J Inf Process Syst, Vol.19, No.4, pp.427~438, August 2023
https://doi.org/10.3745/JIPS.02.0199

ISSN 1976-913X (Print)
ISSN 2092-805X (Electronic)

JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

Real Scene Text Image Super-Resolution Based on Multi-Scale and Attention Fusion

Xinhua Lu[1], Haihai Wei[2], Li Ma[1], Qingji Xue[1,*], and Yonghui Fu[2]

Abstract

Plenty of works have indicated that single image super-resolution (SISR) models relying on synthetic datasets are difficult to be applied to real scene text image super-resolution (STISR) for its more complex degradation. The up-to-date dataset for realistic STISR is called TextZoom, while the current methods trained on this dataset have not considered the effect of multi-scale features of text images. In this paper, a multi-scale and attention fusion model for realistic STISR is proposed. The multi-scale learning mechanism is introduced to acquire sophisticated feature representations of text images; The spatial and channel attentions are introduced to capture the local information and inter-channel interaction information of text images; At last, this paper designs a multi-scale residual attention module by skillfully fusing multi-scale learning and attention mechanisms. The experiments on TextZoom demonstrate that the model proposed increases scene text recognition's (ASTER) average recognition accuracy by 1.2% compared to text super-resolution network.

Keywords

Attention Mechanisms, Multi-Scale, Scene Text Recognition, Text Image Super-Resolution

1. Introduction

Identifying the text in the scene images correctly and efficiently can help people quickly acquire the semantic information in the images, which is very important for some text-related downstream tasks, such as image search, robot navigation, and instant translation [1]. For scene text image recognition, several former models [2-8] have achieved excellent recognition results on clear images, but their performance drops dramatically [9] on low-resolution (LR) text images, for character degradation blurring the shapes and edges of text. Single image super-resolution (SISR) technology is considered as an effective preprocessing method for LR text image recognition.

Some SISR models [10-14] have performed better on synthetic datasets in which LR images are typically made by down-sampling and blurring on high-resolution (HR) images. In reality, the above models cannot achieve satisfactory performance on real LR text images, which its degradation is much more complex. For real LR text images, many existing methods [15,16] started from effectively capturing the sequence characteristics of text images or generating finer image boundaries to generate more recognizable text images. The above methods achieved some results on real LR text image, but they neglect the important role of text images' multi-scale features.

Manuscript received July 26, 2022; first revision October 14, 2022; second revision November 7, 2022; accepted November 14, 2022.

* **Corresponding Author:** Qingji Xue (xue_qj@sina.com)
[1] School of Information Engineering, Nanyang Institute of Technology, Nanyang, China (ieluxinhua@sina.com, ielima@sina.com, xue_qj@sina.com)
[2] School of Computer and Artificial Intelligence, Zhengzhou University, Zhengzhou, China (whh15649863783@163.com, fu8926@qq.com)

This paper proposes TSRMAN, a novel scene text image super-resolution (STISR) model for STISR task that combines multi-scale and attention mechanisms. Many works [17-19] demonstrate that the ability of the model can be further improved by fully utilizing features at different scales of images. In addition, attention mechanisms are widely used to guide the model to concentrate on task-relevant regions. Channel and spatial attention (CA, SA) [20] are two mechanisms commonly used in computer vision. The former could establish the interactive dependencies of the image channel dimension, and the latter could get the information on the image space dimension. This paper designs a multi-scale residual attention (MRA) module by skillfully fusing the above two mechanisms to work on real STISR task.

The contributions of our work are as follow:

- To capture text images' features at different scales, multi-scale learning based on different convolution kernel sizes is introduced.
- A MRA module is designed, which skillfully fuses multi-scale and attention mechanisms to enrich the representation ability of image features, and increase the text image recognition's accuracy.
- The experiments demonstrate that our work increases scene text recognition's (ASTER) average recognition accuracy by 1.2% when compared to text super-resolution network (TSRN) on TextZoom.

2. Related Work

2.1 Single Image Super-Resolution

Restoring a reasonable HR image from an LR image is the primary objective of image super-resolution (SR) technology. Dong et al. [10] introduced the convolution neural network (CNN) into the SISR reconstruction task and proposed a simple three-layer network (SRCNN) to generate HR image and its results indicated the advantage of deep learning in SR techniques. Since then, the image SR task has ushered in a large number of SR models using CNN. Inspired by the residual network [21], Kim et al. [11] constructed a model called VDSR, that significantly improves SISR reconstruction results; Tong et al. [22] proposed SR-DenseNet which uses a dense connection mechanism to connect image features of different depths in the model to each other to improve the reconstruction results; Zhang et al. [23] proposed RCAN, which improves image reconstruction's result by introducing the CA module. The above SISR technologies have achieved good results, but most of them rely on artificially synthesized datasets for training. Baek et al. [9] have shown that the performance of these models on real scene LR images drops drastically because the degradation problems of real-scene images are very complex compared to synthetic LR images.

2.2 Scene Text Image Super-Resolution

SISR techniques can improve scene text recognizers' accuracy as a preprocessing step. However, most of the previous SISR models are trained using LR images that are generated artificially, and it is difficult for such models to reconstruct LR scene text images because the similar degradation problem of real scene text images is more complex than that of synthetic LR images. Due to the lack of datasets for scene text image SR tasks, there are few works on real scene LR text images. Wang et al. [15] proposed the dataset called TextZoom filling the gap of datasets for real scene text-image SR tasks and proposed TSRN for this dataset. The experimental results of this model demonstrated that text image SR could significantly increase real scene LR text images' recognition accuracy.

2.3 Scene Text Recognition

The HR text image generated from the SR model is used for the input of the recognizer to obtain all text in origin LR text image. Shi et al. [24] proposed an image sequence recognition model (CRNN), which extracts the sequence information of text images by jointly CNN and recurrent neural model (RNN) and utilizes the connectionist temporal classification (CTC) to match generated characters to real labels. Shi et al. [5] proposed ASTER, explicitly rectified irregular text by introducing the spatial transformation network (STN) [25] and then used an attention-based approach for decoding; Luo et al. [4] proposed a recognition model called MORAN, which designed a multi-object correction module to correct irregular text. In this work, we choose the above models as the model performance evaluator.

3. Proposed Method

The proposed model (TSRMAN) is introduced as follows. We comprehensively describe the main model architecture, then focus on the proposed MRA module.

3.1 Super-Resolution Model

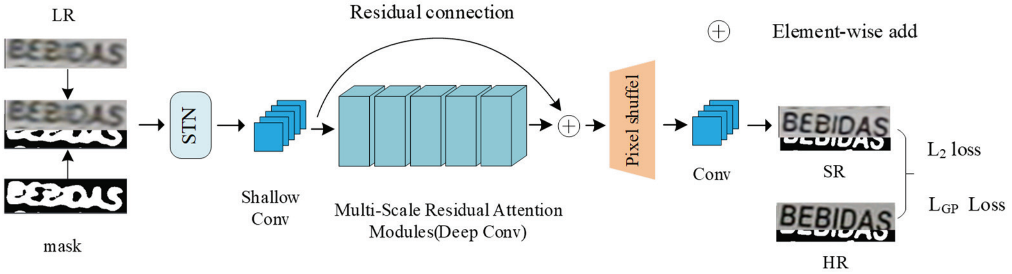

Fig. 1. Architecture of the model (TSRMAN): L_2 loss represents the pixel-wise loss and L_{GP} loss represents the gradient prior loss.

As presented in Fig. 1, our model includes four parts: the STN, the shallow convolution module, the deep feature extraction module consisting of multiple MRA modules connected sequentially, and the upsampling module (pixel shuffle [26]). Firstly, the LR text image and its binarized mask map are combined into a four-channel image and input to STN, the process can be formulated as (1):

$$I_{stn} = F_{stn}(I_{LR}), \tag{1}$$

where $F_{stn}(.)$ represents the STN, which is employed to deal with the issue of pixel alignment between LR-HR text-image pairs and blurring artifacts in reconstructed images, I_{LR} and I_{stn} represent the four-channel LR image and the rectified image, respectively. I_{stn} is input into the shallow convolution module. The shallow feature extraction process can be formulated as (2):

$$X_{SF} = F_{SF}(I_{stn}), \tag{2}$$

where $F_{SF}(.)$ represents the shallow convolution module, which makes use of a single convolutional layer and its convolution kernel size is 9×9, and X_{SF} is the extracted shallow feature map. The large kernel convolution can capture the connection between longer-distance pixels, which is convenient for modeling image features in the global view. Then, X_{SF} is input into the deep feature extraction module, which can be formulated as (3):

$$X_{DF} = F_{DF}(X_{SF}), \tag{3}$$

where $F_{DF}(.)$ represents the deep feature extraction module, and X_{DF} is the extracted deep feature map. $F_{DF}(.)$ can obtain better feature representation, but as the network deepens, it can bring about the problem of exploding or disappearing gradients. Residual connections can solve the above problems by fusing shallow and deep features through element-wise addition. Finally, the fusion map is input into an upsampling module. The upsampling method used in this paper is the subpixel convolution operation [26]. The upsampling process can be formulated as (4):

$$I_{out} = Conv(F_{UP}(add(X_{SF}, X_{DF}))), \tag{4}$$

where $Conv(.)$ represents the operation of generating a four-channel SR image from the unsampled feature map, $F_{UP}(.)$ represents the upsampling module, and $add(.)$ represents element-wise addition, and I_{out} represents the generated SR image. The subpixel convolution operation achieves upsampling by rearranging the pixels. Fig. 2 illustrates the specific implementation.

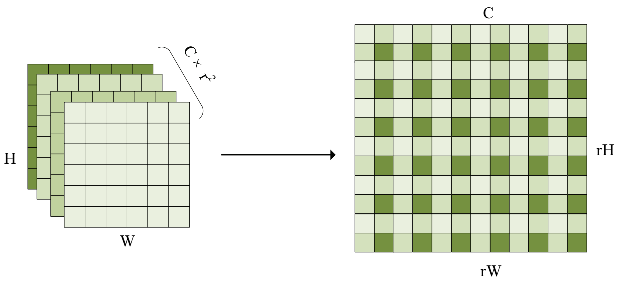

Fig. 2. The sub-pixel convolution operation: H, W, C, and r represent the height, the width, the channel dimension, and the upsampling multiple of the image, respectively (H, W and C represent the same meaning in this paper).

3.2 Multi-Scale Residual Attention Module

To extract LR text images' features fully at different scales, this paper proposes an MRA module. The proposed module is thoroughly introduced in this section.

Fig. 3 is the MRA module which comprises three multi-scale blocks, three BPC (batch normalization [BN] + PReLU+ CA) blocks, one SA module, and a bidirectional gated recurrent unit (GRU). The multi-scale block (MS) is made up of three convolutional layers with different kernel sizes, where the kernel sizes are 1×1, 3×3, and 5×5, respectively. This module could capture the image's feature representation at different scales and use element-wise addition to fuse features. Furthermore, the parameter sharing mechanism is used in this paper to reduce the parameters, that is, all 1×1 convolutions in the multi-scale residual block have the same parameters, and similarly, all 5×5 convolutions also have the same parameters. In BPC, the BN layer in the module is used to reduce the gradient dispersion issue in training deep networks, and it can even accelerate the convergence of the model [27]. The activation layer can increase the model's nonlinearity.

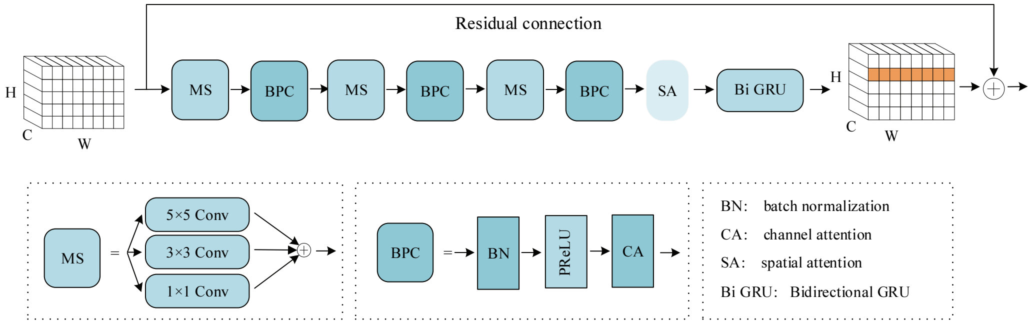

Fig. 3. The multi-scale residual attention module.

Since the convolution operation treats the feature maps of the image channel dimension equally, which limits the representation ability of the model. Therefore, this paper adopts the CA module [23] to establish the dependencies between different channels in Fig. 4.

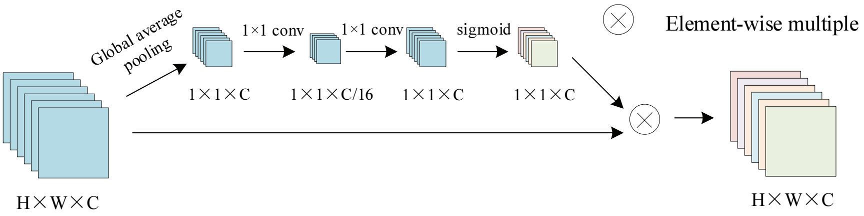

Fig. 4. The channel attention (CA) module.

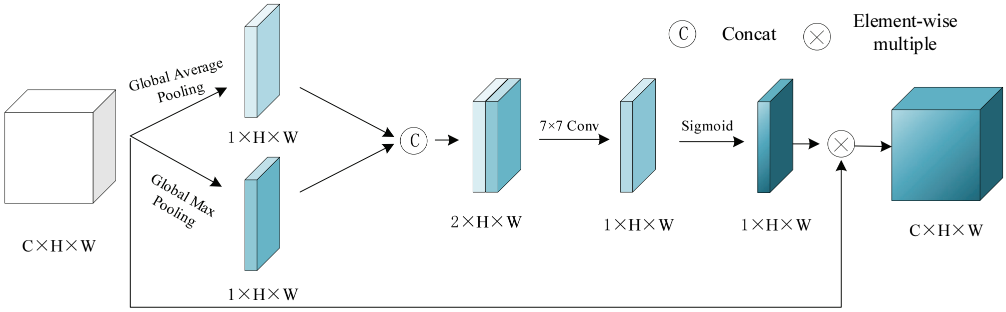

Fig. 5. The spatial attention (SA) module.

Fig. 5 shows SA module. The SA module can direct the model to focus more on image boundaries that contain more high-frequency information, which is helpful for image reconstruction. This paper adopts the SA module proposed in [21].

In [15], the authors demonstrate the effectiveness of modeling the context information using a bidirectional long short-term memory (LSTM). Therefore, this paper takes a similar approach by using GRU with fewer parameters.

3.3 Loss Function

The loss function L includes the pixel loss and the gradient prior loss, which can be express as:

$$L = aL_2 + bL_{GP}, \tag{5}$$

where, coefficients a and b are the weights of different terms of L.

The pixel loss compares the differences between two images. In this paper, the pixel loss is calculated using the mean square error (L_2 loss), which is expressed as Eq. (6):

$$L_2 = \frac{1}{n}\sum_{i=1}^{n}(I_{SR} - I_{HR})^2, \tag{6}$$

where, n represents batch size. I_{SR} represents the SR image and I_{HR} represents the HR image.

Since sharpened characters are more recognizable than smooth ones, this paper adopts the gradient prior loss (L_{GP}) as the same as TSRN to generate sharp image boundaries. L_{GP} is as the equation (7):

$$L_{GP} = \mathbb{E}_i||\nabla I_{HR}(i) - \nabla I_{SR}(i)||_1, \ \ (i \in [i_0, i_1]), \tag{7}$$

where, $\nabla I_{HR}(i), \nabla I_{SR}(i)$ represent the gradient field of HR image and SR image, respectively. i_0, i_1 represent the pixels whose pixel values begin to change and stop changing along the direction of the image gradient, respectively.

4. Experiments

4.1 Dataset

This paper uses TextZoom for model training and testing, including 17,367 pairs of training sets and 4,373 pairs of test sets, in which each data sample contains LR and HR image pairs, and their corresponding text labels. There are three subsets in the test set which include 1,353 image pairs in the hard subset, 1,411 in the medium subset, and 1,619 in the easy subset. Example diagrams are shown in Figs. 6–8.

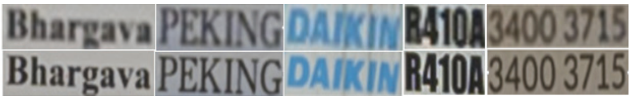

Fig. 6. Easy subset diagram: the top row represents LR text images, and the bottom row represents the corresponding HR text image.

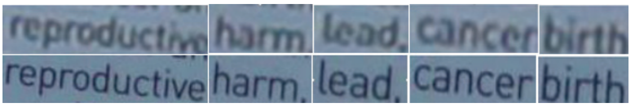

Fig. 7. Medium subset diagram: the top row represents LR text images, and the bottom row represents the corresponding HR text image.

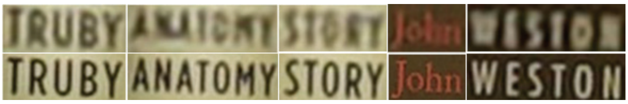

Fig. 8. Hard subset diagram: the top row represents LR text images, and the bottom row represents the corresponding HR text image.

4.2 Implement Details

According to [15], the shape of all LR text images and all HR text images are converted to 16×64 and 32×128, respectively. The coefficients of L_2 and L_{GP} are taken to 1 and 10^{-4}, respectively, and the Adam optimizer with a momentum of 0.9 is used in this paper. The model evaluation metric uses the recognition accuracy obtained from the released PyTorch version of the ASTER [5]. The model is trained on an NVIDIA RTX 3080ti. The training epochs are 500 and the number of images processed at one time is 64. Fig. 9 shows the proposed model's visualization.

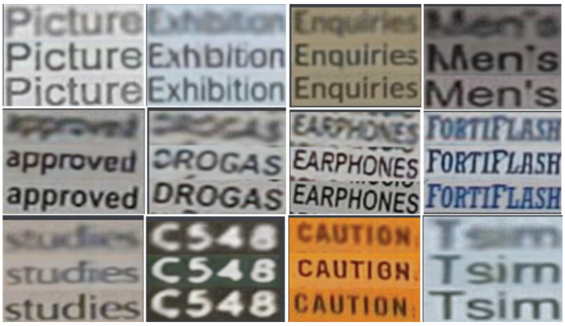

Fig. 9. Image visualization: images with the same character are represented as a group, and from top to bottom are LR text images, SR images generated by the proposed model, and HR text images.

4.3 Ablation Study

This paper analyzes the effectiveness of our work from the following two aspects.

4.3.1 The effect of the number of MRA modules

As shown in Table 1, building deeper networks by increasing the number of MRA can't improve model capability indefinitely. When stacking 6 MRA modules, the accuracy of the model in the text recognizer ASTER begins to decline, and when stacking to 5 MRA modules, the model reaches saturation and obtains the best average recognition accuracy.

Table 1. The effects of the number of MRA modules

Number of modules	ASTER [5] recognition accuracy (%)			
	Easy	Medium	Hard	Average
4	72.8	59.0	40.0	58.3
5	**76.3**	**57.7**	**41.0**	**59.5**
6	75.4	57.8	38.7	58.4

The bold font indicates the best performance in the test on number of MRA modules.

Table 2. The effects of components of MRA module

Configuration		ASTER [5] recognition accuracy (%)			
		Easy	Medium	Hard	Average
Model 1	baseline	75.1	56.3	40.1	58.3
Model 2	+ms	72.8	56.6	39.9	57.5
Model 3	+att	74.1	55.9	40.0	57.8
Model 4	+ms+att	**76.3**	**57.7**	**41.0**	**59.5**

The bold font indicates the best performance in the test on number of MRA modules.

4.3.2 The effect of decomposition components of MRA modules

As shown in Table 2, we take TSRN as the baseline model and obtain four reconstructed models by decomposing the components that constitute the multi-scale residual module. By analyzing the experimental data of Model 1 and Model 2 in Table 2, we found that the model is less effective at reconstructing images with low levels of blur, the reason may be that the reconstruction of this type of image is inhibited by the multi-scale learning features, resulting in a decline in the reconstructed image's quality. We analyze the experimental data of Model 1 and Model 3, and we can conclude that adding the attention module can only achieve a recognition accuracy similar to the benchmark model, which to some extent indicates that the baseline model's feature utilization has reached saturation. We analyzed the experimental data of Model 3 and Model 4 and found that such a skillful fusion of multi-scale and attention mechanisms can significantly enhance the model's performance, demonstrating that such multi-scale module does extract rich feature representations, and also shows that the attention module is indeed possible to filter out the features that are beneficial to the target task from the rich image features.

4.4 Comparison

We compare current SISR models, including SRCNN [10], VDSR [11], SRResNet [28], EDSR [12], LapSRN [17], and TSRN [15] in this paper. The above models are trained with TextZoom, and the model evaluation metric is the text recognition's accuracy. The proposed model achieves an improvement of 1.2% over the baseline model TSRN [15] on TextZoom using the text recognizer ASTER [5], whose specific data are shown in Table 3. The comparison of the recognition accuracy is shown in Fig. 10. Our model achieves relatively high values on images with a large degree of blur, which proves the effectiveness of the model in recovering image details in Table 4.

Table 3. Recognition result of SR models in different recognizers (unit: %)

Model	ASTER [5]				MORAN [4]				CRNN [24]			
	Easy	Medium	Hard	Avg.	Easy	Medium	Hard	Avg.	Easy	Medium	Hard	Avg.
BICUBIC	64.7	42.4	31.2	47.2	60.6	37.9	30.8	44.1	36.4	21.1	21.1	26.8
SRCNN [10]	69.4	43.4	32.2	49.5	63.2	39.0	30.2	45.3	38.7	21.6	20.9	27.7
VDSR [11]	71.7	43.5	34.0	51.0	62.3	42.5	30.5	46.1	41.2	25.6	23.3	30.7
SRResNet [28]	69.6	47.6	34.3	51.3	60.7	42.9	32.6	46.3	39.7	27.6	22.7	30.6
EDSR [12]	72.3	48.6	34.3	53.0	63.6	45.4	32.2	48.1	42.7	29.3	24.1	32.7
LapSRN [17]	71.5	48.6	35.2	53.0	64.6	44.9	32.2	48.3	46.1	27.9	23.6	33.3
TSRN [15]	75.1	56.3	40.1	58.3	**70.1**	**53.3**	**37.9**	**54.8**	52.5	38.2	31.4	41.4
TSRMAN (ours)	**76.3**	**57.7**	**41.0**	**59.5**	70.0	52.4	37.0	54.3	**55.8**	**42.5**	**31.7**	**44.1**

The bold font indicates the best performance in the test to different recognizers.

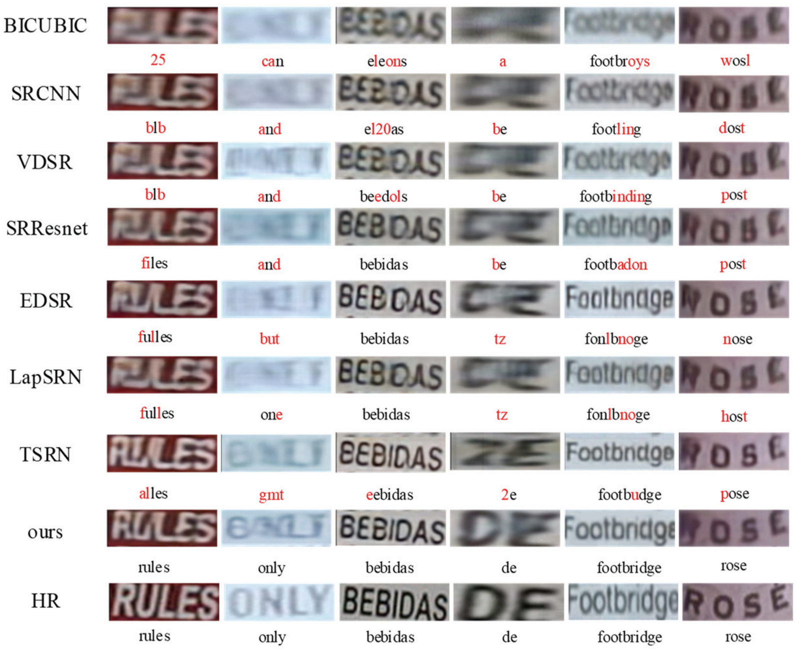

Fig. 10. Different SR methods' visualization and recognition results: the last row represents an HR image, "ours" represents the proposed model (TSRMAN), and "red" presents misrecognition.

Table 4. PSNR and SSIM result comparison

Method	PSNR (dB)				SSIM			
	Easy	Medium	Hard	Avg.	Easy	Medium	Hard	Avg.
BICUBIC	22.35	18.98	19.39	0.7884	0.6254	0.6592	22.35	18.98
SRCNN [10]	23.48	19.06	19.34	0.8379	0.6323	0.6791	23.48	19.06
VDSR [11]	24.62	18.96	19.79	0.8631	0.6166	0.6989	24.62	18.96
SRResnet [28]	24.36	18.88	19.29	0.8681	0.6406	0.6911	24.36	18.88
EDSR [12]	24.26	18.63	19.14	0.8633	0.6440	0.7108	24.26	18.63
LapSRN [17]	24.58	18.85	19.77	0.8556	0.6480	0.7087	24.58	18.85
TSRN [15]	**25.07**	18.86	19.71	**0.8897**	**0.6676**	0.7302	**25.07**	18.86
TSRMAN (ours)	23.48	**19.07**	**19.84**	0.8655	0.6648	**0.7359**	23.48	**19.07**

PSNR=peak signal-to-noise ratio, SSIM=structural similarity index measure.
The bold font indicates the best performance in the test by PSNR and SSIM.

5. Conclusion

In this paper, a novel text image super-resolution model is proposed for real scene text image super resolution. This paper skillfully combines the multi-scale learning and the attention mechanisms, and designs a MRA module, thereby improving the text recognizers' recognition accuracy in LR scene text images, surpassing the existing baseline model TSRN. The proposed model has achieved better results, but the recognition accuracy on extremely blurred and long text images is still low. In the future work, for blurred text, we will introduce deblurring techniques, and for long texts, we can try to build self-attention mechanisms to learn long-distance semantic information.

Acknowledgement

This paper was sponsored by Natural Science Foundation of Henan (No. 222300420504) and Academic Degrees & Graduate Education Reform Project of Henan Province (No. 2021SJGLX262Y).

References

[1] X. Chen, L. Jin, Y. Zhu, C. Luo, and T. Wang, “Text recognition in the wild: a survey,” *ACM Computing Surveys (CSUR)*, vol. 54, no. 2, article no. 42, 2021. https://doi.org/10.1145/3440756

[2] W. Liu, C. Chen, K. Y. K. Wong, Z. Su, and J. Han, “Star-Net: a spatial attention residue network for scene text recognition,” in *Proceedings of the British Machine Vision Conference (BMVC)*, York, UK, 2016.

[3] Z. Cheng, F. Bai, Y. Xu, G. Zheng, S. Pu, and S. Zhou, “Focusing attention: towards accurate text recognition in natural images,” in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, 2017, pp. 5086-5094.

[4] C. Luo, L. Jin, and Z. Sun, “Moran: a multi-object rectified attention network for scene text recognition,” *Pattern Recognition*, vol. 90, pp. 109-118, 2019. https://doi.org/10.1016/j.patcog.2019.01.020

[5] B. Shi, M. Yang, X. Wang, P. Lyu, C. Yao, and X. Bai, “ASTER: an attentional scene text recognizer with flexible rectification,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 9, pp. 2035-2048, 2019. https://doi.org/10.1109/TPAMI.2018.2848939

[6] E. Xie, Y. Zang, S. Shao, G. Yu, C. Yao, and G. Li, “Scene text detection with supervised pyramid context network,” *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 1, pp. 9038-9045, 2019. https://doi.org/10.1609/aaai.v33i01.33019038

[7] W. Wang, E. Xie, X. Li, W. Hou, T. Lu, G. Yu, and S. Shao, “Shape robust text detection with progressive scale expansion network,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, 2019, pp. 9336-9345.

[8] W. Wang, E. Xie, X. Song, Y. Zang, W. Wang, T. Lu, G. Yu, and C. Shen, “Efficient and accurate arbitrary-shaped text detection with pixel aggregation network,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Seoul, South Korea, 2019, pp. 8439-8448.

[9] J. Baek, G. Kim, J. Lee, S. Park, D. Han, S. Yun, S. J. Oh, and H. Lee, “What is wrong with scene text recognition model comparisons? dataset and model analysis,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Seoul, South Korea, 2019, pp. 4714-4722.

[10] C. Dong, C. C. Loy, K. He, and X. Tang, “Image super-resolution using deep convolutional networks,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 2, pp. 295-307, 2016. https://doi.org/10.1109/TPAMI.2015.2439281

[11] J. Kim, J. K. Lee, and K. M. Lee, “Accurate image super-resolution using very deep convolutional networks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, 2016, pp. 1646-1654.

[12] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, “Enhanced deep residual networks for single image super-resolution,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, Honolulu, HI, 2017, pp. 1132-1140.

[13] B. Liu, “Lightweight single image super-resolution by channel split residual convolution,” *Journal of Information Processing Systems*, vol. 18, no. 1, pp. 12-25, 2022. http://doi.org/10.3745/JIPS.02.0168

[14] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, “Residual dense network for image super-resolution,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, 2018, pp. 2472-2481.

[15] W. Wang, E. Xie, X. Liu, W. Wang, D. Liang, C. Shen, and X. Bai, "Scene text image super-resolution in the wild," in *Computer Vision–ECCV 2020*. Cham, Switzerland: Springer, 2020, pp. 650-666. https://doi.org/10.1007/978-3-030-58607-2_38

[16] C. Fang, Y. Zhu, L. Liao, and X. Ling, "TSRGAN: real-world text image super-resolution based on adversarial learning and triplet attention," *Neurocomputing*, vol. 455, pp. 88-96, 2021. https://doi.org/10.1016/j.neucom.2021.05.060

[17] W. S. Lai, J. B. Huang, N. Ahuja, and M. H. Yang, "Deep Laplacian pyramid networks for fast and accurate super-resolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, 2017, pp. 5835-5843.

[18] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, 2016, pp. 2818-2826.

[19] J. Qin, Y. Huang, and W. Wen, "Multi-scale feature fusion residual network for single image super-resolution," *Neurocomputing*, vol. 379, pp. 334-342, 2020. https://doi.org/10.1016/j.neucom.2019.10.076

[20] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: convolutional block attention module," in *Proceedings of the European Conference on Computer Vision (ECCV)*, Munich, Germany, 2018, pp. 3-19).

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, 2016, pp. 770-778.

[22] T. Tong, G. Li, X. Liu, and Q. Gao, "Image super-resolution using dense skip connections," in *Proceedings of the IEEE International Conference on Computer Vision*, Venice, Italy, 2017, pp. 4809-4817.

[23] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, Munich, Germany, 2018, pp. 294-310.

[24] B. Shi, X. Bai, and C. Yao, "An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 11, pp. 2298-2304, 2017. https://doi.org/10.1109/TPAMI.2016.2646371

[25] M. Jaderberg, K. Simonyan, and A. Zisserman, "Spatial transformer networks," *Advances in Neural Information Processing Systems*, vol. 28, pp. 2017-2025, 2015.

[26] W. Shi, J. Caballero, F. Huszar, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, 2016, pp. 1874-1883.

[27] J. W. Liu, H. D. Zhao, X. L. Luo, ad J. Xu, "Research progress on batch normalization of deep learning and its related algorithms," *Acta Automatica Sinica*, vol. 46, no. 6, pp. 1090-1120, 2020.

[28] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, et al., "Photo-realistic single image super-resolution using a generative adversarial network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, 2017, pp. 105-114.

Xinhua Lu **https://orcid.org/0000-0002-2338-7020**

He received the B.E., M.E., and Ph.D. degrees from Zhengzhou University, Zhengzhou, China, in 2003, 2007, and 2019, respectively. Now he is with Nanyang Institute of Technology, Nanyang, China, as a lecturer in the School of Information Engineering. His main research interests include machine learning, Variational Bayesian Inference, and wireless communication. Since September 2015, he was a visiting researcher in Department of Electronic Systems, Aalborg University for 2 years supported by China Scholarship Council.

Haihai Wei **https://orcid.org/0000-0002-0361-6885**

He received his B.E. degree from Henan University of Technology, Zhengzhou, China, in 2020, where he is currently pursuing the M.E. degree from Zhengzhou University. His research interests include computer vision and scene text recognition.

Li Ma **https://orcid.org/0000-0002-7114-3880**

She received the B.E. degree from XiDian University, Xi'an, China, in 2015. Now she is with Nanyang Institute of Technology, Nanyang, China, as an assistant professor in School of Information Engineering. Her main research interests include machine learning, variational Bayesian inference, and wireless communication.

Qingji Xue **https://orcid.org/0000-0001-7516-9984**

He received Ph.D. in School of Computer Science and Technology from Wuhan University of Technology, Wuhan, China, in 2011. He is an in School of Digital Media and Art Design, Nanyang Institute of Technology. His current research interests include computer vision and IoT technology.

Yonghui Fu **https://orcid.org/0000-0001-8196-9400**

He received the B.E. degree from Zhengzhou University of Aeronautics, Zhengzhou, China, in 2020, where he is currently pursuing the M.E. degree from Zhengzhou University. His research interests include computer vision and visual SLAM.