field without increasing the number of convolutions and parameters, we adopt dilated convolution. Finally, we concatenate all features of different branches and depths, and fuse them with input by a residual operation. For a clear presentation, DRB can be formulated as:

$$\begin{aligned}\mathcal{F}_{d-o} =&\mathcal{F}_{d-i} + \mathcal{C}_{1\times1}([\mathcal{D}^1_{3\times3}(\mathcal{F}'), \mathcal{D}^1_{3\times3}(\mathcal{F}'),\\ &\mathcal{D}^2_{3\times3}(\mathcal{F}'), \mathcal{D}^2_{3\times3}(\mathcal{F}')]),\\ \mathcal{F}' =&\mathcal{C}_{1\times1}(\mathcal{F}_{d-i}),\end{aligned} \tag{9}$$

where $\mathcal{F}_{d-i}$ and $\mathcal{F}_{d-o}$ denote the input and output of DRB, respectively. The number of output channels for all dilated convolutionlayer in DRB is 64.

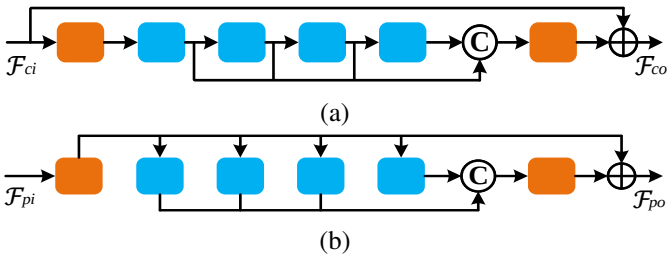


Fig. 3. (a) the cascaded structure; (b) the parallel structure.

We further stack multiple DRBs in a dense connection manner, so that each DRB in MCAB has access to all the previous DRBs' output and could fully utilize them to further distill higher-level contextual features. We then concatenate the outputs of each DRB and integrate them by a 1×1 convolution. The whole process can be expressed as:

$$\mathcal{F}^c_d = \mathcal{C}_{1\times1}([\mathcal{F}_{d-1}, \mathcal{F}'_{d,1}, \cdots, \mathcal{F}'_{d,e}]), \tag{10}$$

where $\mathcal{F}^c_d$ is the output of contextual feature extraction branch in d-th MCAB, $\mathcal{F}'_{d,e}$ denotes the output of e-th DRB in d-th MCAB. We use six DRBs in each MCAB, *i.e.*, $e = 6$, and the dilation rates s of these six DRBs are set to 1, 2, 3, 3, 2, and 1, respectively.

2) the attention branch: The attention branch is designed to make the network focus on more informative features and enhance the discriminative learning ability of the network by considering long-range feature correlations in spatial and channel dimensions. As shown in Figure 2(b), this branch contains three parts: a cutting-splicing block (CSB), a 1^{st}-order triplet, and a 2^{nd}-order triplet.

CSB. To simultaneously capture the spatial dependencies in the local patch and global diagram, we proposed a novel cutting-splicing block (CSB) to extract local patterns and exploit the abundant structure cues and self-similarities in global regions. Formally, the features with a size of $C\times H\times W$ first be cut into $n \times n$ cells ($n = 2$ in Figure 2(b)), and then these cells are concatenated and fed into a 3×3 convolution to extract and aggregate local and non-local information. After that, we splice $n \times n$ cells back into $C \times H \times W$ features. This process can be formulated as:

$$\mathcal{F}^{CSB}_d = \mathcal{O}_s(C_{3\times3}(\mathcal{O}_c(\mathcal{F}_{d-1}))), \tag{11}$$

where $\mathcal{O}_s$ denotes splicing operation, $\mathcal{O}_c$ denotes cutting operation, and $\mathcal{F}^{CSB}_d$ is the output of CSB. The CSB is similar to introduce holes in dilated convolution. The difference is that we also consider the local neighbors. In this way, the local patterns and global diagram are simultaneously guaranteed.

1^{st}-order triplet. After CSB, we design a 1^{st}-order triplet and 2^{nd}-order triplet to model inter-dependencies in different dimensions to find out regions/patterns that should be emphasized in contextual features. As the name implies, each attention triplet consists of three branches which are responsible for capturing cross-dimension interaction between the (C, H), (C, W), and (H, W) dimensions of the input tensor, respectively. By exploiting the inter-dependencies between the channel dimension and the spatial dimension, our network can effectively focus on informative contextual features.

In the 1^{st}-order triplet, we take channel-wise attention as an example, we first aggregate spatial information of (H, W) dimension into a channel-wise descriptor by using average-pooling operation on each channel. Then, the descriptor is forwarded to a shared multi-layer perception ($\mathcal{MLP}$) to produce channel-wise attention maps,

$$\mathcal{A}_{cha}^{1^{st}} = \eta(\mathcal{MLP}(\psi(\mathcal{F}_d^{CSB}))), \tag{12}$$

where η is the sigmoid function, ψ denotes global average pooling operation, $\mathcal{A}_{cha}^{1^{st}}$ denotes the channel-wise attention map. Then we can get channel-wise attentive features $\mathcal{F}_{cha}^{1^{st}}$ by multiplication between $\mathcal{F}_d^{CSB}$ and $\mathcal{A}_{cha}^{1^{st}}$. Similarly, we can obtain the row-wise attentive features $\mathcal{F}_{row}^{1^{st}}$ from (C, H) dimension and the column-wise attentive features $\mathcal{F}_{col}^{1^{st}}$ from (C, W) dimension. Then we add 1^{st}-order triplet attentive features and $\mathcal{F}_d^{CSB}$,

$$\mathcal{F}_d^{1^{st}} = \mathcal{F}_{cha}^{1^{st}} + \mathcal{F}_{row}^{1^{st}} + \mathcal{F}_{col}^{1^{st}} + \mathcal{F}_d^{CSB}, \tag{13}$$

where $\mathcal{F}_d^{1^{st}}$ denotes the output of the 1^{st}-order triplet.

2^{nd}-order triplet. Recent works [54], [26] have shown that second-order statistics in CNNs can provide different information for discriminative representations from the first-order ones. Therefore, we also propose a 2^{nd}-order triplet to learn feature inter-dependencies by cross-dimension interaction like in the 1^{st}-order triplet. In the 2^{nd}-order triplet, we still use channel-wise attention ((H, W) dimension) as an example. Specifically, we first apply average-pooling operation $\mathcal{M}$ along the channel axis on $\mathcal{F}_d^{1^{st}} \in R^{C\times H\times W}$ to generate an efficient feature descriptor. Based on the feature descriptor, we apply a convolution layer and a sigmoid function to generate a spatial attention map $\mathcal{A}_{cha}^{2^{nd}}$ which encodes where to emphasize or suppress. Finally, we perform an multiplication operation between $\mathcal{A}_{cha}^{2^{nd}}$ and $\mathcal{F}_d^{1^{st}}$ to obtain spatial-wise attentive features $\mathcal{F}_{cha}^{2^{nd}}$. The whole process can be expressed as:

$$\begin{aligned} \mathcal{F}_{cha}^{2^{nd}} &= \mathcal{A}_{cha}^{2^{nd}} \times \mathcal{F}_d^{1^{st}}, \\ \mathcal{A}_{cha}^{2^{nd}} &= \eta(\mathcal{C}_{1\times1}(\mathcal{M}(\mathcal{F}_d^{1^{st}}))), \end{aligned} \tag{14}$$

where η is the sigmoid function. Similarly, we generate the row-to-row features $\mathcal{F}_{row}^{2^{nd}}$ from (C, H) dimension and the column-to-column features $\mathcal{F}_{col}^{2^{nd}}$ from (C, W) dimension. Then we add 2^{nd}-order triplet attentive features and $\mathcal{F}_d^{1^{st}}$,

$$\mathcal{F}_d^{2^{nd}} = \mathcal{F}_{cha}^{2^{nd}} + \mathcal{F}_{row}^{2^{nd}} + \mathcal{F}_{col}^{2^{nd}} + \mathcal{F}_d^{1^{st}}, \tag{15}$$

Finally, we can obtain the output of d-th MCAB by,

$$\mathcal{F}_d = \mathcal{F}_d^c \times \eta(\mathcal{C}_{3\times3}(\mathcal{F}_d^{2^{nd}})) + \mathcal{F}_{d-1}. \tag{16}$$

By capturing the inter-dependencies in different dimensions, the attention branch is able to focus on more informative features and enhance discriminative learning ability.

C. Refined Attention Block

As shown in Figure 2(c), our refined attention block (RAB) is proposed to refine a coarse SR result to a more fine-detailed one. The RAB can be simply expressed as,

$$I_2^{SR} = I_1^{SR} \times \eta(\mathcal{C}_{1\times1}(\delta(\mathcal{C}_{3\times3}(I_1^{SR})))) + I_1^{SR}, \tag{17}$$

Note that even though it looks very simple, RAB is essential for reconstructing a visually pleasant SR image with fine details. This is because the information in LR space is limited, and RAB can compensate for the lacked important local information by distilling features in HR space.

D. Implementation Details

We implement our model with Pytorch and run experiments with an NVIDIA Titan V GPU. For training, we use 48×48 RGB patches cropped from LR image as input and its corresponding HR patches as ground truth. Following [1], we preprocess all the images by subtracting the mean RGB value of the DIV2K dataset [55] and augment the training data with random horizontal flips and 90° rotations. We train our model with ADAM optimizer [56] by setting $\beta_1 = 0.9$, $\beta_2 = 0.999$. The mini-batch size is set to 16. The learning rate is initialized as 0.0001 and decreases to half every 200 epochs. And the number of total epochs is 1000. The balancing parameters w_1 and w_2 in Equation 6 are empirically set to 1.

E. Discussions

1) Difference to RDN [2]: Inspired by RDN [2], we introduce dense connection into our MCAB to fully utilize the features information from each DRB in MCAB. There are some differences between RDN [2] and our TSAN. First, RDN [2] mainly cascades convolution layers to enlarge the receptive field, while our TSAN is built based on DRB which can simultaneously distill features with different receptive fields and different context characteristics. Second, RDN [2] focuses on how to exploit and use hierarchical features without considering how to distinguish different feature information. While our TSAN can learn the attention maps used for emphasizing informative contextual features by considering long-range feature correlations in spatial and channel dimensions. Third, RDN belongs to post-upsampling methods which pay attention to exploiting the features information from LR space, while our TSAN is designed to utilize features from LR and HR space simultaneously.

2) Difference to RCAN [6]: We summarize the main differences between RCAN [6] and our TSAN. First, RCAN [6] consists of several residual groups with long skip connections. While, TSAN stack DRBs in a dense connection manner, so each DRB in MCAB has access to all the previous DRBs' output and could fully utilize them to further distill higher-level contextual features. Second, RCAN [6] only extract local information for reconstructing. While TSAN considers non-local operations in CSB to learn long-range feature correlations. Third, RCAN [6] only considers channel attention based first-order feature statistics to enhance the discriminative ability of the network. While our TSAN learns inter-dependencies between different dimensions based on first-order and second-order features.

3) Difference to MSRN [41]: MSRN [41] proposes a multi-scale residual block (MSRB) to detect image features and fuse different scales. There are some main differences between MSRB in MSRN [41] and our proposed DRB. First, we utilize the dilated convolution to widen the receptive field without

additive parameters, which maintains the lightweight structure of DRB. Second, MSRB is adopted to detect the image features at different scales, while we concatenate the outputs of four dilated convolution in DRB, which can simultaneously distill features with different receptive fields and different scales features.

4) Difference to MCERN [57]: MCERN [57] proposes a multi-context block and enhanced reconstruction network for SISR in a coarse-to-fine manner. There are some differences between MCERN [57] and our TSAN. First, MCERN [57] only focuses on how to extract rich contextual information, while our TSAN considers informative features based on MCERN [57]. Second, MCERN [57] only utilize local information for SISR, while TSAN can guarantee local patterns and global diagram with CSB.

5) Difference to SAN [26]: SAN [26] introduces second-order attention operations to learn feature inter-dependencies by global covariance pooling for more discriminative representations in image super-resolution. The main differences between SAN [26] and our TSAN lie in the following aspects. First, SAN [26] pays attention to make full use of the information from the original LR images, while our TSAN values the features information of the LR space and the HR space, and processes the obtained features information in a coarse to fine manner. Second, SAN [26] presents a non-locally enhanced residual group structure based on [58] to capture long-distance contextual information. While we propose a simple and flexible CSB to exploit long-range feature correlation, and we use the triplet structure to capture cross-dimension interaction between the (C, H), (C, W), and (H, W) dimensions.

IV. EXPERIMENT

To verify the effectiveness of the proposed method, we evaluate our SR results with two metrics, *i.e.*, peak signal-to-noise ratio (PSNR) (unit: dB) and structural similarity (SSIM) [59] on Y channel of transformed YCbCr space. For the convenience of a fair comparison, we follow the experiment setup of existing methods. To specify, we use the high-quality DIV2K [55] dataset for training and take four test sets - Set5 [60], Set14 [61], BSDS100 [62], and Urban100 [7] for evaluation.

(a) Original (b) Noisy (c) TSAN

Fig. 4. Qualitative comparisons of color image denoising. The second column shows the noisy images with noise level 25. TSAN recovers fine local details, which is mainly contributed by the abundant hierarchical and contextual features extracted by our proposed DRB. Best viewed in zoom in.

TABLE I
DRB ARCHITECTURE ANALYSIS WITH $\times 2$ SCALE FACTOR. "CASCADED" AND "PARALLEL" DENOTE THE CASCADED ARCHITECTURE AND THE PARALLEL ARCHITECTURE, RESPECTIVELY. $s = 1$ MEANS THE DILATION RATE OF ALL DILATED CONVOLUTIONS IN THE DRB IS 1.

	Set5 [60] PSNR/SSIM	Set14 [61] PSNR/SSIM	BSDS100 [62] PSNR/SSIM	Urban100 [7] PSNR/SSIM
Cascaded	38.14/0.9611	33.78/0.9191	32.27/0.9009	32.50/0.9322
Parallel	38.10/0.9609	33.79/0.9189	32.25/0.9004	32.44/0.9311
DRB(s = 1)	38.17/0.9609	33.80/0.9186	32.27/0.9006	32.54/0.9328
DRB	**38.22/0.9613**	**33.84/0.9196**	**32.32/0.9015**	**32.77/0.9345**

A. Effectiveness of Key Components in MCAB

MCAB contains two branches, *i.e.*, the contextual feature extraction branch and the attention branch.

1) Effectiveness of the Contextual Feature Extraction Branch: The core of this branch is DRB. To verify the effectiveness of the DRB structure, we make three sets of experiments. In the first set of experiments, we replace our DRB structure with the cascaded structure (see Figure 3(a)) or the parallel structure (see Figure 3(b)) and evaluate their performances on the four datasets. For a fair comparison, we set $d = 3$, $e = 6$, and $s = 1$ (that means the dilation rate of all dilated convolutions in the DRB is 1, there are 6 DRBs in each MCAB, and there are 3 MCABs in the whole network). Note that the numbers of these three network parameters are the same. The results are summarized in Table I, from which we can claim that the proposed DRB structure performs best under the same parameters. Such an observation demonstrates that the multiple branches structure in the DRB is more efficient than cascading and parallel structures. This is because DRB can incorporate the advantages of both the cascading and parallel strategies to simultaneously distill features with different receptive fields and different context characteristics.

To demonstrate the importance of different dilated convolution with the dilation rate, we conduct a second set of experiments to compare with a variant structure in which we set the dilation rate s of all dilated convolutions in the DRB to 1 (DRB(s=1)). Similarly, we set $d = 3$ and $e = 6$. The PSNR and SSIM in Table I show that our setting of dilation rate achieves better results on test datasets. This suggests applying different dilated convolution with the dilation rate can obtain a larger receptive field for SISR.

In order to further verify the validity of our proposed DRB structure, we use our network at LR-stage for other low-level computer vision tasks. We provide the results of image denoising in Figure 4. Apparently, our proposed TSAN produces a good result on image denoising because our DRB structure is able to extract abundant hierarchical and contextual features for image reconstruction.

TABLE II
THE ABLATION STUDY OF MCAB COMPONENTS. THE RESULTS ARE EVALUATED ON SET5 [60] FOR A SCALE FACTOR OF $\times 3$.

	w/o CSB	w/o 1^{st}-order triplet	1^{st}-order triplet_HW	w/o 2^{nd}-order triplet	2^{nd}-order triplet_HW	TSAN
PSNR	34.55	34.57	34.60	34.51	34.57	34.64

The above three experiments demonstrate that our proposed DRB is an effective structure that can distill features with different contextual characteristics by two branches and extract features in different receptive fields by two cascaded convolution layers of each branch. And multiple DRBs are combined to integrate contextual features of different levels adaptively.

2) Effectiveness of the Attention Branch: the attention branch consists of three important components, including CSB, 1^{st}-order triplet, and 2^{nd}-order triplet. To verify the effectiveness of different components, we compare TSAN without using CSB, 1^{st}-order triplet, and 2^{nd}-order triplet in Table II. Further, to demonstrate the importance of inter-dependencies between the channel dimension and the spatial dimension, we remove cross-dimension interaction between the (C, H) and (C, W) dimensions in 1^{st}-order triplet and 2^{nd}-order triplet, and only retain the interaction of (H, W) dimension, which denoted as 1^{st}-order triplet_HW, and 2^{nd}-order triplet_HW. It can be found that the CSB contributes to performance improvement. This is mainly because CSB provides local and non-local information to the network, capturing short-distance and long-distance features simultaneously. We can also learn that 1^{st}-order and 2^{nd}-order triplet components contribute to the network ability obviously. This indicates discriminative learning with cross-dimension interaction plays an important role in determining the performance.

B. Effectiveness with Different Number of MCABs

We set different numbers of MCABs in our proposed TSAN and evaluate the performances on different datasets. As shown in Table III, the values of both PSNR and SSIM for our network get better as the number of MCABs increases. Such an observation is consistent with what we expect since the generalization ability will also increase when the number of parameters of our network will go up. As a trade-off between the performance and the complexity of the network, we determine to use three MCABs, which provides strong reconstruction ability and requires not many parameters ($<$ 5.0M).

C. Effectiveness of RAB

To verify that RAB can further improve the reconstruction effect, we conduct a group of experiments without RAB. For a fair comparison, we move the RAB to the front of the upsampling to ensure that the depth of TSAN w/o and w/ RAB are the same. The results are summarized in Table IV. From the results, we can clearly observe that the performance with RAB works better than that without RAB, which suggests

TABLE III
MCABS ANALYSIS BY VARYING THE NUMBER OF MCABS IN TSAN. THE SCALE FACTOR IS $\times 2$.

number of MCABs	Set5 [60] PSNR/SSIM	Set14 [61] PSNR/SSIM	BSDS100 [62] PSNR/SSIM	Urban100 [7] PSNR/SSIM
1	38.01/0.9604	33.55/0.9168	32.13/0.8992	31.92/0.9261
2	38.12/0.9609	33.74/0.9188	32.23/0.9003	32.27/0.9296
3	38.22/0.9613	33.84/0.9196	32.32/0.9015	32.77/0.9345
4	**38.23/0.9614**	**33.87/0.9198**	**32.35/0.9016**	**32.82/0.9348**

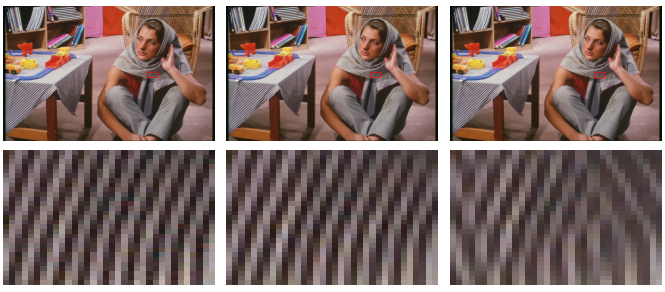
TABLE IV
THE EFFECTIVENESS OF RAB. FOR A FAIR COMPARISON, WE MOVE THE RAB TO THE FRONT OF THE UPSAMPLING TO ENSURE THAT THE DEPTH OF TSAN W/O AND W/ RAB ARE THE SAME. THE RESULTS EVALUATED FOR A SCALE FACTOR OF $\times 2$.

TSAN	Set5 [60] PSNR/SSIM	Set14 [61] PSNR/SSIM	BSDS100 [62] PSNR/SSIM	Urban100 [7] PSNR/SSIM
w/o RAB	38.14/0.9611	33.75/0.9192	32.29/0.9010	32.58/0.9327
w/ RAB†	38.10/0.9610	33.79/0.9193	32.26/0.9007	32.49/0.9320
w/ RAB	**38.22/0.9613**	**33.84/0.9196**	**32.32/0.9015**	**32.77/0.9345**

that our RAB is able to refine a coarse HR result to a more detailed one since it can continue to extract useful features from HR space. We also visualize the visual effects of TSAN w/o and w/ RAB in Figure 5. In this example, our RAB is able to correct the direction for black lines. In order to further prove the effectiveness of our coarse-to-fine method, we designed a variant w/ RAB† that sets the parameters w_1 in the objective function to 0 and w_2 to 1. From Table IV, it can be found that w_1 equal to zero will have a negative impact on the reconstruction results. The reason behind this is that the lack of intermediate coarse results reconstructed from the LR space makes the training process difficult. This demonstrates that multiple MCABs can extract rich attentive contextual features with cross-dimension interaction to obtain a good initial coarse SR result, then our proposed RAB can further improve the coarse SR image into a fine SR image by using LR and HR space information simultaneously.

D. Comparisons with State-of-the-art Methods

We compare our proposed TSAN with eight state-of-the-art light-weighted methods (with $<$ 6M parameters): LapSRN [45], CARN [42], SRMDNF [63], NLRN [64], MSRN [41], FRSR [3], OISR-RK2 [4], and LattienNet [65]. We summarize the quantitative comparisons for $\times 2$, $\times 3$, and $\times 4$ in Table V. As we can see, our TSAN achieves excellent performance on different datasets and different upsampling scales. We also visualize several examples with different upsampling scales in Figures 7, 8, and 9. Obviously, the SR images generated by other methods exhibit visible artifacts, while our proposed TSAN is able to generate a more visually pleasant image with clean details and sharp edges. This can be explained by the fact that rich attentive contextual features with long-range dependencies extracted by multiple MCABs



(a) GT (b) w/ RAB (c) w/o RAB

Fig. 5. Visual comparisons of the effectiveness of RAB on *barbara* from Set14 [61] with scale factor $\times 2$.

TABLE V
PERFORMANCE COMPARISON TO OTHER 8 STATE-OF-THE-ART METHODS WITH THE LIGHT-WEIGHTED MODEL (< 6.0M PARAMETERS). THE BEST AND SECOND BEST RESULTS ARE HIGHLIGHTED IN **BOLD** AND UNDERLINED.

Scale	Method	Set5 [60] PSNR/SSIM	Set14 [61] PSNR/SSIM	BSDS100 [62] PSNR/SSIM	Urban100 [7] PSNR/SSIM
×2	Bicubic	33.66/0.9299	30.24/0.8688	29.56/0.8431	26.88/0.8403
	LapSRN [45] (CVPR'17)	37.52/0.9591	33.08/0.9130	31.80/0.8950	30.41/0.9101
	CARN [42] (ECCV'18)	37.76/0.9590	33.52/0.9166	32.09/0.8978	31.51/0.9312
	SRMDNF [63] (CVPR'18)	37.79/0.9601	33.32/0.9159	32.05/0.8985	31.33/0.9204
	NLRN [64] (NIPS'18)	38.00/0.9603	33.46/0.9195	32.19/0.8992	31.82/0.9249
	MSRN [41] (ECCV'18)	38.08/0.9605	33.74/0.9170	32.23/0.9013	32.22/0.9326
	FRSR [3] (CVPR'19)	37.95v0.9594	33.45/0.9195	32.17/0.8991	32.23/0.9290
	OISR-RK2 [4] (CVPR'19)	38.11/0.9609	33.80/0.9193	32.26/0.9006	32.48/0.9317
	LattienNet [65] (ECCV'20)	38.15/0.9610	33.78/0.9193	32.25/0.9005	32.43/0.9302
	TSAN	**38.22/0.9613**	**33.84/0.9196**	**32.32/0.9015**	**32.77/0.9345**
×3	Bicubic	30.39/0.8682	27.55/0.7742	27.21/0.7385	24.46/0.7349
	LapSRN [45] (CVPR'17)	33.82/0.9227	29.87/0.8320	28.82/0.7980	27.07/0.8280
	CARN [42] (ECCV'18)	34.29/0.9255	30.29/0.8407	29.06/0.8034	27.38/0.8404
	SRMDNF [63] (CVPR'18)	34.12/0.9254	30.04/0.8382	28.97/0.8025	27.57/0.8398
	NLRN [64] (NIPS'18)	34.27/0.9266	30.16/0.8374	29.06/0.8026	27.93/0.8453
	MSRN [41] (ECCV'18)	34.38/0.9262	30.34/0.8395	29.08/0.8041	28.08/0.8554
	FRSR [3] (CVPR'19)	34.38/0.9262	30.27/0.8411	29.11/0.8050	28.33/0.8556
	OISR-RK2 [4] (CVPR'19)	34.55/0.9281	30.46/0.8443	29.18/0.8075	28.50/0.8597
	LattienNet [65] (ECCV'20)	34.53/0.9281	30.39/0.8424	29.15/0.8059	28.33/0.8538
	TSAN	**34.64/0.9282**	**30.52/0.8454**	**29.20/0.8080**	**28.55/0.8602**
×4	Bicubic	28.42/0.8104	26.00/0.7027	25.96/0.6675	23.14/0.6577
	LapSRN [45] (CVPR'17)	31.54/0.8850	28.19/0.7720	27.32/0.7270	25.21/0.7560
	CARN [42] (ECCV'18)	31.92/0.8903	28.42/0.7762	27.44/0.7304	25.63/0.7688
	SRMDNF [63] (CVPR'18)	31.96/0.8925	28.35/0.7787	27.49/0.7337	25.68/0.7731
	NLRN [64] (NIPS'18)	31.92/0.8916	28.36/0.7745	27.48/0.7306	25.79/0.7729
	MSRN [41] (ECCV'18)	32.07/0.8903	28.60/0.7751	27.52/0.7273	26.04/0.7896
	FRSR [3] (CVPR'19)	32.22/0.8950	28.64/0.7830	27.60/0.7370	26.21/0.7910
	OISR-RK2 [4] (CVPR'19)	32.35/0.8970	28.72/0.7843	27.66/0.7390	26.37/0.7953
	LattienNet [65] (ECCV'20)	32.30/0.8962	28.68/0.7830	27.62/0.7367	26.25/0.7873
	TSAN	**32.40/0.8975**	**28.73/0.7847**	**27.67/0.7398**	**26.39/0.7955**

TABLE VI
PERFORMANCE COMPARISON TO 8 STATE-OF-THE-ART METHODS WITH THE HEAVY-WEIGHTED MODEL. THE BEST AND SECOND BEST RESULTS ARE HIGHLIGHTED IN **BOLD** AND UNDERLINED.

Scales	Methods	Set5 [60] PSNR/SSIM	Set14 [61] PSNR/SSIM	BSDS100 [62] PSNR/SSIM	Urban100 [7] PSNR/SSIM
×2	EDSR [1] (CVPRW'17)	38.11/0.9602	33.92/0.9195	32.32/0.9013	32.93/0.9351
	RDN [2] (CVPR'18)	38.24/0.9614	34.01/0.9212	32.34/0.9017	32.89/0.9353
	DBPN [46] (CVPR'18)	38.09/0.9600	33.85/0.9190	32.27/0.9000	32.55/0.9324
	RCAN [6] (ECCV'18)	38.27/0.9614	34.12/0.9216	32.40/0.9025	33.34/0.9384
	RNAN [66] (ICLR'19)	38.17/0.9611	33.87/0.9207	32.32/0.9014	32.73/0.9340
	SAN [26] (CVPR'19)	38.28/0.9618	34.07/0.9213	32.35/0.9019	33.10/0.9370
	Pan [67] (AAAI'20)	38.26/0.9614	33.99/0.9200	32.37/0.9020	33.09/0.9365
	HAN [68] (ECCV'20)	38.27/0.9614	34.16/0.9217	**32.41/0.9027**	33.35/0.9385
	TSAN	38.22/0.9613	33.84/0.9196	32.32/0.9015	32.77/0.9345
	TSAN-L	**38.30/0.9619**	**34.17/0.9218**	32.40/0.9026	**33.45/0.9387**
×3	EDSR [1] (CVPRW'17)	34.65/0.9280	30.52/0.8462	29.25/0.8093	28.80/0.8653
	RDN [2] (CVPR'18)	34.71/0.9296	30.57/0.8468	29.26/0.8093	28.80/0.8653
	DBPN [46] (CVPR'18)	–/–	–/–	–/–	–/–
	RCAN [6] (ECCV'18)	34.74/0.9299	30.65/0.8482	29.32/0.8111	29.09/0.8702
	RNAN [66] (ICLR'19)	34.65/0.9288	30.55/0.8465	29.25/0.8089	28.74/0.8645
	SAN [26] (CVPR'19)	34.75/0.9300	30.59/0.8476	29.33/0.8112	28.93/0.8671
	Pan [67] (AAAI'20)	34.75/0.9298	30.61/0.8466	29.29/0.8102	28.97/0.8683
	HAN [68] (ECCV'20)	34.75/0.9299	**30.67**/0.8483	29.32/0.8110	29.10/0.8705
	TSAN	34.64/0.9282	30.52/0.8454	29.20/0.8080	28.55/0.8602
	TSAN-L	**34.80/0.9301**	30.65/**0.8486**	**29.34/0.8114**	**29.17/0.8720**
×4	EDSR [1] (CVPRW'17)	32.46/0.8968	28.80/0.7876	27.71/0.7420	26.64/0.8033
	RDN [2] (CVPR'18)	32.47/0.8990	28.81/0.7871	27.72/0.7417	26.61/0.8028
	DBPN [46] (CVPR'18)	32.47/0.8980	28.82/0.7860	27.72/0.7400	26.38/0.7946
	RCAN [6] (ECCV'18)	32.63/0.9002	28.87/0.7889	27.77/0.7436	26.82/0.8087
	RNAN [66] (ICLR'19)	32.49/0.8982	28.83/0.7878	27.72/0.7421	26.61/0.8023
	SAN [26] (CVPR'19)	32.64/0.9003	**28.92**/0.7888	27.78/0.7436	26.79/0.8086
	Pan [67] (AAAI'20)	32.56/0.8995	28.80/0.7882	27.73/0.7422	26.72/0.8053
	HAN [68] (ECCV'20)	32.64/0.9002	28.90/**0.7890**	27.80/0.7442	26.85/0.8094
	TSAN	32.40/0.8975	28.73/0.7847	27.67/0.7398	26.39/0.7955
	TSAN-L	**32.65/0.9004**	28.91/0.7888	**27.81/0.7443**	**26.95/0.8110**

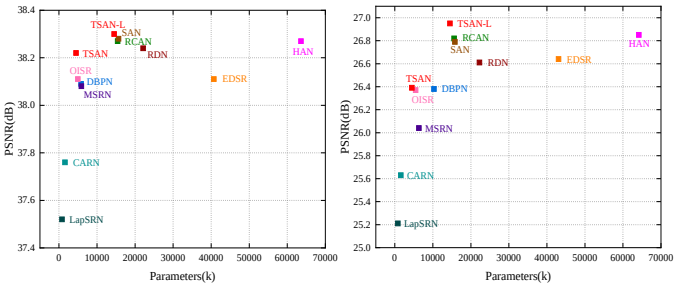


Fig. 6. PSNR performance versus number of parameters. (a) The results are evaluated on Set5 for a scale factor of $\times 2$. (b) The results are evaluated on Urban100 for a scale factor of $\times 4$. Our TSAN and TSAN-L has a better tradeoff between performance and model size.

in the LR space ensure a good initial coarse SR result, and then the SR result can be further improved by RAB.

To further prove the effectiveness of the proposed model, we increase the number of MCAB to 13 (denoted as TSAN-L) to fairly compare with some methods with large parameters or heavy computations. We compare TSAN and TSAN-L with eight current state-of-the-art heavy-weight methods in Table VI, *i.e.*, EDSR [1], RDN [2], DBPN [46], RCAN [6], RNAN [66], SAN [26], Pan [67], and HAN [68]. We can see that our proposed TSAN still achieves comparable performance, and TSAN-L performs favorably against the state-of-the-art methods. For example, the proposed TSAN gains 0.11dB higher than EDSR [1] on Set5 [60] for $\times 2$ scale, and the TSAN-L gains 0.13dB, 0.16dB, and 0.10dB higher than state-of-the-art methods RCAN [6], SAN [26], and Pan [67] on Urban100 [7] for $\times 4$ scale, respectively. This observation suggests that our TSAN with cross-dimension interaction can make better use of more informative contextual features to boost reconstruction performance. We also visualize several examples with different upsampling scales in Figures 10, 11, and 12. As shown, most compared SR methods cannot recover the grids of buildings accurately and suffer from unpleasant blurring artifacts. In contrast, our TSAN-L obtains clearer details and reconstructs sharper high-frequency textures. Again, this strongly demonstrates the superiority of our method.

We also compare the tradeoff between the performance and the number of network parameters from our TSAN network and existing methods. Figure 6 shows the PSNR performances of 12 models versus the number of parameters, where the results are evaluated with Set5 [60] and Urban100 [7] datasets for $2\times$ and $4\times$ upscaling factors, respectively. We can find that our TSAN and TSAN-L network significantly outperforms the relatively small models across all datasets and scales. Moreover, our TSAN-L network performs better than EDSR [1] and RDN [2] across two scales but with about 65% and 34% fewer parameters on average, respectively. Furthermore, compared with RCAN [6] and SAN [26] on two upscaling factors, our TSAN-L has fewer parameters and achieves higher PSNR. We further compute the FLOPs and provide the speed by assuming that the size of LR image is 48×48 and the scale factor is 2. For a fair comparison, all methods are tested on the same CPU. From Table VII, we can see that our network has fewer FLOPs and faster inference speed than compared approaches. These comparisons indicate that our proposed network has a better tradeoff between performance and model size.

TABLE VII
THE FLOPS AND INFERENCE TIME COMPARISONS OF OUR METHOD WITH FIVE STATE-OF-THE-ART NETWORKS.

	EDSR [1]	MSRN [41]	RCAN [6]	SAN [26]	HAN [68]	TSAN
FLOPs (G)	115.78	13.67	36.67	30.04	150.99	10.11
Time (s)	17.00	2.57	9.56	10.26	26.82	2.34

V. CONCLUSION

In this paper, we propose a novel and light-weighted TSAN for SISR in a coarse-to-fine fashion to utilize the attentive contextual information with cross-dimension interaction and emphasize the reconstruction process on both LR and SR space. The DRB with the well designed compact structure can increase the receptive field and get more contextual features. The MCAB can effectively extract attentive contextual features by exploiting inter-dependencies between different dimensions to obtain the coarse result. Also, an RAB is proposed to focus on extracting essential HR space features after upsampling to refine the coarse result. Extensive evaluations on the benchmark datasets have demonstrated the efficacy of our proposed TSAN in terms of metric accuracy and visual effects.

REFERENCES

[1] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee, "Enhanced deep residual networks for single image super-resolution," in *CVPRW*, 2017.
[2] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *CVPR*, 2018.
[3] J. W. Soh, G. Y. Park, J. Jo, and N. I. Cho, "Natural and realistic single image super-resolution with explicit natural manifold discrimination," in *CVPR*, 2019.
[4] X. He, Z. Mo, P. Wang, Y. Liu, M. Yang, and J. Cheng, "Ode-inspired network design for single image super-resolution," in *CVPR*, 2019.
[5] Y. Zhang, K. Li, K. Li, B. Zhong, and Y. Fu, "Residual non-local attention networks for image restoration," in *ICLR*, 2019.
[6] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *ECCV*, 2018.
[7] J.-B. Huang, A. Singh, and N. Ahuja, "Single image super-resolution from transformed self-exemplars," in *CVPR*, 2015.
[8] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE TPAMI*, 2016.
[9] Z. Li, J. Yang, Z. Liu, X. Yang, G. Jeon, and W. Wu, "Feedback network for image super-resolution," in *CVPR*, 2019.
[10] J. Cai, H. Zeng, H. Yong, Z. Cao, and L. Zhang, "Toward real-world single image super-resolution: A new benchmark and a new model," in *ICCV*, 2019.
[11] X. Yang, H. Mei, J. Zhang, K. Xu, B. Yin, Q. Zhang, and X. Wei, "Drfn: Deep recurrent fusion network for single-image super-resolution with large factors," *IEEE TMM*, 2018.
[12] Z. He, Y. Cao, L. Du, B. Xu, J. Yang, Y. Cao, S. Tang, and Y. Zhuang, "Mrfn: Multi-receptive-field network for fast and accurate single image super-resolution," *IEEE TMM*, 2019.
[13] B. Yan, B. Bare, C. Ma, K. Li, and W. Tan, "Deep objective quality assessment driven single image super-resolution," *IEEE TNM*, 2019.
[14] Q. Ning, W. Dong, G. Shi, L. Li, and X. Li, "Accurate and lightweight image super-resolution with model-guided deep unfolding network," *IEEE J-STSP*, 2020.
[15] D. Song, C. Xu, X. Jia, Y. Chen, C. Xu, and Y. Wang, "Efficient residual dense block search for image super-resolution," in *AAAI*, 2020.
[16] W. Song, S. Choi, S. Jeong, and K. Sohn, "Stereoscopic image super-resolution with stereo consistent feature," in *AAAI*, 2020.

Fig. 7. Visual comparison between our **TSAN** and other light-weighted methods on *img062* from Urban100 [7] and *ppt* from Set14 [61] with scale $\times 2$.

Fig. 8. Visual comparison between our **TSAN** and other light-weighted methods on *img004* from Urban100 [7] and *zebra* from Set14 [61] with scale $\times 3$.

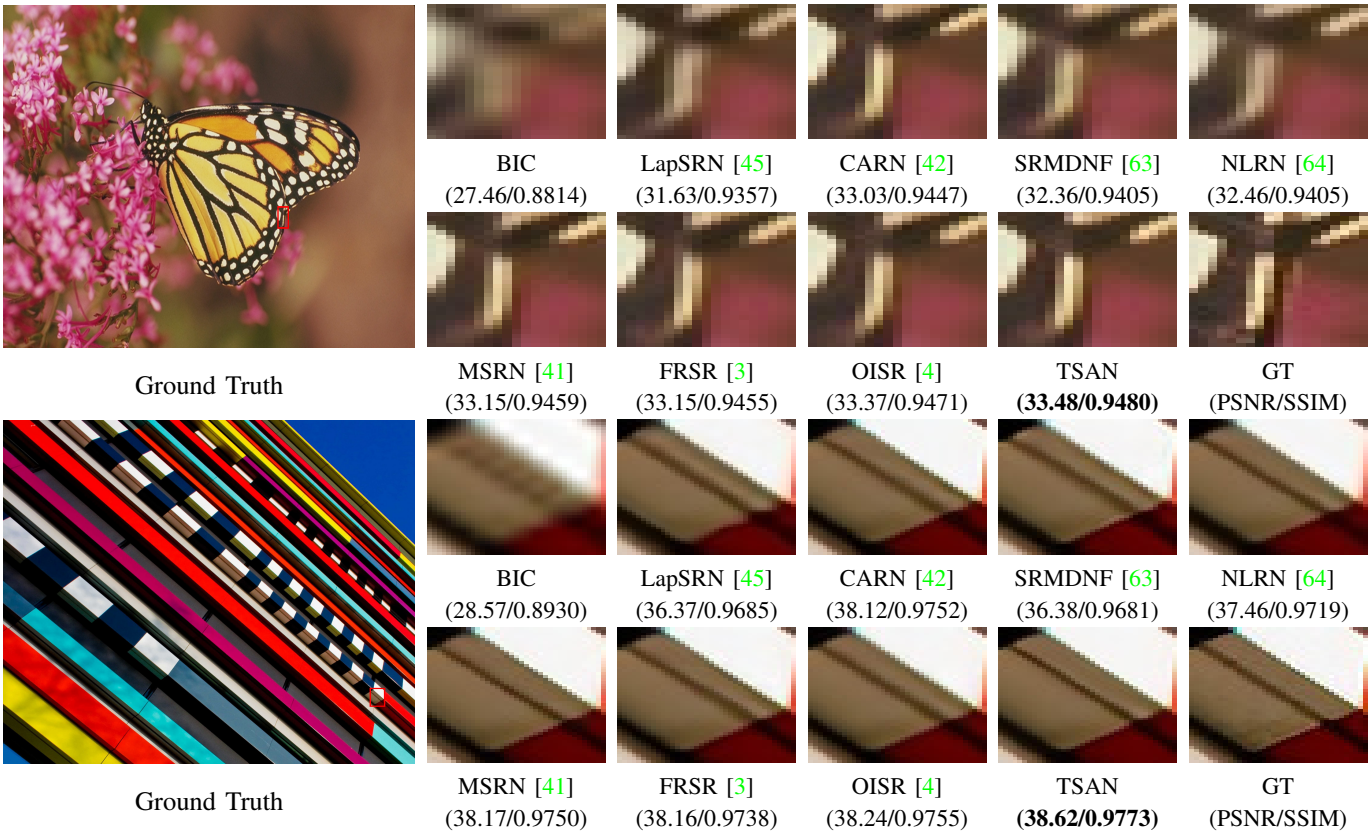


Fig. 9. Visual comparison between our **TSAN** and other light-weighted methods on *monarch* from Set14 [61] and *img081* from Urban100 [7] with scale $\times 4$.

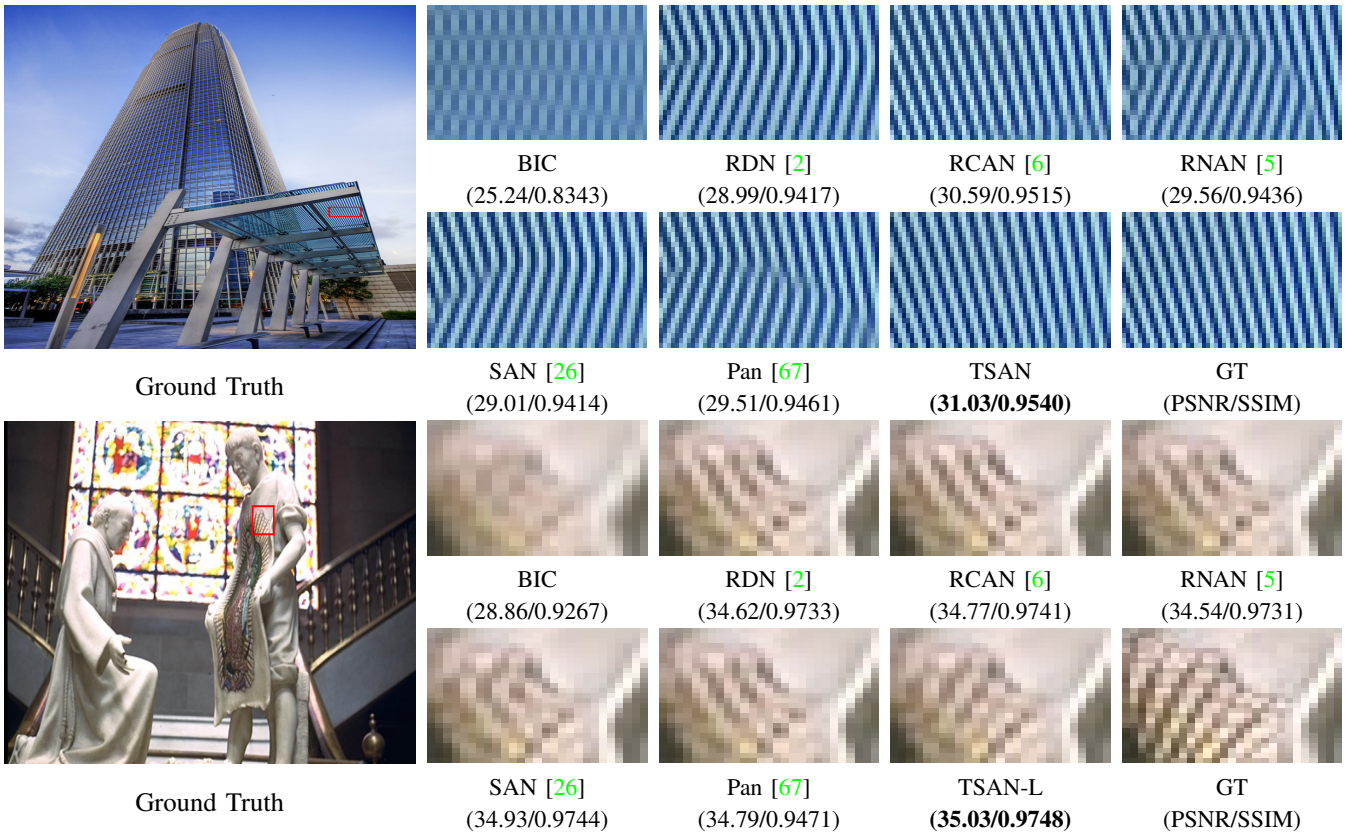


Fig. 10. Visual comparison between our **TSAN-L** and several heavy-weighted methods on *img046* from Urban100 [7] and *24077* from BSDS100 [62] with scale $\times 2$.

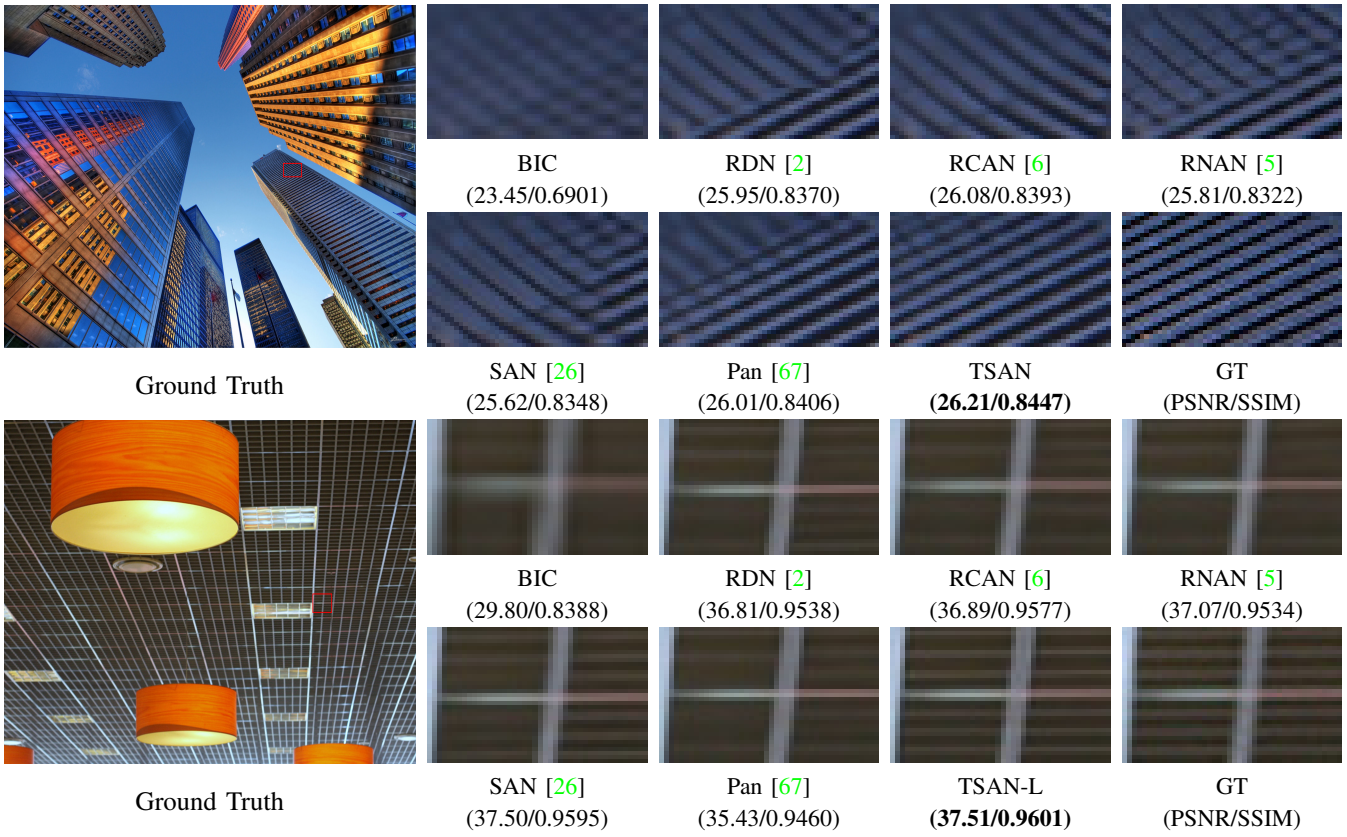


Fig. 11. Visual comparison between our **TSAN-L** and several heavy-weighted methods on *img012* from Urban100 [7] and *img044* from Urban100 [7] with scale $\times 3$.

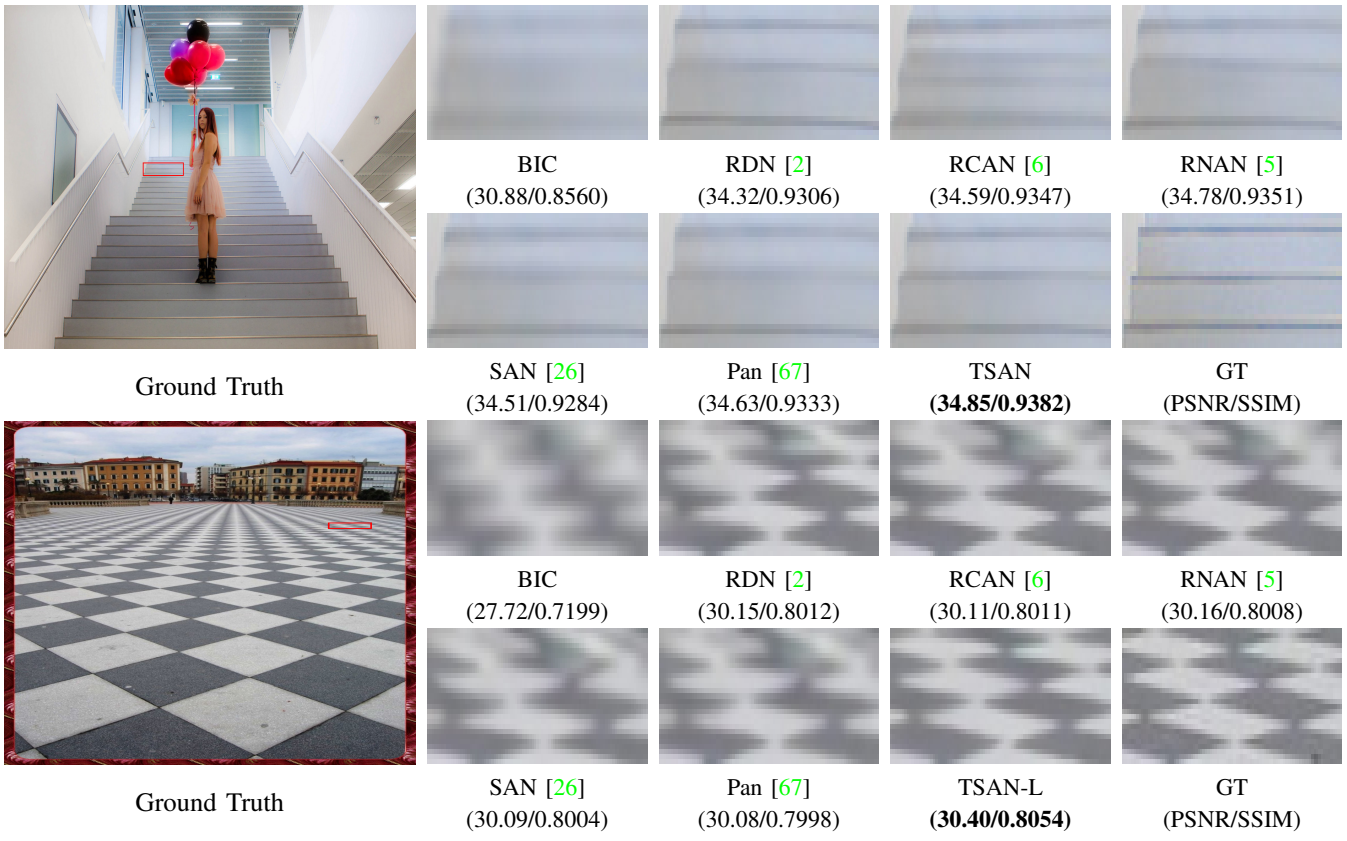


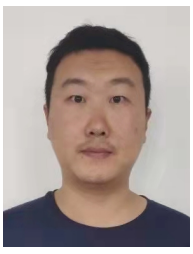
Fig. 12. Visual comparison between our **TSAN-L** and several heavy-weighted methods on *img009* from Urban100 [7] and *img021* from Urban100 [7] with scale $\times 4$.

[17] K. Zhang, L. Van Gool, and R. Timofte, "Deep unfolding network for image super-resolution," in *CVPR*, 2020.
[18] W. Dong, P. Wang, W. Yin, G. Shi, F. Wu, and X. Lu, "Denoising prior driven deep neural network for image restoration," *IEEE TPAMI*, 2018.
[19] J. Yoo, N. Ahn, and K.-A. Sohn, "Rethinking data augmentation for image super-resolution: A comprehensive analysis and a new strategy," in *CVPR*, 2020.
[20] J. W. Soh, S. Cho, and N. I. Cho, "Meta-transfer learning for zero-shot super-resolution," in *CVPR*, 2020.
[21] Y. Hu, J. Li, Y. Huang, and X. Gao, "Channel-wise and spatial feature modulation network for single image super-resolution," *IEEE TCSVT*, 2019.
[22] J.-W. Chang, K.-W. Kang, and S.-J. Kang, "An energy-efficient fpga-based deconvolutional neural networks accelerator for single image super-resolution," *IEEE TCSVT*, 2018.
[23] Y. Wang, L. Wang, H. Wang, and P. Li, "Resolution-aware network for image super-resolution," *IEEE TCSVT*, 2018.
[24] Y. Li, W. Dong, X. Xie, G. Shi, J. Wu, and X. Li, "Image super-resolution with parametric sparse model learning," *IEEE TIP*, 2018.
[25] W. Dong, T. Huang, G. Shi, Y. Ma, and X. Li, "Robust tensor approximation with laplacian scale mixture modeling for multiframe image and video denoising," *IEEE J-STSP*, 2018.
[26] T. Dai, J. Cai, Y. Zhang, S.-T. Xia, and L. Zhang, "Second-order attention network for single image super-resolution," in *CVPR*, 2019.
[27] C. De Boor, "Bicubic spline interpolation," *Journal of Mathematics and Physics*, 1962.
[28] C. E. Duchon, "Lanczos filtering in one and two dimensions," *Journal of applied meteorology*, 1979.
[29] S. Dai, M. Han, W. Xu, Y. Wu, and Y. Gong, "Soft edge smoothness prior for alpha channel super resolution," in *CVPR*, 2007.
[30] R. Fattal, "Image upsampling via imposed edge statistics," *ACM TOG*, 2007.
[31] H. Chang, D.-Y. Yeung, and Y. Xiong, "Super-resolution through neighbor embedding," in *CVPR*, 2004.
[32] K. Xu, X. Wang, X. Yang, S. He, Q. Zhang, B. Yin, X. Wei, and R. W. Lau, "Efficient image super-resolution integration," *The Visual Computer*, 2018.
[33] J. Liu, W. Yang, X. Zhang, and Z. Guo, "Retrieval compensated group structured sparsity for image super-resolution," *IEEE TMM*, 2016.
[34] J. Kim, J. Kwon Lee, and K. Mu Lee, "Accurate image super-resolution using very deep convolutional networks," in *CVPR*, 2016.
[35] J. Kim, J. Kwon Lee, and K. Mu Lee, "Deeply recursive convolutional network for image super-resolution," in *CVPR*, 2016.
[36] Y. Tai, J. Yang, and X. Liu, "Image super-resolution via deep recursive residual network," in *CVPR*, 2017.
[37] Y. Tai, J. Yang, X. Liu, and C. Xu, "Memnet: A persistent memory network for image restoration," in *ICCV*, 2017.
[38] C. Dong, C. C. Loy, and X. Tang, "Accelerating the super-resolution convolutional neural network," in *ECCV*, 2016.
[39] Z. Hui, X. Wang, and X. Gao, "Fast and accurate single image super-resolution via information distillation network," in *CVPR*, 2018.
[40] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *CVPR*, 2017.
[41] J. Li, F. Fang, K. Mei, and G. Zhang, "Multi-scale residual network for image super-resolution," in *ECCV*, 2018.
[42] N. Ahn, B. Kang, and K.-A. Sohn, "Fast, accurate, and lightweight super-resolution with cascading residual network," in *ECCV*, 2018.
[43] C. Xie, W. Zeng, and X. Lu, "Fast single-image super-resolution via deep network with component learning," *IEEE TCSVT*, 2018.
[44] J. Lei, Z. Zhang, X. Fan, B. Yang, X. Li, Y. Chen, and Q. Huang, "Deep stereoscopic image super-resolution via interaction module," *IEEE TCSVT*, 2020.
[45] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Deep laplacian pyramid networks for fast and accurate super-resolution," in *CVPR*, 2017.
[46] M. Haris, G. Shakhnarovich, and N. Ukita, "Deep back-projection networks for super-resolution," in *CVPR*, 2018.
[47] Y. Qiu, R. Wang, D. Tao, and J. Cheng, "Embedded block residual network: A recursive restoration model for single-image super-resolution," in *ICCV*, 2019.
[48] B. Zhao, X. Wu, J. Feng, Q. Peng, and S. Yan, "Diversified visual attention networks for fine-grained object classification," *IEEE TMM*, 2017.
[49] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *CVPR*, 2019.
[50] R. R. A. Pramono, Y.-T. Chen, and W.-H. Fang, "Hierarchical self-attention network for action localization in videos," in *ICCV*, 2019.
[51] B. Chen, W. Deng, and J. Hu, "Mixed high-order attention network for person re-identification," in *ICCV*, 2019.
[52] C. Du, H. Zewei, S. Anshun, Y. Jiangxin, C. Yanlong, C. Yanpeng, T. Siliang, and M. Ying Yang, "Orientation-aware deep neural network for real image super-resolution," in *CVPRW*, 2019.
[53] H. Wu, Z. Zou, J. Gui, W.-J. Zeng, J. Ye, J. Zhang, H. Liu, and Z. Wei, "Multi-grained attention networks for single image super-resolution," *IEEE TCSVT*, 2020.
[54] P. Li, J. Xie, Q. Wang, and W. Zuo, "Is second-order information helpful for large-scale visual recognition?" in *CVPR*, 2017.
[55] R. Timofte, E. Agustsson, L. Van Gool, M.-H. Yang, and L. Zhang, "Ntire 2017 challenge on single image super-resolution: Methods and results," in *CVPRW*, 2017.
[56] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.
[57] J. Zhang, C. Long, Y. Wang, X. Yang, H. Mei, and B. Yin, "Multi-context and enhanced reconstruction network for single image super resolution," in *ICME*, 2020.
[58] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *CVPR*, 2018.
[59] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli *et al.*, "Image quality assessment: from error visibility to structural similarity," *IEEE TIP*, 2004.
[60] M. Bevilacqua, A. Roumy, C. Guillemot, and M. L. Alberi-Morel, "Low-complexity single-image super-resolution based on nonnegative neighbor embedding," in *BMVC*, 2012.
[61] R. Zeyde, M. Elad, and M. Protter, "On single image scale-up using sparse-representations," in *ICCS*, 2010.
[62] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik, "Contour detection and hierarchical image segmentation," *IEEE TPAMI*, 2011.
[63] K. Zhang, W. Zuo, and L. Zhang, "Learning a single convolutional super-resolution network for multiple degradations," in *CVPR*, 2018.
[64] D. Liu, B. Wen, Y. Fan, C. C. Loy, and T. S. Huang, "Non-local recurrent network for image restoration," in *NIPS*, 2018.
[65] X. Luo, Y. Xie, Y. Zhang, Y. Qu, C. Li, and Y. Fu, "Latticenet: Towards lightweight image super-resolution with lattice block," in *ECCV*, 2020.
[66] Y. Zhang, K. Li, K. Li, B. Zhong, and Y. Fu, "Residual non-local attention networks for image restoration," in *ICLR*, 2019.
[67] J. Pan, Y. Liu, D. Sun, J. Ren, M.-M. Cheng, J. Yang, and J. Tang, "Image formation model guided deep image super-resolution," in *AAAI*, 2020.
[68] B. Niu, W. Wen, W. Ren, X. Zhang, L. Yang, S. Wang, K. Zhang, X. Cao, and H. Shen, "Single image super-resolution via a holistic attention network," in *ECCV*, 2020.

Jiqing Zhang received the B.Eng. degree in computer science and technology from Dalian Maritime University, Dalian, China, in 2017. He is currently pursuing the Ph.D. degree in computer application technology at Dalian University of Technology, Dalian, China. His research interests include computer vision, machine learning, and image processing.

Chengjiang Long is currently a Principal Scientist in JD Finance America Corporation (a part of JD.COM) since June 2020. Prior to working at JD.COM, he worked as a Computer Vision Researcher/Senior R&D Engineer at Kitware from February 2016 to April 2020. He also worked as an Adjunct Professor at University at Albany, SUNY from August 2018 to May 2020, and was an Adjunct Professor at Rensselaer Polytechnic Institute (RPI) from Jan 2018 to May 2018. He received the M.S. degree in Computer Science from Wuhan University in 2011 and a B.S degree in Computer Science and Technology from Wuhan University in 2009. He got his Ph.D. degree in Computer Science from Stevens Institute of Technology in 2015. During his Ph.D. study, he worked at NEC Labs America and GE Global Research as a research intern in 2013 and 2015, respectively. To date, he has published 50 papers including top journals such T-PAMI, IJCV, T-IP and T-MM, top international conferences such as CVPR, ICCV, and AAAI, and owns 1 patent. He is also the reviewer for more than 20 top international journals and conferences. His research interests involve various areas of computer vision, computer graphics, multimedia, machine learning, and artificial intelligence. He is a member of IEEE and AAAI.

Yuxin Wang is an associate professor and master tutor of Dalian University of Technology. His main research interests include parallel and distributed computing, big data analysis and application.



Haiyin Piao is currently working toward the Dr.Eng. degree in electronics and information at Northwestern Polytechnical University (NWPU), Xian, China. He is currently also a Senior Engineer at Artificial Intelligence Laboratory, SADRI Institute, Shenyang, China. His research interests include multi-agent reinforcement learning, game theory with particular attention to aerospace applications.

Haiyang Mei is a PH.D. student in the Department of Computer Science at Dalian University of Technology, China. His research interests include computer vision, machine learning, and image processing.

Xin Yang is a Professor in the Department of Computer Science at Dalian University of Technology, China. Yang received his B.S. degree in Computer Science from Jilin University in 2007. From 2007 to June 2012, he was a joint Ph.D. student at Zhejiang University and UC Davis for Graphics and received his Ph.D. degree in July 2012. His research interests include computer graphics and robotic vision.

Baocai Yin is a Professor of Computer Science at Dalian University of Technology and the Dean of the Faculty of Electronic Information and Electrical Engineering. His research concentrates on digital multimedia and computer vision. He received his B.S. degree and Ph.D. degree in Computer Science, each from Dalian University of Technology.