IGARSS 2021 - 2021 IEEE International Geoscience and Remote Sensing Symposium | 978-1-6654-0369-6/21/$31.00 ©2021 IEEE | DOI: 10.1109/IGARSS47720.2021.9553766

SELF-ATTENTION FUSION MODULE FOR SINGLE REMOTE SENSING IMAGE SUPER-RESOLUTION

Han Mei, Haopeng Zhang, Zhiguo Jiang*

Department of Aerospace Information Engineering, School of Astronautics,
Beihang University, Beijing 102206, China
Beijing Key Laboratory of Digital Media, Beijing 102206, China
Key Laboratory of Spacecraft Design Optimization and Dynamic Simulation Technologies,
Ministry of Education, Beijing 102206, China

ABSTRACT

Single image super-resolution (SISR) is an important procedure to improve many remote sensing applications. Global features play an important role in pixel generation of SISR. In this paper, we proposed a self-attention fusion module named as SAF module which combines spatial attention and channel attention in parallel to handle this problem. Our self-attention fusion module can be flexibly added to many popular deep-learning-based SISR models to further improve their representation ability and learn global features. Experiments on UC Merced dataset indicate that SAF module can improve the performance of classic SISR models and achieve state-of-the-art super-resolution results.

Index Terms— super-resolution, spatial attention, channel attention, remote sensing images

1. INTRODUCTION

Single image super-resolution (SISR) is to reconstruct a high-resolution (HR) image with only a single low-resolution (LR) image and the convenience of its application has attracted more and more attention in remote sensing images. With the developing of convolution neural networks (CNN), SR methods based on deep learning, such as SRCNN [1], FSRCNN [2], SRResNet [3], Cycle-CNN [4], etc., have demonstrated superior performance than conventional interpolation-based methods [5].

All CNN-based methods include a large number of convolution operations. These models rely heavily on convolution operations to build relationships between different image regions. Since the convolution operation has a fixed receptive field, long-range dependencies cannot be learned by CNN-based models. Therefore, these CNN-based SISR models are difficult to learn global features so that the visual effect of the reconstructed image is limited. Spatial attention mechanism can improve these models' ability to learn global features by establishing long-range dependencies between different image regions, and has been widely used in image generation tasks. In addition to global features, the indiscriminate processing of different feature channels by the CNN structure also limits the CNN-based SISR models' representation ability. Channel attention mechanism can consider the interdependence between channels by attaching different weights to each channel which have been approved helpful to SISR. For example, [6] proposed a deep residual channel attention networks and [7] proposed a second-order attention network for SISR.

Self-attention is an attention structure with the least information loss and structure-friendly, and can solve CNN-based models' feature expression problem effectively. It has been successfully applied in many low-level vision tasks, such as image synthesis [8] and image reconstruction [9]. Different from these vision tasks, image super-resolution task requires the fusion of spatial attention and channel attention to improve the visual effect of the reconstructed image essentially. Considering that self-attention has outstanding advantages in achieving both types of attention, in this paper, we proposed a self-attention fusion (SAF) module which combines spatial attention and channel attention for remote sensing image super-resolution. Both attention mechanisms are implemented by the self-attention structure. We combined SAF module with several classic CNN-based SISR models, and performed experiments on the UC Merced remote sensing image dataset [10]. Experiments results show that our SAF module can effectively improve the performance of these SISR models.

The rest of this paper is organized as follows. Section 2 describes details about our self-attention fusion module. Section 3 demonstrates the effectiveness and robustness of the self-attention fusion module by comparative experiments. Conclusions are given in Section 4.

Corresponding to Haopeng Zhang: zhanghaopeng@buaa.edu.cn.

This work was supported in part by the National Key Research and Development Program of China (Grant No. 2019YFC1510905) and the Fundamental Research Funds for the Central Universities.

Authorized licensed use limited to: BEIHANG UNIVERSITY. Downloaded on December 23,2021 at 09:14:55 UTC from IEEE Xplore. Restrictions apply.

2. METHOD

2.1. Framework

The general framework of CNN-based SISR model combined with SAF module is shown in Figure 1. The SAF module is placed after CNN-backbone and before the up-sampling module. SAF module achieves long-range dependencies learning and channel weighting on the deep features extracted by CNN-backbone to improve feature expression ability.

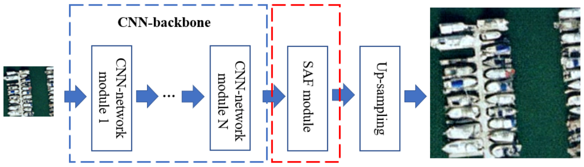


Fig. 1: The CNN-based general SISR model combined with SAF module. Blue dotted frame denotes CNN-backbone. Red dotted frame denotes SAF module.

The whole framework of SAF module is shown in the Figure 2. The whole module combines spatial attention and channel attention in parallel. Putting them in parallel can reduce their mutual interference to make full use of their advantages, thereby further improving the applicability of the output features of this module to SISR. The spatial attention mainly establishes long-range dependence between different image regions so that the model has the ability to learn global features. The channel attention mainly weights the effective channels and weakens the useless channels to improve the representation ability of the model. The output of SAF module is as

$$y = \gamma s + \mu c + x \tag{1}$$

where s is the output of spatial attention, c is the output of channel attention and x is the feature extracted by CNN-backbone. The learnable parameters γ and μ are initially set to 0, in order to avoid destroying the original features. The spatial attention and channel attention will be introduced in Sections 2.2 and 2.3 respectively. The features are optimized by SAF module and then up-sampling can improve the image quality of the reconstructed image.

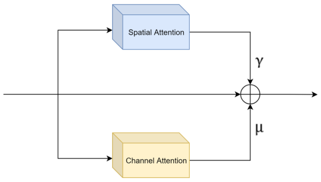


Fig. 2: Framework of the SAF module. $\bigoplus$ denotes feature addition. γ and μ are learnable parameters.

2.2. Spatial Attention

Spatial attention mainly solves the fixed receptive field limitation of convolution operation and improves the model's ability to learn global features. Its structure is shown in Figure 3. Deep feature map $x \in \mathbb{R}^{C\times W\times H}$ is the output of CNN-backbone, where W and H are the width and height of the image respectively, and C is the number of channels. The spatial attention converts deep feature map to two feature spaces f and g through two 1×1 convolutions and reshape operations to calculate spatial attention, where $f(x) \in \mathbb{R}^{C\times N}$, $g(x) \in \mathbb{R}^{C\times N}$, and $N = W \times H$. The calculated spatial attention can be donated as

$$\beta_{j,i} = \frac{\exp\left(s_{ij}\right)}{\sum_{i=1}^{N} \exp\left(s_{ij}\right)}, s_{ij} = f\left(x_i\right)^T g\left(x_j\right) \tag{2}$$

where $\beta_{j,i}$ indicates the influence of the i^{th} pixel on the j^{th} pixel when synthesizing the j^{th} region. This can establish long-range dependencies between different image regions. Then $\beta_{j,i}$ is multiplied by $g(x)$ and reshaped to the initial dimension, in order to get the final spatial attention layer's output $s \in \mathbb{R}^{C\times W\times H}$. Spatial attention can break through the limitation of convolution operation fixed receptive field and make the model to learn global features effectively.

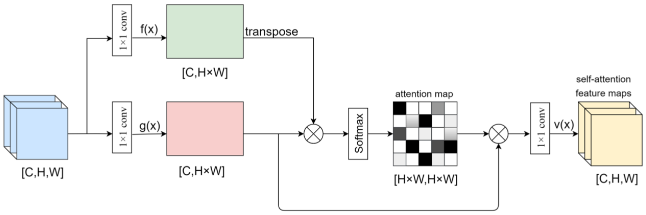


Fig. 3: Self-attention-based spatial attention. $\bigotimes$ denotes multiply.

2.3. Channel Attention

Channel attention mainly selects effective channels for SISR task and improves feature expression ability of SISR models. Its structure is shown in Figure 4. Our channel attention is also implemented through self-attention. It converts deep feature map to two feature spaces k and l through two 1×1 convolutions and reshape operations to calculate channel attention, where $k(x) \in \mathbb{R}^{C\times N}$, $l(x) \in \mathbb{R}^{C\times N}$, and $N = W \times H$. The calculated channel attention can be donated as

$$\alpha_{j,i} = \frac{\exp\left(c_{ij}\right)}{\sum_{i=1}^{N} \exp\left(c_{ij}\right)}, c_{ij} = k\left(x_i\right) l\left(x_j\right)^T \tag{3}$$

where $\alpha_{j,i}$ indicates the influence of the i^{th} channel on the j^{th} channel when synthesizing the j^{th} channel. It selects the most useful channel for the SISR task. Then $\alpha_{j,i}$ is multiplied $g(x)$

Authorized licensed use limited to: BEIHANG UNIVERSITY. Downloaded on December 23,2021 at 09:14:55 UTC from IEEE Xplore. Restrictions apply.

and reshaped to get the final channel attention layer's output. The output of channel attention weights the useful channels and weakens the useless channels for SISR task. This improves the feature expression ability of CNN-based SISR models so that effectively enhances the image quality of the reconstructed images.

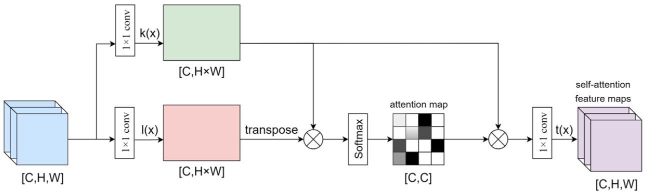


Fig. 4: Self-attention-based channel attention. ⊗ denotes multiply.

3. EXPERIMENTS AND RESULTS

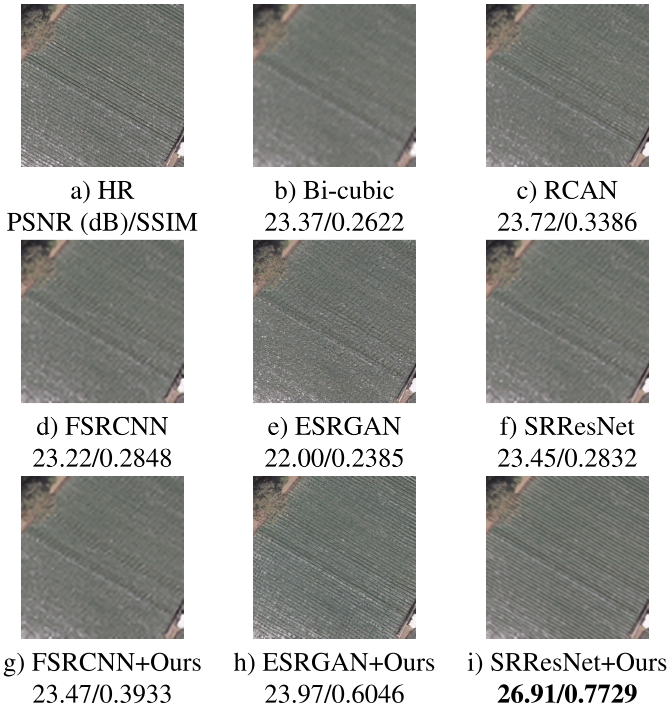
3.1. Experiment Setup

We used the public UC Merced remote sensing image dataset [10], which was widely used in remote sensing image super-resolution. This dataset contains 21 classes of remote sensing scenes and each class consists of 100 images. The size of all images is 256 × 256. The spatial resolution of all images is 0.3m/pixel. We took the original images as high-resolution images, and the low-resolution images were obtained by down-sampling the original images ×4 using bi-cubic interpolation. The dataset was divided into training set, validation set and testing set with the ratios of 70%, 10% and 20% respectively. We used full-reference PSNR and SSIM as evaluation metrics.

3.2. Results

We combined the classic CNN-based SISR models such as FSRCNN [2], SRResNet [3] and ESRGAN [11] with our SAF module, and compared them with original models, bi-cubic interpolation and a state-of-the-art SISR model RCAN [6].

Table 1 and Figure 5 show the experimental results. It can be seen that SRResNet [3] combined with SAF module achieves the state-of-the-art results on UC Merced dataset. The other two models combined with SAF module also improve the performance of the original models, and both PSNR and SSIM have been improved. This proves that SAF module can be applied to different CNN-based SISR models, showing the good generalization of SAF module. From Figure 5, it can be seen that the image reconstructed by ESRGAN [11] combined with SAF module achieves the best visual result. The agricultural texture features of the reconstructed image by this model are almost the same as HR image. The image reconstructed by this combined model shows better texture details and visual effect than the original model. All models combined with SAF module have achieved a substantial improvement in SSIM evaluation metrics, which proves that SAF module can be a useful supplement of convolution operation to enhance CNN-based SISR models' representation ability.



a) HR
PSNR (dB)/SSIM

b) Bi-cubic
23.37/0.2622

c) RCAN
23.72/0.3386

d) FSRCNN
23.22/0.2848

e) ESRGAN
22.00/0.2385

f) SRResNet
23.45/0.2832

g) FSRCNN+Ours
23.47/0.3933

h) ESRGAN+Ours
23.97/0.6046

i) SRResNet+Ours
26.91/0.7729

Fig. 5: Comparison of visual results on UC Merced dataset.

4. CONCLUSIONS

In this paper, we have proposed a self-attention fusion (SAF) module for remote sensing images super-resolution. Our SAF module combines spatial attention and channel attention. Spatial attention can establish long-range dependence between different image regions, and channel attention can weight the useful channels for SISR task. The SAF module can effectively solve the CNN-based SISR models' limitation of inability to learn global features effectively and inferior feature expression ability. Experimental results on UC Merced dataset show that the SRResNet [3] combined with SAF module achieves state-of-the-art results. Different SISR models combined with SAF module can all achieve improvement in the evaluation metrics and image quality, and the visual effect of the reconstructed image has also been enhanced. It demonstrates that our module has good applicability to different SISR models.

Authorized licensed use limited to: BEIHANG UNIVERSITY. Downloaded on December 23,2021 at 09:14:55 UTC from IEEE Xplore. Restrictions apply.

Table 1: Comparison of different SR methods (PSNR/SSIM).

Class	Bi-cubic	RCAN	FSRCNN	FSRCNN+Ours	SRResNet	SRResNet+Ours	ESRGAN	ESRGAN+Ours
agricultural	23.37/0.2622	23.71/0.3362	23.46/0.2933	24.37/0.3848	23.45/0.2832	26.79/0.7629	22.00/0.2385	23.97/0.6046
airplane	30.80/0.8328	34.43/0.8808	32.92/0.8637	33.28/0.8649	34.18/0.8774	34.53/0.8806	30.54/0.7360	31.22/0.8076
baseballdiamond	30.31/0.7601	31.94/0.8143	31.06/0.7931	31.88/0.8051	31.76/0.8093	31.83/0.8112	29.78/0.7243	29.90/0.7251
beach	26.26/0.6693	27.09/0.7383	26.62/0.7046	26.93/0.7237	26.94/0.7274	26.86/0.7271	23.99/0.6161	24.82/0.6458
buildings	29.06/0.8296	32.63/0.8926	30.97/0.8645	30.94/0.8643	32.33/0.8883	32.45/0.8907	30.03/0.8163	29.77/0.8153
chaparral	21.79/0.4543	23.19/0.6305	22.40/0.5439	22.73/0.5792	23.12/0.6004	22.87/0.5858	17.71/0.3390	20.46/0.4742
denseresidential	24.74/0.7806	30.15/0.9085	27.83/0.8617	28.02/0.8679	29.47/0.8965	29.55/0.8997	27.04/0.8221	27.81/0.8364
forest	29.08/0.5829	29.44/0.6265	29.28/0.6185	29.36/0.6200	29.41/0.6230	29.44/0.6275	27.16/0.4936	26.36/0.4486
freeway	28.86/0.8357	33.34/0.9186	31.57/0.8925	31.69/0.9025	32.88/0.9125	33.52/0.9211	29.98/0.8511	30.45/0.8607
golfcourse	29.54/0.7795	31.20/0.8213	30.51/0.8009	30.72/0.8076	31.03/0.8161	31.18/0.8200	29.24/0.7112	28.95/0.7431
harbor	22.77/0.7199	25.94/0.8132	24.65/0.7773	25.01/0.7932	25.81/0.8088	26.43/0.8204	23.03/0.6996	24.33/0.7304
intersection	24.84/0.7425	27.06/0.8213	26.07/0.7913	26.09/0.7914	27.00/0.8169	27.17/0.8224	25.03/0.7429	25.83/0.7621
mediumresidential	23.78/0.5493	24.93/0.6243	24.44/0.6132	24.47/0.6213	24.84/0.6335	24.87/0.6410	23.06/0.5242	22.38/0.5099
mobilehomepark	23.64/0.7051	27.08/0.8380	25.07/0.7616	25.24/0.7922	26.67/0.8252	26.97/0.8316	24.20/0.7293	25.13/0.7651
overpass	24.87/0.7651	30.10/0.8899	27.23/0.8274	27.53/0.8313	29.32/0.8768	29.89/0.8857	26.70/0.7975	27.77/0.8234
parkinglot	22.46/0.7228	24.83/0.8322	23.73/0.7904	24.01/0.7968	24.75/0.8280	27.09/0.8600	21.43/0.6999	24.09/0.7555
river	25.56/0.6369	26.26/0.7037	25.96/0.6792	26.03/0.7012	26.20/0.6952	26.17/0.6989	24.21/0.6133	23.58/0.6069
runway	26.25/0.7708	32.27/0.8673	31.20/0.8518	31.42/0.8601	32.03/0.8621	32.70/0.8664	27.98/0.7508	29.57/0.7894
sparseresidential	24.46/0.4974	25.37/0.5684	25.08/0.5536	25.11/0.5721	25.28/0.5618	25.32/0.5662	23.91/0.4793	23.19/0.4363
storagetanks	22.23/0.6139	25.22/0.7412	23.45/0.6691	23.74/0.6991	24.66/0.7203	24.61/0.7222	23.58/0.6514	24.08/0.6737
tenniscourt	22.41/0.5542	23.92/0.6719	23.27/0.6181	23.42/0.6381	23.85/0.6563	24.31/0.7047	21.64/0.5588	22.15/0.5883
average	25.57/0.6698	28.10/0.7599	26.99/0,7220	27.24/0.7389	27.86/0.7485	**28.31/0.7784**	25.35/0.6474	25.99/0.6858

5. REFERENCES

[1] Chao Dong, Chen Change Loy, Kaiming He, et al., “Image super-resolution using deep convolutional networks,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 2, pp. 295–307, 2015.

[2] Chao Dong, Chen Change Loy, and Xiaoou Tang, “Accelerating the super-resolution convolutional neural network,” in *European Conference on Computer Vision*, 2016, pp. 391–407.

[3] Christian Ledig, Lucas Theis, Ferenc Huszár, et al., “Photo-realistic single image super-resolution using a generative adversarial network,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4681–4690.

[4] Pengrui Wang, Haopeng Zhang, Feng Zhou, et al., “Unsupervised remote sensing image super-resolution using cycle cnn,” in *IEEE International Geoscience and Remote Sensing Symposium*, 2019, pp. 3117–3120.

[5] Lei Zhang and Xiaolin Wu, “An edge-guided image interpolation algorithm via directional filtering and data fusion,” *IEEE Transactions on Image Processing*, vol. 15, no. 8, pp. 2226–2238, 2006.

[6] Yulun Zhang, Kunpeng Li, Kai Li, et al., “Image super-resolution using very deep residual channel attention networks,” in *European Conference on Computer Vision*, 2018, pp. 286–301.

[7] Tao Dai, Jianrui Cai, Yongbing Zhang, et al., “Second-order attention network for single image super-resolution,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11065–11074.

[8] Han Zhang, Ian Goodfellow, Dimitris Metaxas, et al., “Self-attention generative adversarial networks,” in *International conference on machine learning*, 2019, pp. 7354–7363.

[9] Arthur Pajot, Emmanuel de Bezenac, and Patrick Gallinari, “Unsupervised adversarial image reconstruction,” in *International Conference on Learning Representations*, 2019.

[10] Yi Yang and Shawn Newsam, “Bag-of-visual-words and spatial extensions for land-use classification,” in *18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 2010, pp. 270–279.

[11] Xintao Wang, Ke Yu, Shixiang Wu, et al., “Esrgan: Enhanced super-resolution generative adversarial networks,” in *European Conference on Computer Vision*. Springer, 2018, pp. 63–79.

Authorized licensed use limited to: BEIHANG UNIVERSITY. Downloaded on December 23,2021 at 09:14:55 UTC from IEEE Xplore. Restrictions apply.