Second-order Attention Network for Single Image Super-Resolution

Tao Dai[1,2,*,‡], Jianrui Cai[3,*], Yongbing Zhang[1], Shu-Tao Xia[1,2], Lei Zhang[3,4,§]
[1]Graduate School at Shenzhen, Tsinghua University, Shenzhen, China
[2] PCL Research Center of Networks and Communications, Peng Cheng Laboratory, Shenzhen, China
[3]Department of Computing, The Hong Kong Polytechnic University, Hong Kong, China
[4]DAMO Academy, Alibaba Group
{dait14, zhang.yongbing, xiast}@sz.tsinghua.edu.cn, {csjcai, cslzhang}@comp.polyu.edu.hk

Abstract

Recently, deep convolutional neural networks (CNNs) have been widely explored in single image super-resolution (SISR) and obtained remarkable performance. However, most of the existing CNN-based SISR methods mainly focus on wider or deeper architecture design, neglecting to explore the feature correlations of intermediate layers, hence hindering the representational power of CNNs. To address this issue, in this paper, we propose a second-order attention network (SAN) for more powerful feature expression and feature correlation learning. Specifically, a novel trainable second-order channel attention (SOCA) module is developed to adaptively rescale the channel-wise features by using second-order feature statistics for more discriminative representations. Furthermore, we present a non-locally enhanced residual group (NLRG) structure, which not only incorporates non-local operations to capture long-distance spatial contextual information, but also contains repeated local-source residual attention groups (LSRAG) to learn increasingly abstract feature representations. Experimental results demonstrate the superiority of our SAN network over state-of-the-art SISR methods in terms of both quantitative metrics and visual quality.

1. Introduction

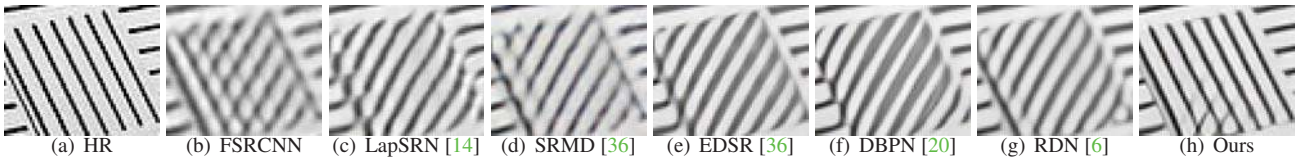
Single image super-resolution (SISR) [5] has recently received much attention. In general, the purpose of SISR is to produce a visually high-resolution (HR) output from its low-resolution (LR) input. However, this inverse problem is ill-posed since multiple HR solutions can map to any LR input. Therefore, a great number of SR methods have been proposed, ranging from early interpolation-based [37] and model-based [4], to recent learning-based methods [32, 39].

The early developed interpolated-based methods (*e.g.*, bilinear and bicubic methods) are simple and efficient but limited in applications. For more flexible SR methods, more advanced model-based methods are proposed by exploiting powerful image priors, such as non-local similarity prior [34] and sparsity prior [4]. Although such model-based methods are flexible to produce relative high-quality HR images, they still suffer from some drawbacks: (1) such methods often involve a time-consuming optimization process; (2) the performance may degrade quickly when image statistics are biased from the image prior.

Deep convolution neural networks (CNNs) have recently achieved unprecedented success in various problems [7, 25]. The powerful feature representation and end-to-end training paradigm of CNN makes it a promising approach to SISR. In the last several years, a flurry of CNN-based SISR methods have been proposed to learn a mapping function from an interpolated or LR input to its corresponding HR output. By fully exploiting the image statics inherent in training datasets, CNNs have achieved state-of-the-art results in SISR [2, 12, 14, 36, 39, 38]. Although considerable progress has been achieved in image SR, existing CNN-based SR models are still faced with some limitations: (1) most of CNN-based SR methods do not make full use of the information from the original LR images, thereby resulting in relatively-low performance; (2) most existing CNN-based SR models focus mainly on designing a deeper or wider network to learn more discriminative high-level features, while rarely exploiting the inherent feature correlations in intermediate layers, thus hindering the representational ability of CNNs.

To address these problems, we propose a deep second-order attention network (SAN) for more powerful feature expression and feature correlation learning. Specifically, we

*The first two authors contribute equally to this work.
§Corresponding author: Lei Zhang
‡This work is supported in part by the National Natural Science Foundation of China under Grant 61771273, the R&D Program of Shenzhen under Grant JCYJ20180508152204044, and the research fund of PCL Future Regional Network Facilities for Large-scale Experiments and Applications (PCL2018KP001).



(a) HR (b) FSRCNN (c) LapSRN [14] (d) SRMD [36] (e) EDSR [36] (f) DBPN [20] (g) RDN [6] (h) Ours

Figure 1. Zoom visual results for $4\times$ SR on "img_092" from Urban100. Our method obtains better visual quality and recovers more image details compared with other state-of-the-art SR methods

propose a second-order channel attention (SOCA) mechanism for better feature correlation learning. Our SOCA adaptively learns feature inter-dependencies by exploiting second-order feature statistics instead of first-order ones. Such SOCA mechnism makes our network focus on more informative feature and improve discriminative learning ability. Moreover, a non-locally enhanced residual group (NLRG) structure is presented to further incorporates non-local operations to capture long-distance spatial contextual information. By stacking the local-source residual attention groups (LSRAG) structure, we can exploit the information from the LR images and allow the abundant low-frequency information to be bypassed. As shown in Fig. 1, our method obtains better visual quality and recovers more image details compared with other state-of-the-art SR methods.

In summary, the main contributions of this paper are listed as follows:

- We propose a deep second-order attention network (SAN) for accurate image SR. Extensive experiments on public datasets demonstrate the superiority of our SAN over state-of-the-art methods in terms of both quantitive and visual quality.
- We propose second-order channel attention (SOCA) mechanism to adaptively rescale features by considering feature statistics higher than first-order. Such SOCA mechanism allows our network to focus on more informative features and enhance discriminative learning ability. Besides, we also utilize an iterative method for covariance normalization to speed up the training of our network.
- We propose non-locally enhanced residual group (NLRG) structure to build a deep network, which further incorporates non-local operations to capture spatial contextual information, and share-source residual group structure to learn deep features. Besides, the share-source residual group structure through share-source skip connections could allow more abundant information from the LR input to be bypassed and ease the training of the deep network.

2. Related Work

During the past decade, a plenty of image SISR methods have been proposed in the computer vision community, including interpolation-based [37], model-based [34], and CNN-based methods [2, 29, 14, 13, 29, 17, 30, 39, 38]. Due to space limitation, we here briefly review works related to CNN-based SR methods and attention mechanism, which is close to our method.

CNN-based SR models. Recently, CNN-based methods have been extensively studied in image SR, due to their strong nonlinear representational power. Generally, such methods cast SR as an image-to-image regression problem, and learn an end-to-end mapping from LR to HR directly. Most existing CNN-based methods mainly focus on designing a deeper or wider network structure [2, 12, 13, 6, 39, 38]. For example, Dong *et al*. [2] first introduced a shallow three-layer convolutional network (SRCNN) for image SR, which achieves impressive performance. Later, Kim *et al*. designed deeper VDSR [12] and DRCN [13] with more than 16 layers based on residual learning. To further improve the performance, Lim *et al*. [20] proposed a very deep and wide network EDSR by stacking modified residual blocks. The significant performance gain indicates the depth of representation plays a key role in image SR. Other recent works like MemNet [30] and RDN [39], are based on dense blocks [10] to form deep networks and focus on utilizing all the hierarchical features from all the convolutional layers. In addition to focusing on increasing the depth of the network, some other networks, such as NLRN [22] and RCAN [38], improve the performance by considering feature correlations in spatial or channel dimension.

Attention mechanism. Attention in human perception generally means that human visual systems adaptively process visual information and focus on salient areas [16]. In recent years, several trials have embeded attention processing to improve the performance of CNNs for various tasks, such as image and video classification tasks [9, 33]. Wang *et al*. [33] proposed non-local neural network to incorporate non-local operations for spatial attention in video classification. On the contrary, Hu *et al*. [9] proposed SENet to exploit channel-wise relationships to achieve significant performance gain for image classification.

Recently, SENet was introduced to deep CNNs to further improve SR performance [38]. However, SENet only explores first-order statistics (*e.g.*, global average pooling), while ignoring the statistics higher than first-order, thus hindering the discriminative ability of the network. In im-

age SR, features with more high-frequency information are more informative for HR reconstruction. To this end, we propose a deep second-order attention network (SAN) by exploring second-order statistics of features.

3. Second-order Attention Network (SAN)

3.1. Network Framework

As shown in Fig. 2, our SAN mainly consists of four parts: shallow feature extraction, non-locally enhanced residual group (NLRG) based deep feature extraction, upscale module, and reconstruction part. Given $\mathbf{I}_{LR}$ and $\mathbf{I}_{SR}$ as the input and output of SAN. As explored in [20, 39], we apply only one convolutional layer to extract the shallow feature $\mathbf{F}_0$ from the LR input

$$\mathbf{F}_0 = H_{SF}(\mathbf{I}_{LR}), \tag{1}$$

where $H_{SF}(\cdot)$ stands for convolution operation. Then the extracted shallow feature $\mathbf{F}_0$ is used for NLRG based deep feature extraction, which thus produces the deep feature as

$$\mathbf{F}_{DF} = H_{NLRG}(\mathbf{F}_0), \tag{2}$$

where H_{NLRG} represents the NLRG based deep feature extraction module, which consists of several non-local modules to enlarge receptive field and G local-source residual attention group (LSRAG) modules (see Fig. 2). So our proposed NLRG obtains very deep depth and thus provides very large receptive field size. Then the extracted deep feature $\mathbf{F}_{DF}$ is upscaled via the upscale module via

$$\mathbf{F}_{\uparrow} = H_{\uparrow}(\mathbf{F}_{DF}), \tag{3}$$

where $H_{\uparrow}(\cdot)$ and $\mathbf{F}_{\uparrow}$ are a upscale module and upscaled feature respectively. There are some choices to act as upscale part, such as transposed convolution [3], ESPCN [28]. The way of embedding upscaling feature in the last few layers obtains a good trade off between computational burden and performance, and thus is preferable to be used in recent CNN-based SR models [3, 6, 39]. The upscaled feature is then mapped into SR image via one convolution layer

$$\mathbf{I}_{SR} = H_R(\mathbf{F}_{\uparrow}) = H_{SAN}(\mathbf{I}_{LR}), \tag{4}$$

where $H_R(\cdot)$, $H_{\uparrow}(\cdot)$ and H_{SAN} are the reconstruction layer, upscale layer and the function of SAN, respectively.

Then SAN will be optimized with a certain loss function. Some loss functions have been widely used, such as L_2 [2, 12, 29, 30], L_1 [14, 15, 20, 39], perceptual losses [11, 26]. To verify the effectiveness of our SAN, we adopt the same loss functions as previous works (*e.g.*, L_1 loss function). Given a training set with N LR images and their HR counterparts denoted by $\{\mathbf{I}^i_{LR}, \mathbf{I}^i_{HR}\}_{i=1}^N$, the goal of training SAN is to optimize the L_1 loss function:

$$L(\Theta) = \frac{1}{N}\sum_{i=1}^{N} ||H_{SAN}(\mathbf{I}^i_{LR}) - \mathbf{I}^i_{HR}||_1, \tag{5}$$

where Θ denotes the parameter set of SAN. The loss function is optimized by stochastic gradient descent algorithm.

3.2. Non-locally Enhanced Residual Group (NLRG)

We now show our non-locally enhanced residual group (NLRG) (see Fig. 2), which consists of several region-level non-local (RL-NL) modules and one share-source residual group (SSRG) structure. The RL-NL exploits the abundant structure cues in LR features and the self-similarities in HR nature scenes. The SSRG is composed of G local-source residual attention groups (LSRAG) with share-source skip connections (SSC). Each LSRAG further contains M simplified residual blocks with local-source skip connection, followed by a second-order channel attention (SOCA) module to exploit feature interdependencies.

It has been verified that stacking residual blocks is helpful to form a deep CNN in [20, 39]. However, very deep network built in such way would suffer from training difficulty and performance bottleneck due to the problem of gradient vanishing and exploding in deep network. Inspired by the work in [15], we propose local-source residual attention group (LSRAG) as the fundamental unit. It is known that simply stacking repeated LSRAGs would fail to obtain better performance. To address this issue, the share-source skip connection (SSC) is introduced in NLRG to not only facilitate the training of our deep network, but also to bypass abundant low-frequency information from LR images. Then a LSRAG in the g-th group is represented as:

$$\mathbf{F}_g = W_{SSC}\mathbf{F}_0 + H_g(\mathbf{F}_{g-1}), \tag{6}$$

where W_{SSC} denotes the weight to the convolution layer, and is initialized as 0, and then gradually learns to assign more weight to the shallow feature. The bias term is omitted for simplicity. $H_g(\cdot)$ is the function of the g-th LSRAG. $\mathbf{F}_g, \mathbf{F}_{g-1}$ denote the input and output of the g-th LSRAG. The deep feature is then obtained as:

$$\mathbf{F}_{DF} = W_{SSC}\mathbf{F}_0 + \mathbf{F}_G. \tag{7}$$

Such SSRG structure can not only ease the flow of information across LSRAGs, but also make it possible to train very deep CNN for image SR with high performance.

Region-level non-local module (RL-NL). The proposed NLRG also exploits the abundant structure cues in LR features and the self-similarities in HR nature scenes by RL-NL modules plugged before and after the SSRG. The non-local neural network [33] is proposed to capture the computation of long-range dependencies throughout the entire

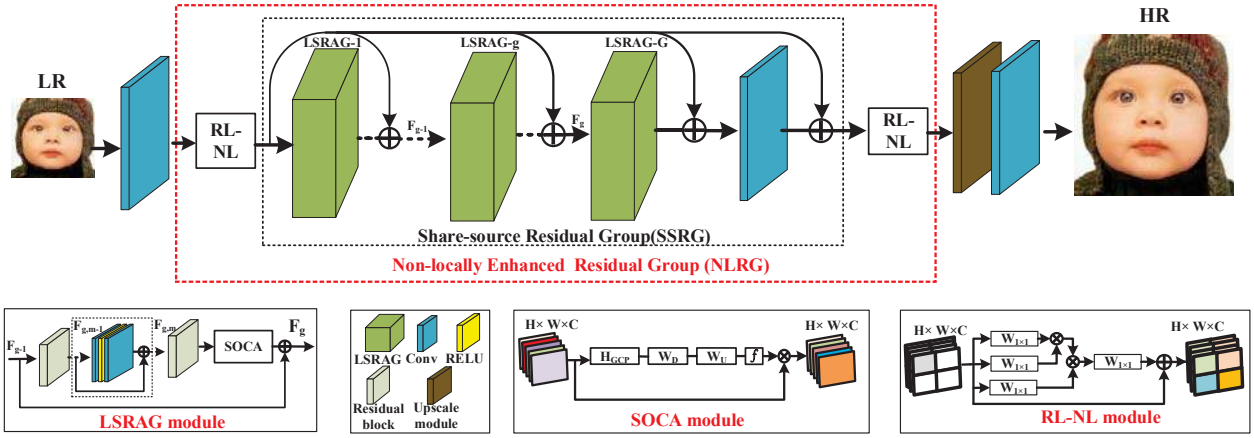


Figure 2. Framework of the proposed second-order attention network (SAN) and its sub-modules.

image for high-level tasks. However, traditional global-level non-local operations may be limited for some reasons: 1) global-level non-local operations require unacceptable computational burden, especially when the size of feature is large; 2) it is empirically shown that non-local operations at a proper neighborhood size are preferable for low-level tasks (*e.g.*, image super-resolution) [22]. Thus for feature with higher spatial resolution or degradation, it is natural to perform region-level non-local operations. For such reasons, we divide the feature map into a grid of regions (see Fig. 2, the $k \times k$ RL-NL indicates the input feature is first divided into a grid of k^2 blocks with the same size.), each of which is then processed by the subsequent layers.

After non-local operations, the feature representation is non-locally enhanced before fedding into the subsequent layers via exploiting the spatial correlations of features.

Local-source residual attention group (LSRAG). Due to our share-source skip connections, the abundant low-frequency information can be bypassed. To go a further step to residual learning, we stack M simplified residual blocks to form a basic LSRAG. The m-th residual block (see Fig. 2) in the g-th LSRAG can be represented as

$$\mathbf{F}_{g,m} = H_{g,m}(\mathbf{F}_{g,m-1}), \tag{8}$$

where $H_{g,m}(\cdot)$ denotes the function of m-th residual block in g-th LSRAG, and $\mathbf{F}_{g,m-1}, \mathbf{F}_{g,m}$ are the corresponding input and output. To make our network focus on more informative features, a local-source skip connection is used to produce the block output via

$$\mathbf{F}_g = W_g \mathbf{F}_{g-1} + \mathbf{F}_{g,M}, \tag{9}$$

where W_g is the corresponding weight. Such local-source and share-source skip connections allow more abundant low-frequency information to be bypassed during training. For more discriminative representations, we propose SOCA mechnism embedded at the tail of each LSRAG. Our SOCA mechnism learns to adaptively rescale channel-wise features by considering second-order statistics of features.

3.3. Second-order Channel Attention (SOCA)

Most previous CNN-based SR models do not consider the feature interdependencies. To utilize such information, SENet [9] was introduced in CNNs to rescale the channel-wise features for image SR. However, SENet only exploits first-order statistics of features by global average pooling, while ignoring statistics higher than first-order, thus hindering the discriminative ability of the network. On the other hand, recent works [19, 21] have shown that second-order statistics in deep CNNs are more helpful for more discriminative representations than first-order ones.

Inspired by the above observations, we propose a second-order channel attention (SOCA) module to learn feature interdependencies by considiering second-order statistics of features. Now we will describe how to exploit such second-order information next.

Covariance normalization. Given a $H \times W \times C$ feature map $\mathbf{F} = [\mathbf{f}_1, \cdots, \mathbf{f}_C]$ with C feature maps with size of $H \times W$. We reshape the feature map to a feature matrix $\mathbf{X}$ with $s = WH$ features of C-dimension. Then the sample covariance matrix can be computed as

$$\mathbf{\Sigma} = \mathbf{X}\bar{\mathbf{I}}\mathbf{X}^T, \tag{10}$$

where $\bar{\mathbf{I}} = \frac{1}{s}(\mathbf{I} - \frac{1}{s}\mathbf{1})$, $\mathbf{I}$ and $\mathbf{1}$ are the $s \times s$ identity matrix and matrix of all ones, respectively.

It is shown in [27, 19] that covariance normalization plays a critical role for more discriminative representations. For this reason, we first perform covariance normalization for the obtained covariance matrix $\mathbf{\Sigma}$, which is symmetric positive semi-definite and thus has eigenvalue decomposition (EIG) as follows

$$\mathbf{\Sigma} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T, \tag{11}$$

where $\mathbf{U}$ is an orthogonal matrix and $\mathbf{\Lambda} = \text{diag}(\lambda_1, \cdots, \lambda_C)$ is diagonal matrix with eigenvalues in non-increasing order. Then convariance normalization

can be converted to the power of eigenvalues:

$$\hat{\mathbf{Y}} = \mathbf{\Sigma}^{\alpha} = \mathbf{U}\mathbf{\Lambda}^{\alpha}\mathbf{U}^{T}, \tag{12}$$

where α is a positive real number, and $\mathbf{\Lambda}^{\alpha} = \text{diag}(\lambda_1^{\alpha}, \cdots, \lambda_C^{\alpha})$. When $\alpha = 1$, there is no normalization; when $\alpha < 1$, it nonlinearly shrinks the eigenvalues larger than 1.0 and streches those less than 1.0. As explored in [19], $\alpha = 1/2$ works well for more discriminative representations. Thus, we set $\alpha = 1/2$ in the following.

Channel attention. The normalized covariance matrix characterizes the correlations of channel-wise features. We then take such normalized covariance matrix as a channel descriptor by global covariance pooling. As illustrated in Fig. 2, let $\hat{\mathbf{Y}} = [\mathbf{y}_1, \cdots, \mathbf{y}_C]$, the channel-wise statistics $\mathbf{z} \in R^{C\times 1}$ can be obtained by shrinking $\hat{\mathbf{Y}}$. Then the c-th dimension of $\mathbf{z}$ is computed as

$$z_c = H_{GCP}(\mathbf{y}_c) = \frac{1}{C}\sum_{i}^{C}\mathbf{y}_c(i), \tag{13}$$

where $H_{GCP}(\cdot)$ denotes the global covariance pooling function. Compared with the commonly used first-order pooling (*e.g.*, global average pooling), our global covariance pooling explores the feature distribution and captures the feature statistics higher than first-order for more discriminative representations.

To fully exploit feature interdependencies from the aggregated information by global covariance pooling, we apply a gating mechanism. As explored in [9], the simple sigmoid function can serve as a proper gating function

$$\mathbf{w} = f(\mathbf{W}_U\delta(\mathbf{W}_D\mathbf{z})), \tag{14}$$

where $\mathbf{W}_D$ and $\mathbf{W}_U$ are the weight set of convolution layer, which set channel dimension of features to C/r and C, respectively. $f(\cdot)$ and $\delta(\cdot)$ are the function of sigmoid and RELU. Finally, we obtain the channel attention map $\mathbf{w}$ to rescale the input

$$\hat{\mathbf{f}}_c = w_c \cdot \mathbf{f}_c, \tag{15}$$

where w_c and $\mathbf{f}_c$ denote the scaling factor and feature map in the c-th channel. With such channel attention, the residual component in the LSRAG is rescaled adaptively.

As is shown above, covariance normalization plays a vital role in our SOCA. However, such covariance normalization relies heavily on eigenvalue decoomposition, which is not well supported on GPU platform, thus leading to inefficient training. To solve this issue, as explored in [18], we also apply a fast matrix normalization method based on Newton-Schulz iteration [8]. In the next section, we briefly describe the covariance normalization.

3.4. Covariance Normalization Acceleration

To date, fast implementation of EIG on GPU is still an open problem. Inspired by [18], we utilize Newton-Schulz iteration to speed up the computation of covariance normalization. Specifically, from Equ. (11), the $\mathbf{\Sigma}$ has square root as $\mathbf{\Sigma}^{1/2} = \mathbf{Y} = \mathbf{U}\text{diag}(\lambda_i^{1/2})\mathbf{U}^T$. Given $\mathbf{Y}_0 = \mathbf{\Sigma}, \mathbf{Z_0} = \mathbf{I}$, for $n = 1, \cdots, N$, as shown in [18], the Newton-Schulz iteration is then updated alternately as follows:

$$\begin{aligned} \mathbf{Y}_n &= \tfrac{1}{2}\mathbf{Y}_{n-1}(3\mathbf{I} - \mathbf{Z}_{n-1}\mathbf{Y}_{n-1}), \\ \mathbf{Z}_n &= \tfrac{1}{2}(3\mathbf{I} - \mathbf{Z}_{n-1}\mathbf{Y}_{n-1})\mathbf{Z}_{n-1}. \end{aligned} \tag{16}$$

After enough iterations, $\mathbf{Y}_n$ and $\mathbf{Z}_n$ quadratically converges to $\mathbf{Y}$ and $\mathbf{Y}^{-1}$. Such iterative operation is suitable for parallel implementation on GPU. In practice, one can achieve approximate solution with few iterations, *e.g.*, no more than 5 iterations in our method.

Since Newton-Schulz iteration only converges locally, to guarantee the convergence, we pre-normalize $\mathbf{\Sigma}$ first via

$$\hat{\mathbf{\Sigma}} = \frac{1}{\text{tr}(\mathbf{\Sigma})}\mathbf{\Sigma}, \tag{17}$$

where $\text{tr}(\mathbf{\Sigma}) = \sum_i^C \lambda_i$ denotes the trace of $\mathbf{\Sigma}$. In such case, it can be inferred that the $||\mathbf{\Sigma} - \mathbf{I}||_{\mathbf{2}}$ equals to the largest singular value of $(\mathbf{\Sigma} - \mathbf{I})$, i.e., $1 - \frac{\lambda_i}{\sum_i \lambda_i)}$ less than 1, which thus satisfies the convergence condition.

After Newton-Schulz iteration, we apply a post-compensation procedure to compensate the data magnitude caused by pre-normalization, thus producing the final normalized covariance matrix

$$\hat{\mathbf{Y}} = \sqrt{\text{tr}(\mathbf{\Sigma})}\mathbf{Y}_N. \tag{18}$$

3.5. Implementations

We set LSRAG number as $G = 20$ in the SSRG structure, and embed RL-NL modules ($k = 2$) at the head and tail of SSRG. In each LSRAG, we use $m = 10$ residual blocks plus single SOCA module at the tail. In SOCA module, we use 1×1 convolution filter with reduction ratio $r = 16$. For other convolution filter outside SOCA, the size and number of filter are set as 3×3 and $C = 64$, respectively. For upscale part $H_{\uparrow}(\cdot)$, we follow the works in [20, 39] and apply ESPCNN [28] to upscale the deep features, followed by one final convolution layer with three filters to produce color images (RGB channels).

3.6. Discussions

Difference to Non-local RNN (NLRN). NLRN [22] introduces non-local operations to capture long-distance spatial contextual information in image restoration. There are some differences between NLRN and our SAN. First, NLRN embeds non-local operations in a recurrent neural

network (RNN) for image restoration, while our SAN incorporates non-local operations in deep convolutional neural network (CNN) framework for image SR. Second, NLRN only considers spatial feature correlations between each location and its neighborhood, but ignores the channel-wise feature correlations. While our SAN mainly focuses on learning such channel-wise feature correlations with second-order statistics of features for more powerful representational ability.

Difference to Residual Dense Network (RDN). We summarize the main differences between RDN [39] and our SAN. The first one is the design of basic block. RDN mainly combines dense blocks with local feature fusion by using local residual learning, while our SAN is built on the basis of residual blocks. The second one is the way of enhancing discriminative ability of the network. Channel attention [9, 38] has been shown to be effective for better discriminative representations. However, RDN does not consider such information, but pays attention to exploiting the hierarchical features from all the convolutional layers. On the contrary, our SAN heavily relies on channel attention for better discriminative representations. Thus, we propose second-order channel attention (SOCA) mechanism to effectively learn channel-wise feature interdependencies.

Difference to Residual Channel Attention Network (RCAN). Zhang et al. [38] proposed a residual in residual structure to form a very deep network. RCAN is close to our SAN, and the main differences lie in the following aspects. First, RCAN consists of several residual groups with long skip connections. While, SAN stacks repeated residual groups through share-source skip connections, which allows more abundant low-frequency information to be bypassed. Second, RCAN can only exploit the contextual information in a local receptive field, but is unable to exploit the information outside of the local region. While SAN can alleviate this problem by incorporating non-local operations to not only capture long-distance spatial contextual information, but enlarge the receptive field. Third, to enhance the discriminative ability of the network, RCAN only considers channel attention based first-order feature statistics by global average pooling. While our SAN learns channel attention based on second-order feature statistics.

To the best of our knowledge, it is the first attempt to investigate the effect of such attention based on second-order feature statistics for image SR. More analysis about the effect of such attention mechanism are shown next.

4. Experiments

4.1. Setup

Following [20, 6, 39, 38], we use 800 high-resolution images from DIV2K dataset [31] as training set. For testing, we adopt 5 standard benchmark datasets: Set5, Set14, BSD100, Urban100 and Manga109, each of which has different characteristics. We carry out experiments with Bicubic (BI) and Blur-downscale (BD) degradation models [36]. All the SR results are evaluated by PSNR and SSIM metrics on Y channel of transformed YCbCr space.

During training, we augment the training images by randomly rotating $90°, 180°, 270°$ and horizontally flipping. In each min-batch, 8 LR color patches with size 48×48 are provided as inputs. Our model is trained by ADAM optimizor with $\beta_1 = 0.9, \beta_2 = 0.99$, and $\varepsilon = 10^{-8}$. The learning rate is initialized as 10^{-4} and then reduced to half every 200 epochs. Our proposed SAN has been implemented on the Pytorch framework [23] on an Nvidia 1080Ti GPU.

4.2. Ablation Study

As discussed in Section 3, our SAN contains two main components including non-locally enhanced residual group (NLRG) and second-order channel attention (SOCA).

Non-locally Enhanced Residual Group (NLRG). To verify the effectiveness of different modules, we compare NLRG with its variants trained and tested on Set5 dataset. The specific performance is listed in Table 1.

$Base$ refers to a very basic baseline which only contains the convolution layers with 20 LSRAGs and 10 residual blocks in each LSRAG, thus resulting in deep network with over 400 convolution layers. As in [38], we also add long and short skip connections in $Base$ model. From Table 1 we can see that $Base$ reaches PSNR=32.00 dB on Set5 ($\times 4$). Results from R_a to R_e verify the effectiveness of individual module, since the module used alone improves the performance over $Base$ model. Specifically, R_a and R_b that add a single RL-NL in shallow (before SSRG) or deep layers (after SSRG) obtain similar SR results and outperform $Base$, which verifies the effectiveness of RL-NL. When share-source skip connection (SSC) is added alone (R_c), the performance can be improved from 32.00 dB to 32.07 dB. The main reason lies in that share-source skip connections allows more abundant low-frequency information from the LR images to be bypassed. When both of R_a and R_b are used (leading to R_f), the performance can be further improved. It is found more RL-NL modules cannot obtain much better performance than R_f in our method, and thus we apply R_f in our method to balance the performance and efficiency.

Second-order channel attention (SOCA). We also show the effect of our SOCA from the results of R_d, R_e, R_h and R_i. Specifically, R_d means that channel attention is based on first-order feature statistics by global average pooling, thus leading to first-order channel attention (FOCA). R_e means that channel attention is based on second-order feature statistics, thus leading to our second-order channel attention (SOCA). It can be found that both of R_d and R_e obtain better performance than methods of R_a to R_c with-

Table 1. Effects of different modules. We report the best PSNR (dB) values on Set5 ($4\times$) in 5.6×10^5 iterations.

	$Base$	R_a	R_b	R_c	R_d	R_e	R_f	R_g	R_h	R_i
RL-NL(before SSRG)		✓					✓	✓	✓	✓
RL-NL(after SSRG)			✓				✓	✓	✓	✓
share-source skip connection (SSC)				✓				✓	✓	✓
First-order channel attention (FOCA)					✓				✓	
Second-order channel attention (SOCA)						✓				✓
	32.00	32.04	32.06	32.07	32.12	32.16	32.08	32.10	32.14	32.20

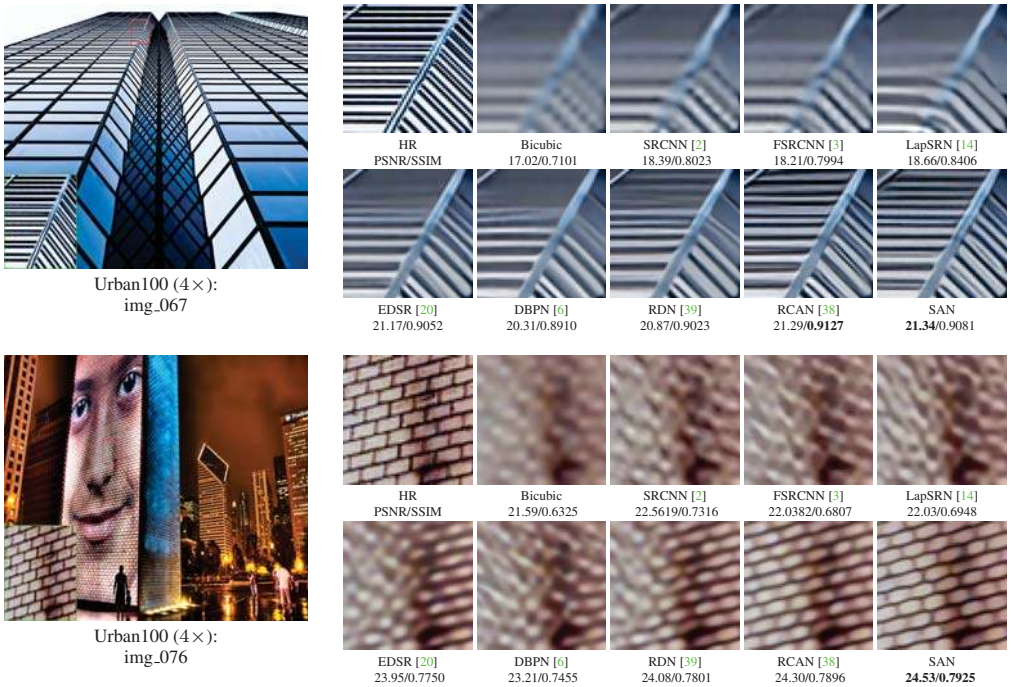


Figure 3. Visual comparison for $4\times$ SR with BI model on Urban100 dataset. The best results are **highlighted**

out channel attention. This indicates that channel attention plays a more important role in determining the performance. Furthermore, compared with FOCA, our SOCA achieves consistently better results, no matter if combined with other modules (*e.g.*, RL-NL and SSC). These observations demonstrate the superiority of our SOCA.

4.3. Results with Bicubic Degradation (BI)

To test the effectiveness of our SAN, we compare our SAN with 11 state-of-the-art CNN-based SR methods: SRCNN [1], FSRCNN [3], VDSR [12], LapSRN [14], MemNet [30], EDSR [20], SRMD [36], NLRN [22], DBPN [6], RDN [39] and RCAN [38]. As in [20, 39, 38], we also adopt self-ensemble method to further improve our SAN denoted as SAN+. All the quantitative results for various scaling factors are reported in Table 2. Compared with other methods, our SAN+ performs the best results on all the datasets on various scaling factors. Without self-ensemble, SAN and RCAN obtain very similar results and outperform other methods. This is mainly because both of them adopt channel attention to learn feature interdependencies, thus making the network focus on more informative features. Compared with RCAN, our SAN obtains better results for datasets (*e.g.*, such as Set5, Set14 and BSD100) with rich texture information, while obtaining a little worse results for datasets(*e.g.*, Urban100 and Manga109) with rich repeated edge information. It is known that textures are high-order patterns and have more complex statistic characteristics, while edges are first-order patterns that can be extracted by first-order gradient operators. Thus our SOCA based on second-order feature statistics works better on images with more high-order information (*e.g.*, textures).

Visual quality. We also show the zoomed results of various methods in Fig. 3, from which we can see that most compared SR models cannot reconstruct the lattices accurately and suffer from serious blurring artifact. In contrast, our SAN obtains sharper results and recovers more high-frequency details, such as high contrast and sharp edges. Take "img_076" for example, most compared methods output heavy blurring artifacts. The early developed bicubic, SRCNN, FSRCNN and LapSRN even lose the main structure. More recent methods (*e.g.*, EDSR, DBPN and RDN) can recover the main outlines but fail to recover more image details. Compared with the ground-truth, RCAN and

Table 2. Quantitative results with BI degradation model.

Method		Set5	Set14	BSD100	Urban100	Manga109
		PSNR/ SSIM	PSNR/ SSIM	PSNR/ SSIM	PSNR/ SSIM	PSNR/ SSIM
Bicubic	2	33.66/9299	30.24/8688	29.56/8431	26.88/8403	30.80/9339
SRCNN	2	36.66/9542	32.45/9067	31.36/8879	29.50/8946	35.60/9663
FSRCNN	2	37.05/9560	32.66/9090	31.53/8920	29.88/9020	36.67/9710
VDSR	2	37.53/9590	33.05/9130	31.90/8960	30.77/9140	37.22/9750
LapSRN	2	37.52/9591	33.08/9130	31.08/8950	30.41/9101	37.27/9740
MemNet	2	37.78/9597	33.28/9142	32.08/8978	31.31/9195	37.72/9740
EDSR	2	38.11/9602	33.92/9195	32.32/9013	32.93/9351	39.10/9773
SRMD	2	37.79/9601	33.32/9159	32.05/8985	31.33/9204	38.07/9761
NLRN	2	38.00/9603	33.46/9159	32.19/8992	31.81/9246	−−/ −−
DBPN	2	38.09/9600	33.85/9190	32.27/9000	32.55/9324	38.89/9775
RDN	2	38.24/9614	34.01/9212	32.34/9017	32.89/9353	39.18/9780
RCAN	2	38.27/9614	34.11/9216	32.41/9026	33.34/9384	39.43/9786
SAN	2	38.31/9620	34.07/9213	32.42/9028	33.10/9370	39.32/9792
SAN+	2	**38.35/9619**	**34.44/9244**	**32.50/9038**	**33.73/9416**	**39.72/9797**
Bicubic	3	30.39/8682	27.55/7742	27.21/7385	24.46/7349	26.95/8556
SRCNN	3	32.75/9090	29.30/8215	28.41/7863	26.24/7989	30.48/9117
FSRCNN	3	33.18/9140	29.37/8240	28.53/7910	26.43/8080	31.10/9210
VDSR	3	33.67/9210	29.78/8320	28.83/7990	27.14/8290	32.01/9340
LapSRN	3	33.82/9227	29.87/8320	28.82/7980	27.07/8280	32.21/9350
MemNet	3	34.09/9248	30.01/8350	28.96/8001	27.56/8376	32.51/9369
EDSR	3	34.65/9280	3.52/ .8462	29.25/8093	28.80/8653	34.17/9476
SRMD	3	34.12/9254	30.04/8382	28.97/8025	27.57/8398	33.00/9403
NLRG	3	34.27/9266	30.16/8374	29.06/8026	27.93/8453	−−/ −−
RDN	3	34.71/9296	30.57/8468	29.26/8093	28.80/8653	34.13/9484
RCAN	3	34.74/9299	30.64/8481	29.32/8111	29.08/8702	34.43/9498
SAN	3	34.75/9300	30.59/8476	29.33/8112	28.93/8671	34.30/9494
SAN+	3	**34.89/9306**	**30.77/8498**	**29.38/8121**	**29.29/8730**	**34.74/9512**
Bicubic	4	28.42/8104	26.00/7027	25.96/6675	23.14/6577	24.89/7866
SRCNN	4	30.48/8628	27.50/7513	26.90/7101	24.52/7221	27.58/8555
FSRCNN	4	30.72/8660	27.61/7550	26.98/7150	24.62/7280	27.90/8610
VDSR	4	31.35/8830	28.02/7680	27.29/0726	25.18/7540	28.83/8870
LapSRN	4	31.54/8850	28.19/7720	27.32/7270	25.21/7560	29.09/8900
MemNet	4	31.74/8893	28.26/7723	27.40/7281	25.50/7630	29.42/8942
EDSR	4	32.46/8968	28.80/7876	27.71/7420	26.64/8033	31.02/9148
SRMD	4	31.96/8925	28.35/7787	27.49/7337	25.68/7731	30.09/9024
DBPN	4	32.47/8980	28.82/7860	27.72/7400	26.38/7946	30.91/9137
NLRG	4	31.92/8916	28.36/7745	27.48/7346	25.79/7729	−−/ −−
RDN	4	32.47/8990	28.81/7871	27.72/7419	26.61/8028	31.00/9151
RCAN	4	32.62/9001	28.86/7888	27.76/7435	26.82/8087	31.21/9172
SAN	4	32.64/9003	28.92/7888	27.78/7436	26.79/8068	31.18/9169
SAN+	4	**32.70/9013**	**29.05/7921**	**27.86/7457**	**27.23/8169**	**31.66/9222**
Bicubic	8	24.40/6580	23.10/5660	23.67/5480	20.74/5160	21.47/6500
SRCNN	8	25.33/6900	23.76/5910	24.13/5660	21.29/5440	22.46/6950
FSRCNN	8	20.13/5520	19.75/4820	24.21/5680	21.32/5380	22.39/6730
SCN	8	25.59/7071	24.02/6028	24.30/5698	21.52/5571	22.68/6963
VDSR	8	25.93/7240	24.26/6140	24.49/5830	21.70/5710	23.16/7250
LapSRN	8	26.15/7380	24.35/6200	24.54/5860	21.81/5810	23.39/7350
MemNet	8	26.16/7414	24.38/6199	24.58/5842	21.89/5825	23.56/7387
MSLap	8	26.34/7558	24.57/6273	24.65/5895	22.06/5963	23.90/7564
EDSR	8	26.96/7762	24.91/6420	24.81/5985	22.51/6221	24.69/7841
DBPN	8	27.21/7840	25.13/6480	24.88/6010	22.73/6312	25.14/**7987**
SAN	8	27.22/7829	25.14/6476	24.88/6011	22.70/6314	24.85/7906
SAN+	8	**27.30/7849**	**25.23/6493**	**24.97/6031**	**22.91/6369**	**25.17**/7964

SAN obtain more faithful results and recover more image details, but SAN has sharper results. These observations verify the superiority of SAN with more powerful representational ability. Although the recovery of high-frequency information is difficult due to limited information available in LR input (scaling factor $> 4\times$), our SAN can still make full use of the limited LR information through share-source skip connections, and simultaneously utilize both spatial and channel feature correlations for more powerful feature expressions, thus producing more finer results.

4.4. Results with Blur-downscale Degradation (BD)

Following [36, 39], we also compare various SR methods on image with blur-down degradation (BD) model. We compare our method with 8 state-of-the-art SR methods: SPMSR [24], SRCNN [2], FSRCNN [3], VDSR [12], IRCNN [35], SRMD [36], RDN [39], and RCAN [38]. All the results on $3\times$ are shown in Table 3, from which we can observe that our SAN achieves consistently better performance than other methods even without self-ensemble. Specifically, the PSNR gain of SAN over RDN is up to 0.4 dB on Urban100 and Manga109 datasets.

Table 3. Quantitative results with BD degradation model. Best and second best results are **highlighted** and underlined

Method		Set5	Set14	BSD100	Urban100	Manga109
		PSNR/ SSIM	PSNR/ SSIM	PSNR/ SSIM	PSNR/ SSIM	PSNR/ SSIM
Bicubic	3	28.78/8308	26.38/7271	26.33/6918	23.52/6862	25.46/8149
SPMSR	3	32.21/9001	28.89/8105	28.13/7740	25.84/7856	29.64/9003
SRCNN	3	32.05/8944	28.80/8074	28.13/7736	25.70/7770	29.47/8924
FSRCNN	3	26.23/8124	24.44/7106	24.86/6832	22.04/6745	23.04/7927
VDSR	3	33.25/9150	29.46/8244	28.57/7893	26.61/8136	31.06/9234
IRCNN	3	33.38/9182	29.63/8281	28.65/7922	26.77/8154	31.15/9245
SRMD	3	34.01/9242	30.11/8364	28.98/8009	27.50/8370	32.97/9391
RDN	3	34.58/9280	30.53/8447	29.23/8079	28.46/8582	33.97/9465
RCAN	3	34.70/9288	30.63/8462	29.32/8093	28.81/8645	34.38/9483
SAN	3	34.75/9290	30.68/8466	29.33/8101	28.83/8646	34.46/9487
SAN+	3	**34.86/9297**	**30.77/8481**	**29.39/8112**	**29.03/8674**	**34.76/9501**

Table 4. Computational and parameter comparison ($2\times$ Set5).

	EDSR	MemNet	NLRG	DBPN	RDN	RCAN	SAN
Para.	43M	677k	330k	10M	22.3M	16M	15.7M
PSNR	38.11	37.78	38.00	38.09	38.24	38.27	**38.31**

4.5. Model Size Analyses

The Table 4 shows the performance and model size of recent deep CNN-based SR methods. Among these methods, MemNet and NLRG contain much less parameters at the cost of performance degradation. Instead, our SAN has lless parameters than RDN and RCAN, but obtains higher performance, which implies that our SAN can have a good trade-off between performance and model complexity.

5. Conclusions

We propose a deep second-order attention network (SAN) for accurate image SR. Specifically, the non-locally enhanced residual group (NLRG) structure allows SAN to capture the long-distance dependencies and structural information by embedding non-local operations in the network. Meanwhile, NLRG allows abundant low-frequency information from the LR images to be bypassed through share-source skip connections. In addition to exploiting the spatial feature correlations, we propose second-order channel attention (SOCA) module to learn feature interdependencies by global covariance pooling for more discriminative representations. Extensive experiments on SR with BI and BD degradation models show the effectiveness of our SAN in terms of quantitative and visual results.

References

[1] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *ECCV*, 2014. 7

[2] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *TPAMI*, 2016. 1, 2, 3, 7, 8

[3] Chao Dong, Chen Change Loy, and Xiaoou Tang. Accelerating the super-resolution convolutional neural network. In *ECCV*. Springer, 2016. 3, 7, 8

[4] Weisheng Dong, Lei Zhang, Guangming Shi, and Xiaolin Wu. Image deblurring and super-resolution by adaptive sparse domain selection and adaptive regularization. *TIP*, 2011. 1

[5] William T Freeman, Egon C Pasztor, and Owen T Carmichael. Learning low-level vision. *IJCV*, 40(1):25–47, 2000. 1

[6] Muhammad Haris, Greg Shakhnarovich, and Norimichi Ukita. Deep backprojection networks for super-resolution. In *CVPR*, 2018. 2, 3, 6, 7

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1

[8] Nicholas J Higham. *Functions of matrices: theory and computation*. SIAM, 2008. 5

[9] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, 2018. 2, 4, 5, 6

[10] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *CVPR*, 2017. 2

[11] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016. 3

[12] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *CVPR*, 2016. 1, 2, 3, 7, 8

[13] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Deeply-recursive convolutional network for image super-resolution. In *CVPR*, 2016. 2

[14] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate superresolution. In *CVPR*, 2017. 1, 2, 3, 7

[15] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Fast and accurate image super-resolution with deep laplacian pyramid networks. *arXiv preprint arXiv:1710.01992*, 2017. 3

[16] Christof Koch Laurent Itti and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *PAMI*, 1998. 2

[17] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew P Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, 2017. 2

[18] Peihua Li, Jiangtao Xie, Qilong Wang, and Zilin Gao. Towards faster training of global covariance pooling networks by iterative matrix square root normalization. In *CVPR*, 2018. 5

[19] Peihua Li, Jiangtao Xie, Qilong Wang, and Wangmeng Zuo. Is second-order information helpful for large-scale visual recognition. In *ICCV*, 2017. 4, 5

[20] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *CVPRW*, 2017. 2, 3, 5, 6, 7

[21] Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji. Bilinear cnn models for fine-grained visual recognition. In *ICCV*, 2015. 4

[22] Ding Liu, Bihan Wen, Yuchen Fan, Chen Change Loy, and Thomas S Huang. Non-local recurrent network for image restoration. In *NIPS*, 2018. 2, 4, 5, 7

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

[24] Tomer Peleg and Michael Elad. A statistical prediction model based on sparse representations for single image super-resolution. *TIP*, 2014. 8

[25] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015. 1

[26] Mehdi SM Sajjadi, Bernhard Schölkopf, and Michael Hirsch. Enhancenet: Single image super-resolution through automated texture synthesis. In *ICCV*, 2017. 3

[27] Jorge Sánchez, Florent Perronnin, Thomas Mensink, and Jakob Verbeek. Image classification with the fisher vector: Theory and practice. *IJCV*. 4

[28] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *CVPR*, 2016. 3, 5

[29] Ying Tai, Jian Yang, and Xiaoming Liu. Image super-resolution via deep recursive residual network. In *CVPR*, 2017. 2, 3

[30] Ying Tai, Jian Yang, Xiaoming Liu, and Chunyan Xu. Memnet: A persistent memory network for image restoration. In *CVPR*, pages 4539–4547, 2017. 2, 3, 7

[31] Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, Lei Zhang, Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, Kyoung Mu Lee, et al. Ntire 2017 challenge on single image super-resolution: Methods and results. In *CVPRW*, 2017. 6

[32] Shenlong Wang, Lei Zhang, Yan Liang, and Quan Pan. Semi-coupled dictionary learning with applications to image super-resolution and photo-sketch synthesis. In *CVPR*, 2012. 1

[33] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, 2018. 2, 3

[34] K. Zhang, X. Gao, D. Tao, and X. Li. Single image super-resolution with non-local means and steering kernel regression. *TIP*, 21(11):4544–4556, 2012. 1, 2

[35] Kai Zhang, Wangmeng Zuo, Shuhang Gu, and Lei Zhang. Learning deep cnn denoiser prior for image restoration. In *CVPR*, 2017. 8

[36] Kai Zhang, Wangmeng Zuo, and Lei Zhang. Learning a single convolutional super-resolution network for multiple degradations. In *CVPR*, 2018. 1, 2, 6, 7, 8

[37] Lei Zhang and Xiaolin Wu. An edge-guided image interpolation algorithm via directional filtering and data fusion. *TIP*, 2006. 1, 2

[38] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *ECCV*, 2018. 1, 2, 6, 7, 8

[39] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *CVPR*, 2018. 1, 2, 3, 5, 6, 7, 8