Lightweight Single Image Super-Resolution With Multi-Scale Spatial Attention Networks

JAE WOONG SOH®, (Student Member, IEEE), AND NAM IK CHO®, (Senior Member, IEEE)
Department of Electrical and Computer Engineering, INMC, Seoul National University, Seoul 08826, South Korea
Corresponding author: Nam Ik Cho (nicho@snu.ac.kr)

This work was supported in part by the Ministry of Trade, Industry, and Energy (MOTIE), South Korea, under the Regional Innovation Cluster Development Program Research and Development, under Grant P0002072, in part by the Korea Institute for Advancement of Technology (KIAT), and in part by the Samsung Electronics Company, Ltd.

ABSTRACT Convolutional neural networks (CNNs) generally provide higher performance gain for single image super-resolution (SISR) as the depth and number of parameters are increasing. However, just increasing the layers of straightforward deep networks has a problem that it requires an impractically large number of parameters for obtaining state-of-the-art performance. Instead, some researchers proposed lightweight networks, which is designed with more sophisticated network structures for achieving better performance than the straightforward networks at the same parameter requirement. In this paper, we propose new lightweight Multi-scale Spatial Attention Networks (MSAN) for SISR, which attempt to bring out a better performance from the relatively small number of parameters. Specifically, we adopt a dense connection with feature fusion layers to broadcast abundant features to every level of layers, and propose a double residual structure that provides an additional skip-connection. We also design a Multi-scale Spatial Attention Block (MSAB) to exploit multi-scale spatial contextual information. Furthermore, we introduce a spatial attention module which adaptively focuses on the most informative feature scale in a given region of the image. In the experiments, we validate that the proposed MSAN achieves significant accuracy compared to recent lightweight models and comparable performance to the state-of-the-art methods.

INDEX TERMS Convolutional neural network (CNN), lightweight, multi-scale spatial attention, single image super-resolution (SISR).

I. INTRODUCTION

Recently, deep convolutional neural networks (CNNs) have shown great success in most of computer vision tasks, including single-image super-resolution. Since Dong et al. [1] first proposed a three-layer light CNN for the SISR, many researchers have proposed better and deeper CNNs for SISR [1]–[16]. Earlier works mostly focused on training deeper models with straightforward structure (stack of single-size small-kernel convolution layers) or a large number of filters per layer to achieve state-of-the-art performance without considering the undue increase of parameters. For instance, one of the state-of-the-art networks, EDSR [7] has about 43 M parameters with 256 number of filters for a layer. Also, RCAN [9] has more than 400 layers, which is enormously large compared to the three-layer SRCNN that needs just 57 K parameters.

Hence, rather than just deepening the network, some researchers tried to find more sophisticated structures and methods to achieve better performance with less number of parameters. Specifically, there are relatively lightweight networks which attempt to achieve high performance while requiring the number of parameters under 2 M.® For example, IDN [19] was proposed as a fast and accurate SISR network with information distillation. DSRN [17] adopted dual-state recurrent network and achieved competitive performance. CARN [18] was proposed as a fast, accurate, and lightweight CNN for SISR with some sophisticated techniques.

We note that conventional networks for SISR are based on stacked $3 \times 3$ convolution layers, which seem to lack flexibility in dealing with the diverse scale of spatial contexts. Precisely, they cannot capture both local and global

® Although there is no exact criterion for the lightweight, we consider the limit as 2 M because most of the recent sophisticated lightweight networks keep the number of parameters under 2 M [17]–[19].
features at the same time, and rarely learn multi-scale non-local self-similarity due to the limited receptive fields. Hence, the features from CNNs are often redundant to deal with image structure or not very useful to exploit multi-scale structure. Moreover, as evidenced in [20], internal structure and multi-scale recurrence are sometimes more important than the evidence from external training patches. To fully utilize such properties, it is required to capture both local and global contextual information simultaneously.

Regarding the above issues, we propose Multi-scale Spatial Attention Networks (MSAN) that can adaptively give attention to the most appropriate scale of features in a specific region of the image. Precisely, we design a Multi-scale Spatial Attention Block (MSAB) as a basic building block for the MSAN. The MSAB includes a multi-scale structure with the spatial attention mechanism and local residual learning (LRL). Also, we adopt a dense connection with feature fusion (FF) to push useful information to the overall layers. In addition, instead of using only one global residual learning (GRL), we propose a double residual learning (DRL) scheme to enable the network to learn more sparse features. We also introduce a variant model with the recursive scheme, namely MSAN-X (eXtremely lightweight model). Experimental results show that the proposed architecture yields superior performance while maintaining the number of parameters to be comparable to the existing lightweight models such as [17]–[19].

In summary, we propose a new lightweight CNN, where we attempt to reduce the number of parameters as much as possible while achieving a performance gain over the other lightweight models. In developing the structure, our main contributions are as follows:

- We propose a new network architecture (MSAN) that extracts dynamic multi-scale features for feature enrichment.
- This is achieved by developing a new basic building block, namely MSAB, which adopts multi-scale structure with dilated convolution that enables the use of local residual learning and spatial attention for the dynamic feature selection.
- We also propose an extremely lightweight model, namely MSAN-X, with the recursive scheme.
- Extensive experiments show that both MSAN and MSAN-X achieve state-of-the-art performance compared to the other lightweight SISR methods.

Finally, it is worth to note that increasing the layers of lightweight models such as [17]–[19], and ours does not work well; specifically, they do not converge well when they are deepened. Unlike the classical models that can generally achieve better performance with deeper layers, the structures of lightweight models are so complicated that they are not well trained when deepened. Rather, lightweight models have the virtue of requiring a practically small number of parameters while achieving comparable performance to the classical deeper networks.

II. RELATED WORK

A. CNN-BASED SINGLE IMAGE SUPER-RESOLUTION

As a pioneer work, Dong et al. introduced SRCNN [1] with a three-layer light network. VDSR [2] adopted a very deep structure of 20 convolution layers with residual learning and has proven “the deeper, the better” in SISR. ESPCN [21] first introduced an efficient sub-pixel convolution layer for the upsampling to save the computational cost. DRCN [22] and DRRN [6] adopted the recursive structure and achieved significant results compared to the networks with similar depths but with a smaller number of parameters. Also, MemNet [23] exploited recursive units for general image restoration problem, and LapSRN [3] adopted the Laplacian pyramids in CNN. SRGAN [4] was proposed as a generative approach for SISR based on generative adversarial net (GAN) [24], which focused on the perceptual quality rather than the mean squared error. Recently, EDSR [7], RDN [8] and RCAN [9] have been proposed and achieved very high-performance gain by using very deep architecture with a huge number of parameters more than 10 M (For EDSR, more than 40 M.)

B. LIGHTWEIGHT CNNS FOR SISR

Compared to the networks with a large number of parameters and depth, some of the recent works are focusing on the lightweight but competitive SISR architectures. For example, CARN [18] was proposed as a fast, accurate, and lightweight structure with less number of parameters. Also, IDN [19] achieved the same goal via information distillation network. Dual CNN [25] achieved considerable performance gain with a reduced number of parameters by addressing the dual domains separately. DSRN [17] employed dual-state recurrent structure for the efficient SISR. Liu et al. [10] introduced non-local module within a recurrent network, which adopts non-local self-similarity of natural images. Our approach also focuses on the designs for lightweight networks with competitive performance.

C. MULTI-SCALE SPATIAL FEATURES

Szegedy et al. proposed the Inception module [26], which consists of multiple-size kernels at the same level of layer. Hence, multi-scale contexts in images can be utilized at a layer, and various size of features can be extracted at the deeper layers. As a result, the Inception network has shown outstanding performance compared to the VGG-style networks with much less number of parameters in the image classification task. They also modified Inception module to a better version, proposed “Inception-v4” [27], which leverages the accuracy. For the SISR, Shi et al. [28] proposed a modification of the Inception module as a dilated convolution layer. Recently, MSRN [29] adopted the multi-scale structure for SISR and have shown competitive performances. In this paper, we also focus on multi-scale spatial features to increase and enrich the quality of features.
D. ATTENTION MECHANISM

It is well-known that human visual system captures information not from the whole scene but the salient local parts for efficiency, which is the attention mechanism. Since it plays an essential role in image perception, many researchers developed neural network modules that can incorporate the attention mechanism [9], [16], [30]–[32]. For example, Wang et al. [30] proposed Residual Attention Network for image classification problem by controlling whole feature maps with attention module. Also, squeeze and excitation network with channel-wise attention was proposed in [31]. RCAN [9] adopted a channel-attention mechanism for SISR and achieved state-of-the-art performance. Recently, CBAM [32] was introduced, which considers both spatial and channel attention with a small number of parameters and thus boosts image classification and object detection performances. Very recently, SAN [16] was introduced, which considers second-order statistics for channel attention to further enhance the performance in SISR. Inspired by CBAM [32], we also consider spatial attention with some modification. The original spatial attention module from the CBAM examines “where to focus on,” which is not proper for a low-level vision problem. To be specific with SISR, it is important to focus on the correct scale of the structure to predict its high-frequency component. Hence, it is essential to focus on the most informative features that reflect the appropriate spatial scale, rather than just the saliencies. Thus, our multi-scale spatial attention module is designed to give attention to the features that have the most appropriate scale for the given region in the image.

III. PROPOSED METHODS

In this section, we present our MSAN and MSAN-X for SISR. Our MSAN is designed to extract and exploit enriched features with a reduced number of parameters. Moreover, MSAN-X is designed to further reduce the number of parameters by employing a recursive scheme.

A. SINGLE IMAGE SUPER-RESOLUTION

In this part, we define the SISR problem and present some notations for the rest of the paper. The SISR is to predict a plausible super-resolved image $I_{SR} \in \mathbb{R}^{rH \times rW \times C}$ from a single low-resolution image $I_{LR} \in \mathbb{R}^{H \times W \times C}$, where $r$ is the scaling factor. In this paper, we wish to minimize the difference between the predicted $I_{SR}$ and the true high-resolution image $I_{HR} \in \mathbb{R}^{H \times W \times C}$. For the CNN-based methods, the goal is to find a mapping function $f(I_{LR}; \theta)$ parameterized by $\theta$ where its output is $I_{SR}$. The parameter of CNN is trained by using a paired training dataset $\{(I_{LR}^{(i)}, I_{HR}^{(i)})\}_{i=1}^{N}$, where $N$ is the number of training image pairs.

B. NETWORK ARCHITECTURE

The overall architecture of MSAN is shown in FIGURE 1. The proposed network mostly consists of MSABs with some convolution layers and feature fusion (FF) layers. The MSAB will be explained in the later subsection, which is designed to extract dynamic multi-scale features via local residual learning and spatial attention. The figure also shows that we adopt the dense connection to push informative intermediate features to the higher layers. The global feature fusion (GFF) layer in our network fuses all output features from previous MSABs. Similar to recent SISR works [4], [7], [8], [18], the upsampling process is located at the latter part of the network. Hence most of the computation is done in the LR scale, which greatly reduces computational cost compared to the ones that require bicubic interpolation [1], [2].

The proposed MSAN can be divided into three functional parts: feature extractor, feature enhancer, and reconstructor. The first convolution layer works as the feature extractor, which extracts the low-level feature $F_0$ from the input LR image $I_{LR}$, described as

$$F_0 = C_{FE}(I_{LR}).$$

Then, the central part, which consists of FF layers, MSAB, GFF, and the convolution layer for double residual learning (DRL), works as the feature enhancer. The FF layer fuses...
concatenated features from the previous layers and outputs fused features $H_{d-1}$, which are fed to the next MSAB. The output feature $F_d$ of the $d$-th MSAB is computed as

$$H_{d-1} = C_{FF-d}([F_0, \cdots, F_{d-1}]).$$  

$$F_d = C_{MSAB-d}(H_{d-1}).$$

where $C_{FF-d}()$ and $C_{MSAB-d}()$ denote the $d$-th FF layer and the $d$-th MSAB, respectively ($d = 1, 2, \cdots, D$). After the last MSAB, all output features from the previous MSABs and $F_0$ are concatenated, and then aggregated by GFF. Additionally, we adopt GRL and DRL to provide additional gradient paths. Between two residual learnings, one convolution layer $C_E()$ enhances the features from the previous GRL. Hence, we have

$$F_{GFF} = F_0 + C_{GFF}([F_0, \cdots, F_D]).$$

$$F_E = F_0 + C_E(F_{GFF}),$$

where $F_{GFF}$ is the features after GFF and GRL, and $F_E$ denotes the final enhanced features after the feature enhancer. Then, we feed $F_E$ to the reconstructor which consists of an upsampler $U()$ and the last convolution layer $C_R()$ to upsample and reconstruct the final super-resolved image $I_{SR}$, i.e.,

$$I_{SR} = C_R(U(F_E)).$$

C. MULTI-SCALE SPATIAL ATTENTION BLOCK

The MSAB is the main building block of our proposed network, which is shown in FIGURE 2a. As shown in FIGURE 1, the role of the $d$-th MSAB is to take mid-level features after the $d$-th FF layer $H_{d-1} \in \mathbb{R}^{H \times W \times G}$ as the input, where $G$ denotes the number of features. At the first part of MSAB, four parallel $1 \times 1$ convolution layers $C_{1 \times 1}()$ followed by ReLU compress the input features for each branch, which is expressed as

$$h_{1,d-1}^l = ReLU(C_{1 \times 1}(H_{d-1})).$$

where $l = 1, 2, 3, 4$, denotes each branch which corresponds to the scale of the receptive field, and $h_{1,d-1}^l$ denotes output feature of the $l$-th branch. Then, features from each branch are fed to the dilated convolution layers with different dilation rates followed by ReLU activation function. By doing so, we get

$$h_{2,d-1}^l = ReLU(D_{l \times 3}^3(h_{1,d-1}^l)).$$

where $D_{3 \times 3}^l$ denotes $3 \times 3$ dilated convolution layer with dilation rate $l$. As a result, the dilated convolution layer provides various sizes of receptive fields without increasing the number of parameters and hence plays a key role in our method.

As state previously, we note that the important scale of context is spatially variant, i.e., a specific scale of the context in a region might be more informative than the other scales for reconstructing the details. To reflect such property, we adopt an attention module to focus on the scale of features which are the most informative at a given region. Specifically, the features $h_{2,d-1}^l$ are fed to the spatial attention module to generate the spatial attention map $M^l$ so that the features are further rescaled by $M^l$ to get “attention-applied features $h_{3,d-1}^l$.” In short, the final output of each branch is

$$M^l = SA(h_{2,d-1}^l)$$

$$h_{3,d-1}^l = h_{2,d-1}^l \otimes M^l.$$ 

FIGURE 2. The architecture of (a) the multi-scale attention block and (b) the spatial attention module for our MSAN, where dConv/ denotes dilated convolution with $l$ dilation rate. SA and LFF denote spatial attention module and local feature fusion, respectively.
Finally, all scale features are concatenated and fused by local feature fusion (LFF) layer $C_{LFF}()$ which works as a bottleneck layer. In addition, local residual learning (LRL) is adopted, which is proven to boost performances [4], [7], [33]. Formally, the overall output feature $F_d$ is obtained as

$$F_d = H_{d-1} + C_{LFF}(h^1_{3,d-1}, h^2_{3,d-1}, h^3_{3,d-1}, h^4_{3,d-1}).$$  \hspace{1cm} (11)

D. SPATIAL ATTENTION MODULE

For dynamic feature selection regarding the scale-significance, we adopt the spatial attention module as shown in FIGURE 2a, which is adopted from [32]. The spatial attention map is obtained as

$$M^l = \sigma(C_{7 \times 7}([\text{AvgPool}(h_2^l), \text{MaxPool}(h_0^l)])),$$  \hspace{1cm} (12)

where $C_{7 \times 7}()$ denotes a convolution layer with kernel size of 7, and $\sigma()$ denotes the sigmoid function (subscript $d - 1$ is omitted for brevity). ‘MaxPool’ and ‘AvgPool’ stand for channel-wise max pooling and average pooling operations, respectively.

E. IMPLEMENTATION DETAILS

The kernel size of the convolution layers is $3 \times 3$ unless otherwise stated. The FF layers are implemented with $1 \times 1$ convolution layer followed by ReLU. LFF and GFF are implemented by $1 \times 1$ convolution layer. For the upsamplemer $U(\cdot)$, we adopt one sub-pixel convolution layer [21] for scaling factor $\times 2$ and $\times 3$, while two sub-pixel convolution layers are adopted for $\times 4$. We set the number of features $G = 64$, and the number of MSABs $D = 16$.

For the loss function, we use the mean absolute error (MAE) which is described as

$$\mathcal{L}(\theta) = \frac{1}{N} \sum_{i=1}^{N} ||f(HR^i; \theta) - f(LR^i; \theta)||_1.$$  \hspace{1cm} (13)

IV. DISCUSSION

A. COMPARISONS TO THE RECENT SISR MODELS

a: COMPARISONS TO RCAN

RCAN [9] adopts the residual-in-residual structure and channel-attention module. Our proposed model adopts skip-connections and attention module similarly to RCAN. The skip-connection is currently a widely used scheme, but we note that the difference comes from where to make connections. We adopt LRL and GRL, where we add one more skip-connection, i.e. DRL. Also, we combine dense connection with skip-connection to boost performance. For the attention module, RCAN focuses on channel attention with global average pooling, whereas our model focuses on spatial attention to select spatially important features.

b: COMPARISONS TO CARN

CARN [18] was proposed as a lightweight structure with cascading residual blocks. The key idea of lightweight structure design lies in the construction of group convolution layers. Since our main building block MSAB processes each branch in parallel, it can be considered a cascaded group convolution. Specifically, if we alter dilated convolutions to normal convolution layers, it is just the same as a group convolution with the group size of 4. The group convolution is known to reduce parameters and computations approximately proportional to the group size, which enables the lightweight design for both CARN and our method. However, different from CARN, we adopt dilated convolution and spatial attention module, and hence, our MSAB can be viewed as an extension of a cascade of group convolution layers that can have many informative features.

c: COMPARISONS TO MSRN

MSRN [29] also adopts the multi-scale scheme like the methods mentioned above, but there is a big difference between MSRN and our MSAN. MSRN adopts the multi-scale scheme by increasing the size of kernels and bypassing intermediate features to other branches. On the other hand, our MSAN adopts the dilated convolution to increase the receptive field size, and the features are not shared with other branches but processed within a parallel structure. Eventually, our method has advantages against MSRN in terms of the number of parameters and computation. Thus, our MSAN shows comparable results to MSRN with a small number of parameters.

V. EXPERIMENTAL RESULTS

We present the results of two models: MSAN and MSAN-X. The MSAN-X has recursive MSABs, in which one MSAB is shared $D$ times. We train our networks with DIV2K [34], which is a high-quality dataset with 2 K resolution, containing 800 images for training, 100 for validation, and 100 for the test. ADAM optimizer [35] is used with the initial learning rate of $4 \times 10^{-4}$ and halved every 100,000 iterations until it reaches $5 \times 10^{-5}$. The RGB input LR training patches are generated by using MATLAB bicubic downsampling function, and the size is set to $48 \times 48$. We evaluate our models with four benchmarks: Set5 [36], Set14 [37], BSD100 [38], and Urban100 [39]. We assess our models by using PSNR and SSIM on the luminance channel of YCbCr color model. Three scaling factors are evaluated: $\times 2$, $\times 3$, and $\times 4$.

A. COMPARISONS AND RESULTS ON BENCHMARK

First, we compare our MSAN and MSAN-X with several recent lightweight CNN models under 2 M parameters, including SRCNN [1], VDSR [2], LapSRN [3], DRRN [6], MemNet [23], IDN [19], DSRN [17], CARN [18]. TABLE 1 lists the average PSNR and SSIM results on benchmarks. For the best performance, we also present MSAN+ and MSAN-X+ results with self-ensemble [40] on MSAN and MSAN-X. As shown, our MSAN outperforms other lightweight SISR methods in both PSNR and SSIM. Notably, both of our MSAN and MSAN-X show significant performance gains in Urban100 set which contains images with repetitive patterns with spatially varying scale, like FIGURE 3. As MSAN and MSAN-X are designed to focus on locally appropriate scales, they achieve large performance gains.
gain in Urban100 set; sometimes MSAN-X even outperforms CARN with a large margin. For example, on Urban100 (×2), our MSAN-X (870 K) has PSNR gain about 0.23 dB against CARN (1592 K). The comparison of lightweight models for Urban100 dataset (×2) is also graphically shown in FIGURE 6(b).

### B. COMPARISONS WITH THE STATE-OF-THE-ART METHODS

For further analysis, we compare our methods with several state-of-the-art methods that need a large number of parameters: EDSR [7], RDN [8], MSRN [29], and RCAN [9]. TABLE 2 shows the PSNR difference compared to our
TABLE 1. Benchmark SISR results. Average PSNR/SSIM values for the scaling factor × 2. Red indicates the best and blue indicates the second best. (For the choice of the best model, we exclude ‘+’ models for fair comparisons.).

| Method      | Scale | Set5 PSNR | Set5 SSIM | Set14 PSNR | Set14 SSIM | BSD100 PSNR | BSD100 SSIM | Urban100 PSNR | Urban100 SSIM |
|-------------|-------|-----------|-----------|------------|------------|-------------|--------------|---------------|---------------|
| Bicubic     | 2     | 33.66/0.9299 | 30.24/0.8688 | 29.56/0.8431 | 26.88/0.8403 |
| SRCNN [1]   | 2     | 36.66/0.9542 | 32.45/0.9067 | 31.36/0.8879 | 29.50/0.8946 |
| VDSR [2]    | 2     | 37.53/0.9587 | 33.03/0.9124 | 31.90/0.8960 | 30.76/0.9140 |
| LapSRN [3]  | 2     | 37.52/0.9590 | 33.08/0.9130 | 31.80/0.8950 | 30.41/0.9100 |
| DRRN [6]    | 2     | 37.74/0.9591 | 33.23/0.9136 | 32.05/0.8973 | 31.23/0.9188 |
| MemNet [23] | 2     | 37.78/0.9597 | 33.28/0.9142 | 32.08/0.8978 | 31.31/0.9195 |
| IDN [19]    | 2     | 37.83/0.9600 | 33.30/0.9148 | 32.08/0.8985 | 31.27/0.9196 |
| DSRN [17]   | 2     | 37.66/0.9590 | 33.15/0.9130 | 32.10/0.8970 | 30.97/0.9160 |
| CARN [18]   | 2     | 37.76/0.9590 | 33.52/0.9166 | 32.09/0.8978 | 31.51/0.9312 |
| MSAN-X (Ours) | 2     | 37.86/0.8909 | 33.52/0.9167 | 32.12/0.8983 | 31.91/0.9255 |
| MSAN (Ours)  | 2     | 37.98/0.9597 | 33.67/0.9184 | 32.21/0.8997 | 32.37/0.9298 |
| MSAN-X+ (Ours) | 2     | 37.95/0.9595 | 33.63/0.9176 | 32.17/0.8989 | 32.10/0.9271 |
| MSAN+ (Ours) | 2     | 38.05/0.9600 | 33.79/0.9192 | 32.24/0.9003 | 32.54/0.9312 |
| Bicubic     | 3     | 30.39/0.8682 | 27.55/0.7742 | 27.21/0.7385 | 24.46/0.7349 |
| SRCNN [1]   | 3     | 32.75/0.9090 | 29.30/0.8215 | 28.41/0.7863 | 26.24/0.7989 |
| VDSR [2]    | 3     | 33.66/0.9213 | 29.77/0.8314 | 28.82/0.7976 | 27.14/0.8279 |
| DRRN [6]    | 3     | 34.03/0.9244 | 29.96/0.8349 | 28.95/0.8004 | 27.53/0.8378 |
| MemNet [23] | 3     | 34.09/0.9248 | 30.00/0.8350 | 28.96/0.8001 | 27.56/0.8376 |
| IDN [19]    | 3     | 34.11/0.9253 | 29.99/0.8354 | 28.95/0.8013 | 27.42/0.8359 |
| DSRN [17]   | 3     | 33.88/0.9220 | 30.26/0.8370 | 28.81/0.7970 | 27.16/0.8280 |
| CARN [18]   | 3     | 34.29/0.9255 | 30.29/0.8407 | 29.06/0.8034 | 27.38/0.8404 |
| MSAN-X (Ours) | 3     | 34.19/0.9246 | 30.27/0.8403 | 29.03/0.8030 | 27.96/0.8473 |
| MSAN (Ours)  | 3     | 34.47/0.9265 | 30.44/0.8432 | 29.14/0.8055 | 28.36/0.8554 |
| MSAN-X+ (Ours) | 3     | 34.34/0.9258 | 30.38/0.8419 | 29.10/0.8042 | 28.12/0.8501 |
| MSAN+ (Ours) | 3     | 34.55/0.9271 | 30.50/0.8440 | 29.19/0.8063 | 28.51/0.8577 |
| Bicubic     | 4     | 28.42/0.8104 | 26.00/0.7027 | 25.96/0.6675 | 23.14/0.6577 |
| SRCNN [1]   | 4     | 30.48/0.8628 | 27.50/0.7513 | 26.90/0.7101 | 24.52/0.7221 |
| VDSR [2]    | 4     | 31.35/0.8838 | 28.01/0.7674 | 27.29/0.7251 | 25.18/0.7524 |
| LapSRN [3]  | 4     | 31.54/0.8850 | 28.19/0.7720 | 27.32/0.7280 | 25.21/0.7560 |
| DRRN [6]    | 4     | 31.68/0.8888 | 28.21/0.7720 | 27.38/0.7284 | 25.44/0.7638 |
| MemNet [23] | 4     | 31.74/0.8893 | 28.26/0.7723 | 27.40/0.7281 | 25.50/0.7630 |
| IDN [19]    | 4     | 31.82/0.8903 | 28.25/0.7730 | 27.41/0.7297 | 25.41/0.7632 |
| DSRN [17]   | 4     | 31.40/0.8830 | 28.07/0.7700 | 27.25/0.7240 | 25.08/0.7470 |
| CARN [18]   | 4     | 32.13/0.8937 | 28.60/0.7806 | 27.58/0.7349 | 26.07/0.7837 |
| MSAN-X (Ours) | 4     | 31.99/0.8909 | 28.55/0.7797 | 27.53/0.7338 | 25.91/0.7789 |
| MSAN (Ours)  | 4     | 32.26/0.8948 | 28.70/0.7834 | 27.62/0.7375 | 26.25/0.7904 |
| MSAN-X+ (Ours) | 4     | 32.16/0.8931 | 28.64/0.7815 | 27.58/0.7352 | 26.04/0.7823 |
| MSAN+ (Ours) | 4     | 32.38/0.8963 | 28.78/0.7849 | 27.67/0.7386 | 26.39/0.7935 |

MSAN with other methods, including MSAN-X, which shows that there is a trade-off between the number of parameters and accuracy. Moreover, it is notable that the trade-off appears strongly between depth and accuracy. With EDSR, to get the average PSNR gain of 0.195 dB, about 23.5 times more parameters than MSAN are required. Also, RDN
TABLE 2. APSNR comparisons with the state-of-the-art methods with scaling factor ×4.

| Methods      | Parameters | Depths | Set14 | BSD100 | Urban100 | Average |
|--------------|------------|--------|-------|--------|----------|---------|
| EDSR [7]     | 43.1 M     | 69     | 0.29  | 0.10   | 0.09     | 0.39    | 0.195  |
| RDN [8]      | 22.3 M     | 151    | 0.21  | 0.11   | 0.10     | 0.34    | 0.195  |
| MSRNN [9]    | 6.3 M      | 28     | -0.19 | -0.10  | -0.10    | -0.31   | -0.195 |
| DRRN [10]    | 15.6 M     | 415    | 0.27  | 0.17   | 0.15     | 0.57    | 0.215  |
| MSAN-X (Ours)| 1018 K     | 70     | -0.27 | -0.15  | -0.00    | -0.34   | -0.213 |
| MSAN (Ours)  | 1827 K     | 70     | -     | -      | -        | -       | -      |

TABLE 3. The number of Mult-Adds compared to recent lightweight models. We only present the case of ×4.

| Methods      | DRRN [6] | MemNet [23] | CARN [18] | MSAN (Ours) |
|--------------|----------|--------------|------------|-------------|
| Parameters   | 297 K    | 677 K        | 1592 K     | 1827 K      |
| Mult-Adds    | 6796.9 G | 2662.4 G     | 90.9 G     | 148.3 G     |

C. ANALYSIS OF COMPUTATIONAL COMPLEXITY

Since the number of operations is more closely related to the computational complexity than the number of parameters, we analyze the number of operations in terms of Mult-Adds in TABLE 3. We directly follow the same setting and the method from CARN [18] to calculate the number of Mult-Adds. Like the method in [18], we assume the size of HR image to be 1280 × 720. The table shows that our MSAN uses comparable parameters and computational costs to CARN [18]. Compared to other lightweight models such as DRRN and MemNet [6], [23], MSAN requires much fewer Mult-Adds while using more parameters. In conclusion, our MSAN has advantages in terms of both the number of parameters and the number of operations.

D. VISUALIZED RESULTS

For qualitative comparisons, we visualize some of super-resolved results on FIGURE 3, 4, 5. As shown, MSAN predicts multi-scale high-frequency details better than other methods.

VI. CONCLUSION

In this paper, we have proposed a new lightweight CNN-based SISR method by designing new schemes that can extract multi-scale spatial features. To achieve the performance gain with a limited number of parameters, we focused on the quality of intermediate features. For this, we designed a building block that addresses multi-scale receptive fields, which enables to capture local to global contextual information, and also reflect multi-scale self-similarity. Moreover, we adopt the spatial attention module to focus on the region-adaptive informative spatial features. It is shown that the dense connection, along with the skip connection, boosts the performance. In our extensive experiments, we have demonstrated that the proposed networks show competitive performance against other recent CNN-based methods. Our codes are publicly available at https://github.com/JWSoh/MSAN.

REFERENCES
T. Dai, J. Cai, Y. Zhang, S.-T. Xia, and L. Zhang, “Second-order attention networks for single image super-resolution,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 8122–8131.

J. W. Soh, N. I. Cho: Lightweight SISR With MSAN

Y. Tai, J. Yang, X. Liu, and C. Xu, “MemNet: A persistent memory network for image super-resolution via information distillation network,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 723–731.

Z. Hui, X. Wang, and X. Gao, “Fast, accurate, and lightweight super-resolution with cascading residual network,” in Proc. Eur. Conf. Comput. Vis., 2018, pp. 252–268.

J. Pan, S. Liu, D. Sun, J. Zhang, Y. Liu, J. Ren, Z. Li, J. Tang, H. Lu, Y.-W. Tai, and M.-H. Yang, “Learning dual convolutional neural networks for low-level vision,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 3070–3079.

C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2015, pp. 1–9.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in Proc. Adv. Neural Inf. Process. Syst., 2014, pp. 2672–2680.

J. Han, S. Liu, D. Sun, J. Zhang, Y. Liu, J. Ren, Z. Li, J. Tang, H. Lu, Y.-W. Tai, and M.-H. Yang, “Learning deep convolutional neural networks for low-level vision,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2015, pp. 3070–3079.

C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2015, pp. 1–9.

J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 7132–7141.