SCAN: Spatial Color Attention Networks for Real Single Image Super-Resolution

Xuan Xu and Xin Li
Lane Department of Computer Science and Electrical Engineering,
West Virginia University, Morgantown, WV 26506, USA
xuxu@mix.wvu.edu xin.li@mail.wvu.edu

Abstract

Conceptually similar to adaptation in model-based approaches, attention *has received increasing more attention in deep learning recently. As a tool to reallocate limited computational resources based on the importance of informative components, attention mechanism has found successful applications in both high-level and low-level vision tasks which includes channel attention, spatial attention, non-local attention and etc. However, to the best of our knowledge, attention mechanism has not been studied for the R,G,B channels of* color *images in the open literature. In this paper, we propose a spatial color attention networks (SCAN) designed to jointly exploit the spatial and spectral dependency within color images. More specifically, we present a spatial color attention module that calibrates important color information for individual color components from output feature maps of residual groups. When compared against previous state-of-the-art method Residual Channel Attention Networks (RCAN), SCAN has achieved superior performance in terms of both subjective and objective qualities on the dataset provided by NTIRE2019 real single image super-resolution challenge.*

1. Introduction

Attention mechanism, originally inspired by the behavior and the neuronal architecture of primate visual systems [15, 14], has received increasingly more attention by computer vision and machine learning communities. Since the breakthrough in machine translation application [32], attention has been found to be useful to many high-level vision tasks including image captioning [5, 38], lip reading[6], image classification [34, 11, 35] and image understanding [4, 16]. The success of attention mechanism is generally attributed to prioritize the allocation of available processing resources towards the most informative components (e.g., salient regions) in an image.

By contrast, attention mechanism has been under-researched for low-level vision tasks. The only few exceptions all deal with single image super-resolution (SISR) (e.g., channel attention [39], channel-spatial attention [12], non-local attention [40]). The common theme behind so-called spatial or channel attention mechanism is to adaptively rescale each spatial-domain or channel-wise feature by modeling their interdependency, that will help networks pay more attention to specific features. Such attention mechanism allows a network to concentrate its computational resources on the most useful features and enhance the discriminative learning ability.

However, existing study about attention mechanism has not been extended for color images or across spectral bands to the best of our knowledge. The only studies about color attention we can find are [18, 19] which have focused on the application of object recognition for high-level vision tasks. The issue of how to jointly exploit spatial and spectral dependencies [8] for low-level vision tasks such as SISR seems to have not been addressed in the open literature. All previous attention strategies for SISR have only considered to directly use R,G,B color channels as input training data. In other words, the networks will simply treat all the color information among R,G,B channels equally. One potential risk of this strategy is the lack of optimization - e.g., exploiting the spectral dependency among color channels might benefit the task of deep residual learning.

In this paper, we propose to address the above issue by developing a new architecture named Spatial Color Attention Networks (SCAN). Conceptually similar to bilateral filtering [30] in which spatial and color are treated as two independent domains, we treat spatial and color features as two complementary channels. So instead of considering channel-wise and spatial feature modulation in [12], we have developed a spatial color attention module (SCAM) to calibrate important color information from output feature maps of residual groups. Unlike existing works which treat channel-wise features across spectral bands equally, we propose to make the networks focus on informative fea-

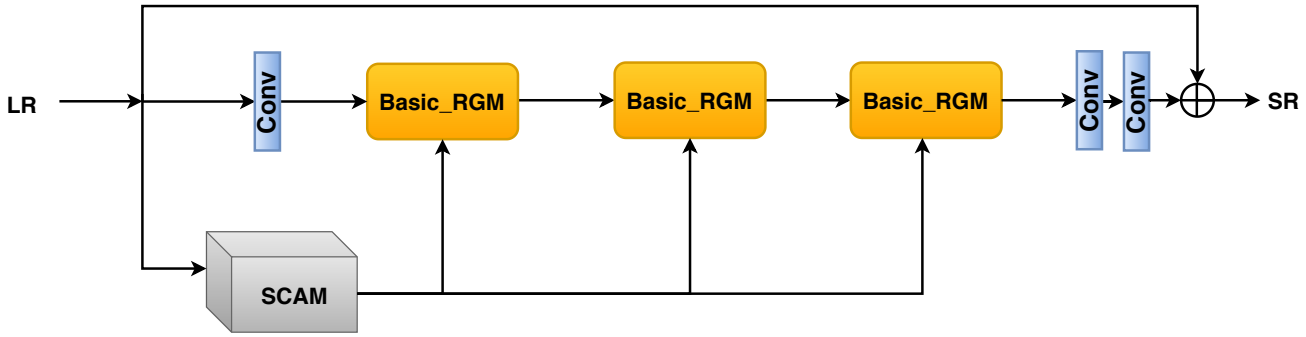

Figure 1. Overview of proposed networks architecture, Basic_RGM stands for the basic residual group module which includes several residual groups, SCAM is proposed spatial color attention module where to generate R,G,B spatial color attention map, $\oplus$ denotes element-wise sum.

tures and exploit interdependencies among color channels. The newly developed color attention mechanism enables the network to not only focus on recovering spatially high frequency components (e.g., edges and textures) but also pay attention to vivid and sharp color information (e.g., colorful flowers and texts) in the generated HR image.

A summary of our key contributions include:

• We propose to address the issue of *color attention* for SISR and demonstrate it is supplementary to the spatial and channel attention mechanisms studied in the literature;

• Our proposed spatial color attention module (SCAM) and residual channel-spatial attention (RCSA) can be easily integrated to most existing SISR networks;

• Experimental results have shown our SCAN can significantly outperform previous state-of-art RCAN [39] on real SISR competition dataset.

2. Related Works

Deep learning-based approaches toward single image super-resolution (SISR) have shown the reliability and advantages compared with the traditional model-based methods. SRCNN [7] first introduced a simple three layers CNN architecture to solve SISR problems; VDSR [20] utilized the concept of deep residual networks [10] to make the deeper networks (20 layers) trainable and significantly improve the results; LapSRN [22] proposed to upscale low resolution image by a pyramid structure which has a better performance on large scale factors (ex. $8\times$). EDSR [24] introduced to use residual blocks without batch-norm layer and get the significant improvement. Most recent advances include deep recursive residual network (DRRN)[28], SR-DenseNet [31] and Residual Dense Network (RDN) [41] combined the state-of-art deep learning approaches ResNet [10] and DenseNet [13] to further improve SISR performance.

Inspired by [11], Residual Channel Attention Networks (RCAN) [39] first considered attention mechanism - channel attention to improve the representational ability of the network and get the state-of-art performance with a very deep networks. Besides objective measures such as PSNR/SSIM [37], SRGAN [23] introduced a novel generative adversarial networks (GAN) [9] based architecture to optimize the perceptual quality of SR images. An enhanced version of SRGAN named ESRGAN [36] using relativistic average GAN (RaGAN) was developed in [17] as well as [33], which demonstrated improved visual quality than standard GAN.

It is worth highlighting previous works on attention mechanism in the existing literature. Generally speaking, the common principle underlying various attention mechanisms is to bias limited computational resources based on the importance of informative components. For example, channel attention [39] adaptively rescale the channel-wise feature by modeling their interdependency; channel-spatial attention addresses the issue of channel-wise and spatial feature modulation [12]; non-local attention [40] attempts to simultaneously exploit the local and non-local dependency within an image for the task of image restoration. To the best of our knowledge, the issue of color attention - i.e., the modeling of interdependency across different spectral channels - had not been studied in the open literature. Therefore, we propose to address this issue and develop specially tailored modules for color image restoration.

3. Proposed Approach

3.1. Network Design

We present the designed networks in the following hierarchy: SCAN (Fig. 1) $\rightarrow$ Subnetwork of SCAM and Basic_RGM (Fig. 2) $\rightarrow$ Residual Channel-Spatial Attention (RCSA, Fig. 3). It should be noted that our SCAN and previous state-of-the-art RCAN [39] are similar at the coarsest level. Both SCAN and RCAN are decomposed of the residual group (RG) and residual channel attention block (RCAB). This is because we want to evaluate the validity of proposed SCAM (refer to Sec. 3.2) and RCSA (refer to

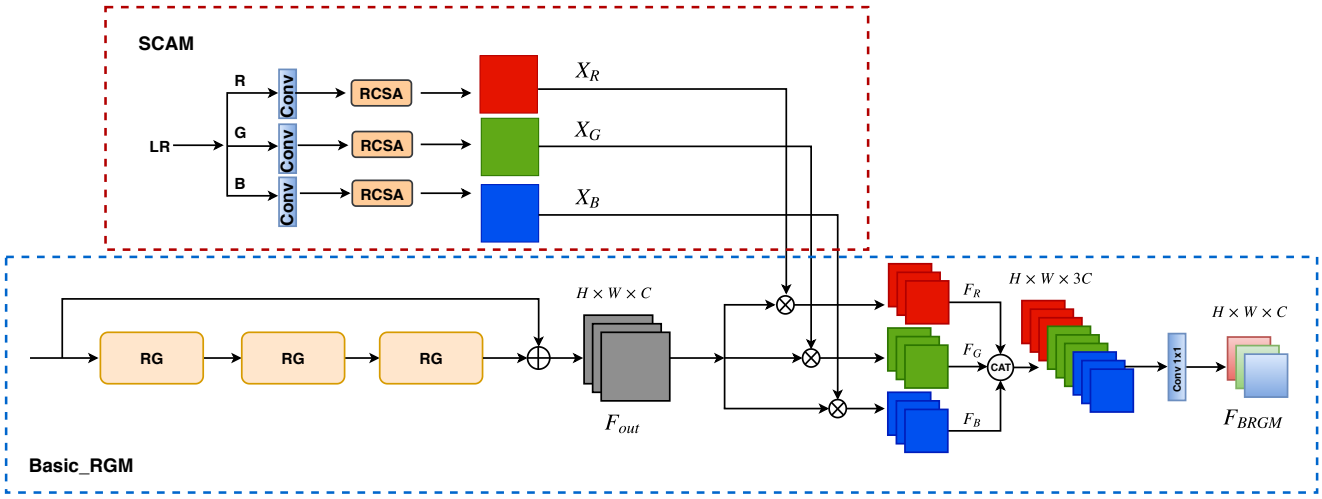

Figure 2. The structure of proposed Basic_RGM and SCAM modules. In the block of SCAM, RCSA stands for proposed residual channel-spatial attention module (the details are demonstrated in Fig. 3); $\oplus$ denotes element-wise sum, $\otimes$ denotes element-wise product, CAT denotes feature-concatenation.

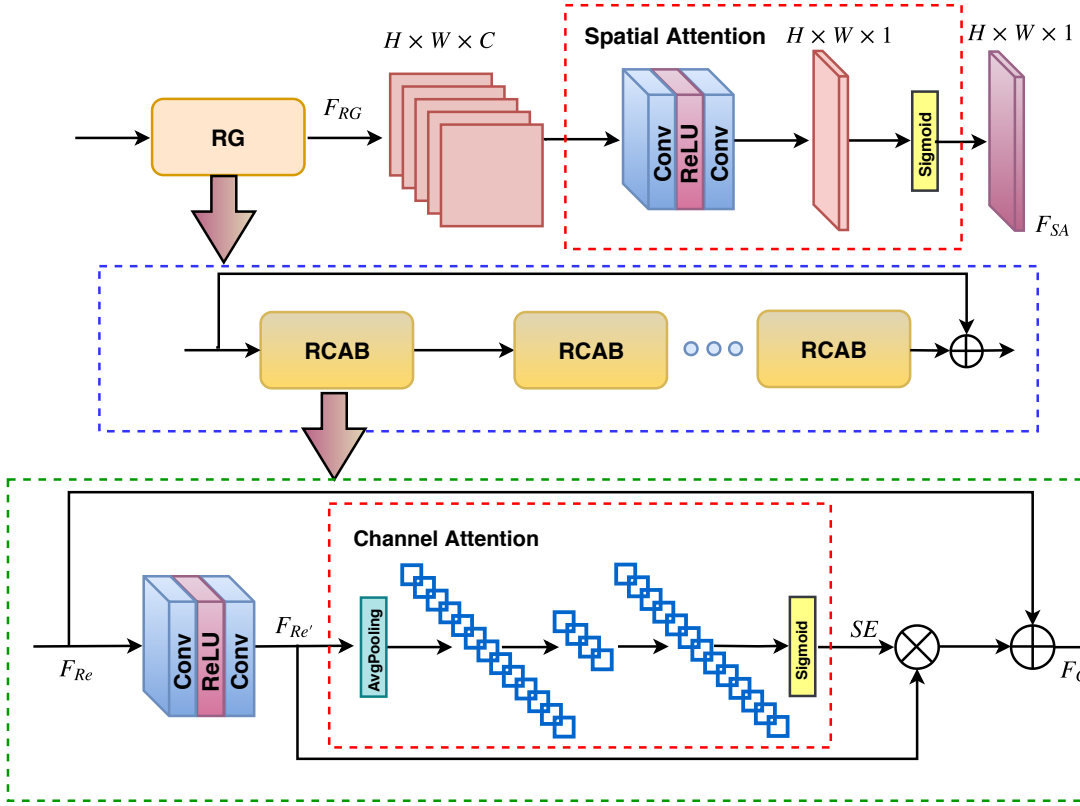

Figure 3. The structure of proposed RCSA module which includes the implementation of RG and RCAB blocks. The two red blocks show the structures of channel and spatial attentions. $\otimes$ denotes element-wise product and $\oplus$ denotes element-wise sum.

Sec. 3.3) modules under the same conceptual framework.

We are hoping that this way of presentation can facilitate our explanation about why SCAN can outperform RCAN (similar to the popular ablation study) - i.e., without changing the overall structure, we can still improve the performance of SISR by designing novel fine-scale modules (SCAM and RCSA) in a plug-and-play fashion (e.g., the number of Basic_RGM modules in Fig. 1 can be reduced as we will show in our ablation study in Table 2). Meanwhile, we will focus on the key difference be-

tween the design of RCAN and SCAN - i.e., the desirable color attention mechanism. Across different hierarchies (SCAN→SCAM→RCSA), we will show how color attention mechanism is the theme unifying our network design and optimizing the task of deep residual learning.

3.2. Spatial Color Attention Module (SCAM)

In SCAM module (see Fig. 2), we organize the input training data (LR images) into two parts: 1) similar to the normal SISR architectures, the whole LR image is supplied as the input to the main network (see Basic_RGM block in Fig. 2; 2) the LR image is divided to R,G,B channels separately, which then serve as the input to the proposed RCSA module for generating spatial color attention maps X_R, X_G, X_B, for R,G,B channels respectively. Note that the second part (our new contribution) is absent in previous works on SISR because they treat all R,G,B channels equally.

Let F_{out} denote the output feature maps of the Basic_RGM block (see Fig. 2, the gray-colored feature map with the dimension of $H \times W$ that contains C feature maps), which is generated from the whole LR input image and fused all R,G,B information into each feature map. To re-calibrate F_{out}, we apply element-wise product between each spatial color attention map X_R, X_G, X_B and F_{out}. The process of introducing color attention can be expressed as follows:

$$F_R = F_{out} \cdot X_R \tag{1}$$
$$F_G = F_{out} \cdot X_G \tag{2}$$
$$F_B = F_{out} \cdot X_B \tag{3}$$

where F_R, F_G, F_B are the re-calibrated feature maps from F_{out} to represent spatial color information for each R,G,B channel (e.g., F_R represents the spatial information from red channel).

Next, to get the final output feature-map F_{BRGM} with the dimension of $H \times W \times C$, we first concatenate R,G,B feature maps and then use a 1×1 Conv layer to reduce the feature-map dimension from $3C$ to C:

$$F_{BRGM} = \mathbf{W}_D([F_R, F_G, F_B]) \tag{4}$$

where $\mathbf{W}_D \in \mathbb{R}^{1\times1\times C}$ is a 1×1 Conv layer used for dimensionality reduction.

By applying SCAM to the basic_RGM, the networks can fuse channel attention (already considered in RCAN [39]) and spatial color attention (new module introduced by this work) to better re-calibrate input feature maps based on the pair of training data. Note that the real SISR challenge still belongs to strongly supervised learning; therefore the objective here is the same as the original idea of applying ResNet [20] to SISR (i.e. to learn a more accurate residual representation). The new insight we attempt to bring through this work is that residual representations across spectral channels are not independent, which implies the potential of jointly learning them (as we will elaborate next).

3.3. Residual Channel-Spatial Attention (RCSA)

To implement RCSA module (see Fig. 3), we have followed the basic structure of SENet[11] and RCAN[39] which sets up a regular residual block including channel attention mechanism. More specifically, we first squeeze input feature maps with global average pooling:

$$Q_C = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} F_C(i, j) \tag{5}$$

where C is the number of feature maps, Q_C is the c-th element of $Q \in \mathbb{R}^C$, $F_C(i, j)$ is the pixel value of the c-th feature at position (i, j) from input feature maps $F_{Re'} \in \mathbb{R}^{H\times W\times C}$. Then we propose to implement a simple gating mechanism as adopted by previous works including SENet [11] and RCAN [39]:

$$SE = \sigma(\mathbf{W}_E(\delta(\mathbf{W}_S(Q)))) \tag{6}$$

where σ refers to a sigmoid function, δ denotes the ReLU function, $\mathbf{W}_S \in \mathbb{R}^{1\times1\times\frac{C}{r}}$ is the *squeeze* Conv layers with weights and $\mathbf{W}_E \in \mathbb{R}^{1\times1\times C}$ is the *expand* Conv layers with weights, r is the reduction ratio to reduce the dimension of Q (the parameter r controls the trade-off between the capacity and the complexity [11]). Finally, we can rescale the feature maps F_{Re} by:

$$F_{CA} = SE \cdot F_{Re'} + F_{Re} \tag{7}$$

where F_{Re} is the input feature map to RCAB block, F_{CA} is the output from RCAB block which is a rescaled feature maps by channel attention module (see Fig. 3). Note that Eq. (7) is different from previous works such as RCAN because we will have a separate channel/spatial attention mechanism for each R,G,B channel respectively.

Next, we can apply spatial attention to the rescaled feature map. Unlike previous works in which R,G,B feature maps are treated equally, we note that the input of RCSA is a single channel of RGB image. Therefore our approach generates the output feature map (so-called spatial color attention map) which focus on re-calibrating single color channel information from the feature maps of F_{out}. In this fashion, the outputs of SCAM module naturally fit the grey-colored feature map in Fig. 2 (refer to section 3.2). More specifically, we have

$$F_{SA} = \sigma(\mathbf{W}_{SA}(\delta(\mathbf{W}_{SA}(F_{RG})) \tag{8}$$

where F_{SA} is the output spatial color attention feature map which can be represented as X_R, X_G, X_B based on the corresponding R,G,B channels, $\mathbf{W}_{SA} \in \mathbb{R}^{1\times1\times C}$ is the Conv

layer with weight, σ refers to a sigmoid function, δ denotes the ReLU function. F_{RG} is the output of RG (refer to Fig. 3). In summary, newly designed spatial color attention map is expected to more effectively learn the joint residual representations across spectral channels.

4. Experiments

4.1. Dataset

In this work, we have used the real-world paired image dataset provided by NTIRE2019 challenge. It includes 60 pairs of images for training, 20 pairs of images for validation and another 20 pairs of images for testing; both HR and LR images are collected by standard DSLR cameras, which means the LR image is not synthetic but captured from the real-world (likely with a different focal length). This is in sharp contrast with previous SISR challenges which generate LR images from HR images (e.g., DIV2K [1]) using model-based methods (e.g., bicubic interpolation). The new real-world dataset is arguably more closely related to the real-world SISR tasks - e.g., the scaling factors between LR and HR images are unknown (in theory it is determined by the ratio of focal lengths).

We also note that HR images for test data (i.e., the ground-truth) is not provided; therefore LR images in test data have already been scaled to the same size/resolution as the corresponding HR images by the competition organizer. Accordingly, we have opted to report our PSNR/SSIM experimental results based on 20 paired validation data (for which ground-truth is available) and report perceptual index (PI) score [3] based on 20 test data since PI is a no-reference image quality metric (no HR image is needed).

4.2. Training

In our proposed SCAN networks, we have set Basic_RGM to 3, each Basic_RGM includes 3 residual groups (RG). And every RG contains 20 RCAB blocks which is the same as the original RCAN [39]. In RCSA module, we have used one RG with 6 RCAB blocks inside. Most of kernel size of Conv layers are 3×3 with 64 filters ($C = 64$) except few exceptions as shown in Fig. 3 (e.g., in the spatial attention block, the two Conv layers have only 1 filter that means $C = 1$, and 1×1 of kernel size). In channel attention block, the reduction ratio is $r = 16$. The last layer filter of the whole networks is set to be 3 in order to output super-resolved color images. Note that the original RCAN has 10 RGs; due to the limitation of GPU memory, we have only adopted 9 RGs in our current implementation of SCAN (3 Basic_RGM, totally amount to 9 RGs).

In our training process, we first randomly crop both the input and ground-truth RGB images with small patches such as 128×128, with a batch size of 16; then we augment the training set by standard geometric transformations (e.g., flipping and rotation). Our model is trained and optimized by ADAM [21] with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$. The initial learning rate is set to 1×10^{-4}, the decay factor is set to 5, which decreases the learning rate by half after $[384k, 576k, 768k, 883k, 998k]$ steps; the MSE loss function is applied to minimize the error between HR and SR images. All reported experiment results are trained by PyTorch [27] on 4 NVIDIA TITAN Xp GPUs. The total training time is around 35 hours.

Patch size	48×48	96×96	128×128
PSNR	29.37	29.55	29.59

Table 1. The influence of different cropped patch-size used (48×48, 96×96 and 128×128) for training process.

4.3. Effect of Patch Size for Training

We explored the effect of different patch-size cropped for the model training. Table 1 shows the results of 48×48, 96×96 and 128×128 patch-size used. Note that all the training settings are exactly same besides the patch size of training data. From the results we find that large patch size leads a better PSNR performance. However, due to the constraint with limited GPU memory, 128×128 is the largest patch size we can train at this point.

4.4. Ablation Study

In order to better illustrate the benefit of spatial color attention map step by step, we have compared different strategies to evaluate the validity of proposed SCAM. We have implemented four competing models in our experiments (trained by the same dataset): 1) baseline RCAN [39]: training without SCAM (all settings follow the original RCAN); 2) SCAN_1: training with only one-time calibration with SCAM (one Basic_RGM with 9 RGs inside); 3) SCAN_2: training with two-times calibration with SCAM (two Basic_RGM, first one has 5 RGs and the second one has 4 RGs); 4) SCAN_3: training with three-times calibration with SCAM (the proposed SCAN, please refer to Fig. 1).

Table. 2 shows the results of the four strategies mentioned above. Without SCAM, the RCAN can achieve the average PSNR of 29.31 dB; after adding SCAM, SCAN_1 can improve the initial PSNR results to 29.49 dB (**0.18 dB** gained when compared with RCAN); keep increasing calibration time to 2 and 3, we observe that the PSNR results are further improved. Finally we have achieved the best PSNR result of 29.59 dB with the proposed SCAN (i.e., SCAN_3 in Table 2).

4.5. Comparison Against State-of-the-Art

We have compared our proposed SCAN with current state-of-art SISR approach RCAN [39]. The original RCAN is trained to super-resolve LR image by a specific

Figure 4. Visual results for validation data "cam1_07" and "cam2_09".

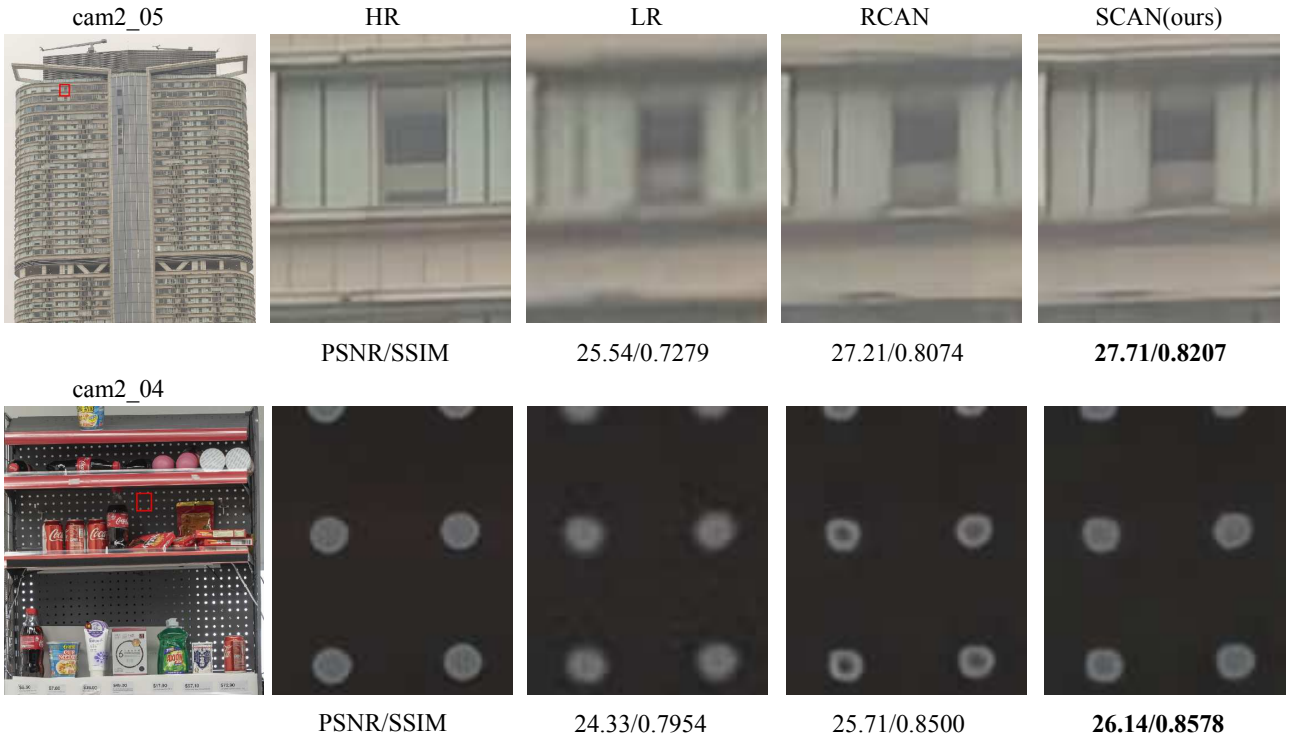

Figure 5. Visual results for validation data "cam2_05" and "cam2_04".

scale factor (i.e., 2×, 3×), but the LR and HR image in the new real-world dataset from NTIRE2019 have the same size. Therefore, we have to remove the upscale module from the original RCAN to make sure both the input and the output have the same size. The comparison results in term of PSNR is shown in Table 3.

The baseline result is the average PSNR between the (scaled) LR images and the corresponding HR images. The "+" in RCAN+ and SCAN+ stands for self-ensemble strategy used to further improve results (similar strategies have been adopted in previous works [24, 29, 41, 39]). From Table 3, our proposed SCAN and SCAN+ have the best

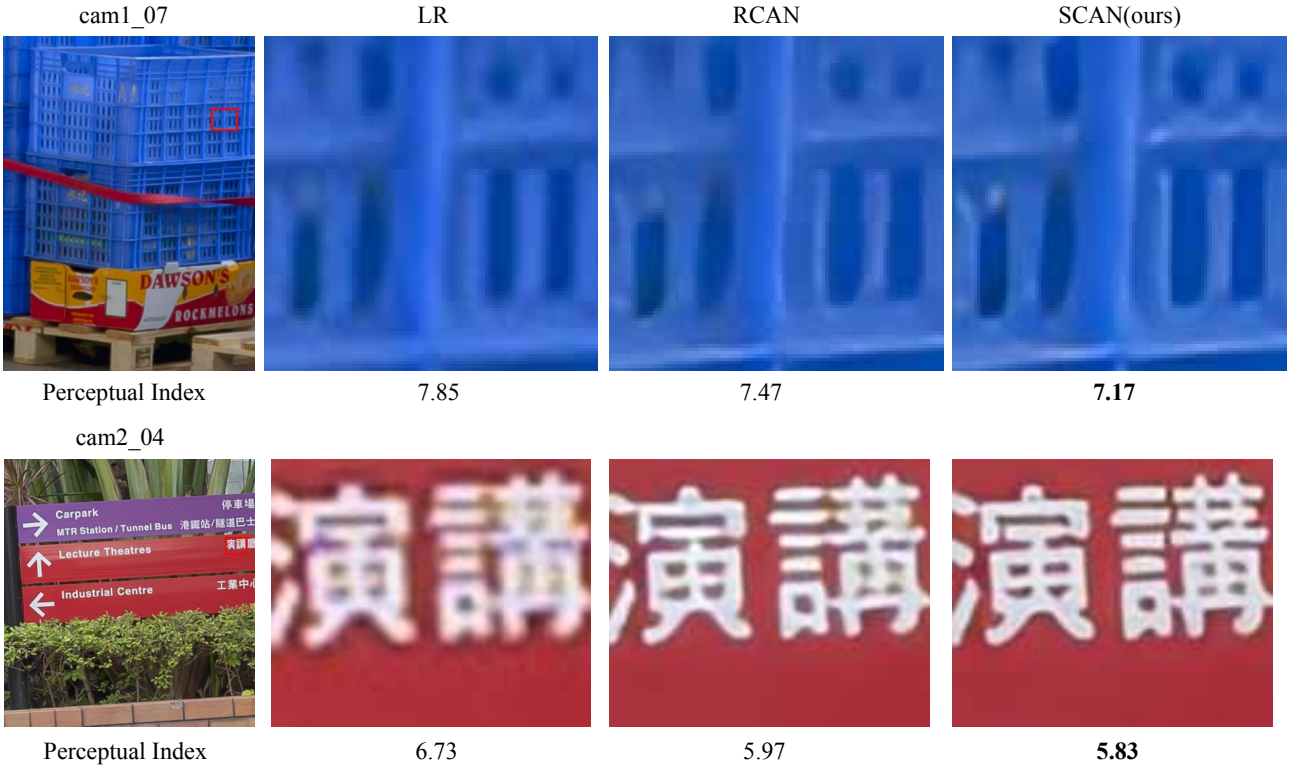

Figure 6. The visual results for test data "cam1_07" and "cam2_04". The results is based on perceptual index (PI) score since the HR image is not available. The lower PI score indicates the better perceptual quality.

Method	No. of RGs	No. of SCAM calibration used	PSNR	SSIM
RCAN	10	N/A	29.31	0.8606
SCAN_1	9	1	29.49	0.8628
SCAN_2	9	2	29.52	0.8641
SCAN_3	9	3	29.59	0.8650

Table 2. Investigations of how to set spatial color attention modules (SCAM) .

	Baseline	RCAN	RCAN+	SCAN	SCAN+
PSNR	27.78	29.31	29.42	<u>29.59</u>	**29.75**
SSIM	0.8163	0.8606	0.8632	<u>0.8650</u>	**0.8687**

Table 3. Quantitative results of PSNR and SSIM for all methods. The higher is better. **Bold** font indicates the best result and underline indicates the second.

PSNR/SSIM performance. When compared with RCAN and RCAN+, our proposed SCAN+ can significantly improve the PSNR performance by as much as 0.44 dB and 0.33 dB respectively. Even without activating the strategy of self-ensemble, SCAN is still noticeably better than RCAN and RCAN+.

Beside quantitative PSNR/SSIM results, we have also included the subjective quality results comparison in Fig. 4 and Fig. 5. For image "cam2_09" in Fig. 4, we can see that RCAN suffers from severe edge blurring artifacts and text color distortions. Our proposed SCAN can reconstruct colorful texts with fewer blurring artifacts and less color distortion. For image "cam1_07" in Fig. 4, our SCAN is capable of recovering more edge details than RCAN (e.g., the sharpness of wall-pattern). For another image "cam2_05" in Fig. 5 (note that this example is really challenging - even ground-truth HR image has suffered a little bit of edge blurring), our SCAN can reconstruct the large-scale building structure details much better (e.g., the horizontal roof structure above the window and vertical edges on both sides of the window). For image "cam2_04" in Fig. 5, we can see that the dots in ground-truth image contain solid color; while in RCAN reconstructed image, they become hollow dots. One possible interpretation is that for fine structures like small dots, it takes both spatial and color attention mechanism to ensure the structural consistency among them. By contrast, our SCAN can still faithfully reconstruct

	Baseline	RCAN	SCAN
PI Score	7.36	6.79	**6.68**

Table 4. Quantitative results of perceptual index scores for all methods. The lower score is better. **Bold** font indicates the best result.

those solid dots.

Finally, because the HR images for test data are not released, we cannot report the PSNR based results on test data. Alternatively, to evaluate the quantitative results among our method, baseline and RCAN on test data, we have used a new objective metric called Perceptual Index (PI) [3] (a no-reference image quality metric) which was recently developed to measure perceptual quality for SISR (e.g., the 2018 PIRM Challenge [2]). The PI score is defined by

$$\text{PI} = \frac{1}{2}((10 - \text{MA}) + \text{NIQE}) \tag{9}$$

where MA denotes a no-reference quality metric [25] and NIQE refers to Naturalness Image Quality Evaluator [26]. Unlike PSNR or SSIM [37], the lower PI score, the better perceptual quality.

Table 4 includes the PI comparison between ours and other competing methods. SCAN reaches the lowest PI score, which implies the highest perceptual quality. Fig. 6 includes the PI comparison among baseline (LR), RCAN and SCAN on images "cam1_07" and "cam2_04" from test dataset (HR images are not released so we will not be able to evaluate the fidelity or accuracy of SR reconstruction). But it can still be observed that SCAN is capable of delivering the most visually pleasant reconstruction of fine-detailed structures in basket on image "cam1_07". On another image "cam2_04", our proposed SCAN can significantly reduce the blurring of white-colored texts when compared with RCAN.

5. Conclusion

In this paper, we proposed a spatial color attention networks (SCAN) to tackle the problem of single image super-resolution based on real-world image dataset from NTIRE2019 challenge. The newly designed spatial color attention module (SCAM) can enable the networks to learn the joint representations across spectral channels and better calibrate the feature maps with R,G,B spatial color attention maps. When compared with start-of-the-art RCAN, our method SCAN can significantly improve both objective (including PSNR/SSIM/PI) and subjective results. Meantime, the designed SCAM module can easily be integrated with other existing super-resolution networks. Under the framework of NTIRE challenge, one issue that remains to be addressed is the modeling/learning of real-world degradation (the forward process). We expect that exploiting a priori information about the degradation process can offer new insight to the problem of real SISR.

References

[1] Eirikur Agustsson and Radu Timofte. NTIRE 2017 challenge on single image super-resolution: Dataset and study. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017. 5

[2] Yochai Blau, Roey Mechrez, Radu Timofte, Tomer Michaeli, and Lihi Zelnik-Manor. 2018 PIRM challenge on perceptual image super-resolution. *arXiv preprint arXiv:1809.07517*, 2018. 8

[3] Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 5, 8

[4] Chunshui Cao, Xianming Liu, Yi Yang, Yinan Yu, Jiang Wang, Zilei Wang, Yongzhen Huang, Liang Wang, Chang Huang, Wei Xu, et al. Look and think twice: Capturing top-down visual attention with feedback convolutional neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2956–2964, 2015. 1

[5] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua. SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5659–5667, 2017. 1

[6] Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Lip reading sentences in the wild. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3444–3453. IEEE, 2017. 1

[7] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *European conference on computer vision*, pages 184–199. Springer, 2014. 2

[8] J-M Geusebroek, Rein Van den Boomgaard, Arnold W. M. Smeulders, and Hugo Geerts. Color invariance. *IEEE Transactions on Pattern analysis and machine intelligence*, 23(12):1338–1350, 2001. 1

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 2

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[11] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1, 2, 4

[12] Yanting Hu, Jie Li, Yuanfei Huang, and Xinbo Gao. Channel-wise and spatial feature modulation network for single image super-resolution. *arXiv preprint arXiv:1809.11130*, 2018. 1, 2

[13] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 2

[14] Laurent Itti and Christof Koch. Computational modelling of visual attention. *Nature reviews neuroscience*, 2(3):194, 2001. 1

[15] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (11):1254–1259, 1998. 1

[16] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *Advances in neural information processing systems*, pages 2017–2025, 2015. 1

[17] Alexia Jolicoeur-Martineau. The relativistic discriminator: a key element missing from standard GAN. *arXiv preprint arXiv:1807.00734*, 2018. 2

[18] Fahad Shahbaz Khan, Joost Van De Weijer, and Maria Vanrell. Top-down color attention for object recognition. In *2009 IEEE 12th International Conference on Computer Vision*, pages 979–986. IEEE, 2009. 1

[19] Fahad Shahbaz Khan, Joost Van de Weijer, and Maria Vanrell. Modulating shape features by color attention for object recognition. *International Journal of Computer Vision*, 98(1):49–64, 2012. 1

[20] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1646–1654, 2016. 2, 4

[21] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[22] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate superresolution. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, page 5, 2017. 2

[23] 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*, volume 2, page 4, 2017. 2

[24] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017. 2, 6

[25] Chao Ma, Chih-Yuan Yang, Xiaokang Yang, and Ming-Hsuan Yang. Learning a no-reference quality metric for single-image super-resolution. *Computer Vision and Image Understanding*, 158:1–16, 2017. 8

[26] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a" completely blind" image quality analyzer. *IEEE Signal Process. Lett.*, 20(3):209–212, 2013. 8

[27] 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. In *NIPS-W*, 2017. 5

[28] Ying Tai, Jian Yang, and Xiaoming Liu. Image super-resolution via deep recursive residual network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, page 5, 2017. 2

[29] Radu Timofte, Rasmus Rothe, and Luc Van Gool. Seven ways to improve example-based single image super resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1865–1873, 2016. 6

[30] Carlo Tomasi and Roberto Manduchi. Bilateral filtering for gray and color images. In *ICCV*, volume 98, page 2, 1998. 1

[31] Tong Tong, Gen Li, Xiejie Liu, and Qinquan Gao. Image super-resolution using dense skip connections. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 4809–4817. IEEE, 2017. 2

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017. 1

[33] Thang Vu, Tung M Luu, and Chang D Yoo. Perception-enhanced image super-resolution via relativistic generative adversarial networks. In *European Conference on Computer Vision*, pages 98–113. Springer, 2018. 2

[34] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3156–3164, 2017. 1

[35] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7794–7803, 2018. 1

[36] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. ESRGAN: Enhanced super-resolution generative adversarial networks. In *The European Conference on Computer Vision Workshops (ECCVW)*, September 2018. 2

[37] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 2, 8

[38] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015. 1

[39] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *The European Conference on Computer Vision (ECCV)*, September 2018. 1, 2, 4, 5, 6

[40] Yulun Zhang, Kunpeng Li, Kai Li, Bineng Zhong, and Yun Fu. Residual non-local attention networks for image restoration. *arXiv preprint arXiv:1903.10082*, 2019. 1, 2

[41] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 6