2022 21st International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES)

Multi-Scale Multi-Stage Single Image Super-Resolution Reconstruction Algorithm Based on Transformer

Wei Wang[1], Yinfang Zhu[1], Dewu Ding[1], Jing Li[1], Yu Luo[2]
[1]*School of Mathematics and Computer Science, Yichun University, Yichun 336000*
[2]*School of Computer Science and Technology, Guangdong University of Technology, Guangzhou 510006*
Email:2111905068@mail2.gdut.edu.cn

***Abstract*—In this paper, creatively combining Transformer with image super-resolution reconstruction, we proposes a multi-scale multi-stage single image super-resolution reconstruction algorithm based on Transformer (MSTN). The algorithm uses Transformer as a feature sharing module, thus it realizes network parameter sharing, dynamically focuses on the correlation between feature information of adjacent stages, and then extracts the high-frequency texture information embedded in the current stage features from the feature information learned in the previous stage, which achieves a coarse-to-fine enhancement of image reconstruction. Experiments show that our method can not only perform better image super-resolution reconstruction compared with other advanced methods, but also reduce the network parameters to a great extent.**

***Keywords*-Super resolution; Transformer; computer vision; deep learning; image processing;**

I. Introduction

Image super-resolution reconstruction is a popular research subject in the field of computer vision. It refers to interpolating low-resolution images to obtain super-resolution(SR) images. But mapping a low-resolution(LR) image to a SR image is a highly uncertain solution problem. Under this solution space, one LR image can generate several different SR images, and one SR image can also get several different LR images. Therefore, for the above problem, scholars have proposed a variety of SR methods to solve this phenomenon.

Dong et al. firstly proposed a three-layer neural network SRCNN [1] based on image super-resolution reconstruction and achieved good results compared with the previous conventional super-resolution methods. Kim et al. proposed VDSR[2] with 20 layers in depth. Tai et al. designed DRRN [3] by combining skip connections with recursive structure. Kim et al. proposed DRCN [4] with recursive structure. These network structures are based on first interpolating the images to the same size as the high-resolution(HR) before feeding them into the network structure for training, but this increases the number of parameters in the network as well as greatly increases the computation time. In order to solve the above mentioned problems, Dong et al. proposed the FSRCNN [5] network structure and Shi et al. proposed the ESPCN [6] network structure. These two network structures represent two different methods of upsampling: FSRCNN uses deconvolution to implement the image upsampling operation, and ESPCN uses subpixel convolution to implement the image upsampling operation.These two strategies greatly reduce the network parameters and save computing time.

As the network deepens it can cause the network to be prone to gradient explosion/gradient disappearance during training. To avoid the above problems, Lim et al. proposed deep EDSR [7] network architecture by skipping connections. Tong et al. proposed the SRDenseNet network architecture by densely connecting each layer of the network. Zhang et al. proposed the RDN network architecture by combining residual connectivity and dense connectivity. All these network architectures not only avoid the gradient explosion/gradient disappearance problem but also effectively allow the flow of information from the bottom layer to the top layer in the network by using skip connections or dense connections, which avoid the problem of information loss as the depth of the network is superimposed.

Although super-resolution reconstruction has made wonderful advances in convolutional neural networks, there are some limitations on CNN-based network models.

- Most super-resolution network structures are single-stage end-to-end forms, while ignoring the multi-stage feature information in the network reconstruction process.
- Most network models train specific scales to produce the corresponding super-resolution results, and therefore need to train corresponding network models for different desired scales, which not only increases the time cost but also increases the calculation cost.

To address these problems, we propose the multi-scale multi-stage single image super-resolution reconstruction (MSTN) algorithm based on Transformer to reconstruct SR images, which can effectively use feature information from different stages to reconstruct multi-scale SR images. To achieve this, the model uses Transformer as a feature sharing module (TFM) thus it focues on the coupling relationship between $T-1$ stage image feature information and T stage feature information, which in turn allows the highly coupled feature information from $T-1$ stage to further play a role in the reconstruction process of T stage, and therefore it perform image super-resolution reconstruction from coarse to fine.

II. Methodology

Fig. 1 shows the proposed image super-resolution network architecture. Our network takes the LR image after bilinear interpolation as input and outputs the multi-scale

2473-3636/22/$31.00 ©2022 IEEE
DOI 10.1109/DCABES57229.2022.00044

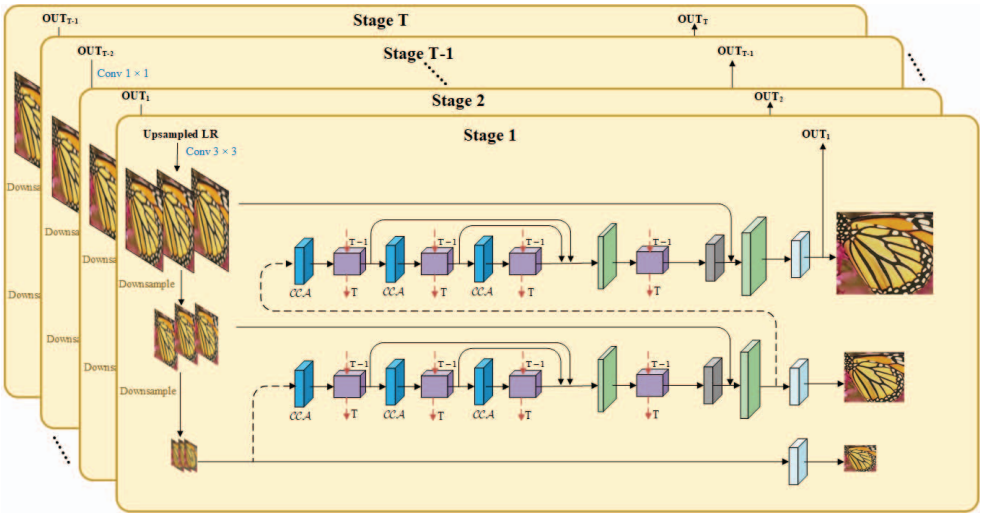

Figure 1. The overall architecture of the network, where the purple module represents the feature sharing module, is shown in Fig. 2. The network architecture is optimized through multiple stages, and each stage outputs multiple high-resolution SR patches, while the feature sharing module is used to dynamically focus on the correlation of feature information in adjacent stages, which shares multi-stage network parameters and accelerates network training.

SR images simultaneously. The whole network architecture is optimized via T phases. Each stage starts with feature extraction through convolutional layers with channel attention (CA) in a residual dense connection. The network is further optimized by feature sharing module (TFM) for $T-1$ stage feature information extraction, which allows high coupling feature information for SR reconstruction in T stage. Finally the overall network architecture is constrained with the help of loss of attention mechanism.

A. Network architecture

In terms of network architecture, the network architecture proposed in this paper is similar to the C^{3}Net [8]. Both of them are optimized by iteratively T stages to implement image super-resolution reconstruction. Meanwhile, the network parameters of each stage are shared, which reduces the network parameters. To improve the network feature extraction capability, our network architecture also adopts the $\mathcal{CCA}$ [8] module. Therefore, this paper will not describe the $\mathcal{CCA}$ in detail and focus on the feature sharing module (TFM).

B. Feature sharing module

Our network makes full use of the high-frequency texture information in $T-1$ stage, and therefore we propose a new feature sharing module (TFM) as shown in Fig. 2, which reduces the network parameters and performs SR reconstruction from coarse to fine. The network first records the feature information of $T-1$ stage by Transformer, followed by calculating the correlation between T stage and $T-1$ stage features by inner product, then outputting the correlation attention map between $T-1$ stage feature information and T stage feature information by Softmax function. Finally it multiplies the feature information of stage T and the attention map by dot product operation to obtain the high frequency texture

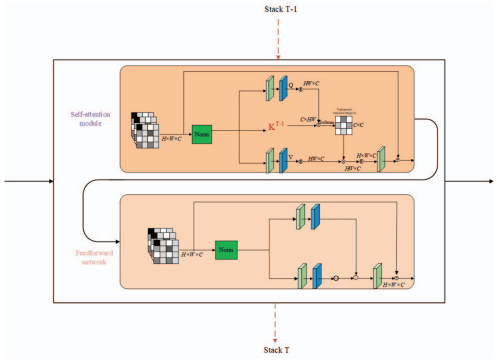

Figure 2. With Transformer dynamically focusing on the correlation between the multi-scale feature information of the T^{th} stage and the feature information of the $(T-1)^{th}$ stage, the higher the correlation between the features of adjacent stages, the more high-frequency texture information can be reconstructed.

information of the current stage. Meanwhile, for the better acquisition of feature information of the network during propagation, the network still uses $\mathcal{CCA}$ for feature extraction to enhance the network feature extraction capability.

$$\mathcal{F}_{TFM}^{(T)} = \mathcal{G}[\mathcal{X}_{\mathcal{CCA}}, \mathcal{F}_{TFM}^{(T-1)}], \tag{1}$$

where $\mathcal{G}$ is feature sharing unit. $\mathcal{X}_{\mathcal{CCA}}$ is the output of the $\mathcal{CCA}$ module.

C. Loss attention mechanism

In order to constrain the network reconstruction loss at different stages at different scales, the network uses the loss-attention mechanism proposed by C^{3}Net to constrain the network, thus improving the SR reconstruction capability of the model.

$$\mathcal{L}_{\Theta}^{(\times 1)} = \frac{1}{T}\frac{1}{N}\sum_{t=1}^{T}\sum_{m=1}^{N} \|\mathcal{F}_{\Theta}(I_{LR}^{(m)}) - I_{HR}^{(m)}\|_1, \tag{2}$$

Table I
THE AVERAGE PSNR/SSIM RESULTS OF DIFFERENT METHODS UNDER $\times 2$ AND $\times 4$ MODELS ARE COMPARED ON FOUR BENCHMARK DATASETS.

Scale	Methods	Set5		Set14		BSD100		Urban100	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
$\times 2$	SRCNN[1]	36.66	0.9542	32.45	0.9067	31.36	0.8879	29.50	0.8946
	ESPCN[6]	37.00	0.9559	32.75	0.9098	31.51	0.8939	29.87	0.9065
	FSRCNN[5]	37.05	0.9560	32.66	0.9090	31.53	0.8920	29.88	0.9020
	VDSR[2]	37.53	0.9590	33.05	0.9130	31.90	0.8960	30.77	0.9140
	DRCN[4]	37.63	0.9588	33.04	0.9118	31.85	0.8942	30.75	0.9133
	LapSRN[9]	37.52	0.9590	33.08	0.9130	31.08	0.8950	30.41	0.9101
	DRRN[3]	37.74	0.9591	33.23	0.9136	32.05	0.8973	31.23	0.9188
	Ours	**37.98**	**0.9610**	**33.40**	**0.9158**	**32.15**	**0.9000**	**32.04**	**0.9269**
$\times 4$	SRCNN[1]	30.50	0.8573	27.50	0.7513	26.90	0.7103	24.52	0.7226
	ESPCN[6]	30.66	0.8646	23.45	0.5980	23.92	0.5740	21.20	0.5540
	FSRCNN[5]	30.73	0.8601	27.59	0.7535	26.96	0.7128	24.60	0.7258
	VDSR[2]	31.36	0.8796	28.02	0.7678	27.29	0.7252	25.18	0.7525
	DRCN[4]	31.56	0.8810	28.15	0.7627	27.24	0.7150	25.15	0.7530
	LapSRN[9]	31.54	0.8811	28.09	0.7694	27.32	0.7264	25.21	0.7553
	DRRN[3]	31.68	0.8888	28.21	0.7720	27.38	0.7284	25.44	0.7638
	Ours	**32.14**	**0.8920**	**28.41**	**0.7761**	**27.30**	**0.7330**	**25.95**	**0.7798**

Table II
COMPARE THE PARAMETERS OF EACH METHOD AND THE AVERAGE SSIM VALUE OF THE SET5 DATASET UNDER THE $\times 2$ SCALE

	EDSR[7]	SRMDNF[10]	CARN[11]	MSRN[12]	CRN[13]	C^3Net[8]	Ours
SSIM	0.9602	0.9600	0.9590	0.9605	0.9610	**0.9612**	**0.9611**
Parameters	40.7 Mb	1.51 Mb	1.59 Mb	5.89 Mb	9.47 Mb	2.41 Mb	**1.68 Mb**

$$\mathcal{L}_{\Theta}^{(\times 2)} = \frac{1}{T}\frac{1}{N}\sum_{t=1}^{T}\sum_{m=1}^{N} \|\mathcal{F}_{\Theta}(I_{LR}^{(m)}) - I_{\times 2HR}^{(m)}\|_1, \quad (3)$$

$$\mathcal{L}_{\Theta}^{(\times 4)} = \frac{1}{T}\frac{1}{N}\sum_{t=1}^{T}\sum_{m=1}^{N} \|\mathcal{F}_{\Theta}(I_{LR}^{(m)}) - I_{\times 4HR}^{(m)}\|_1, \quad (4)$$

where $\mathcal{F}$ is the proposed network, and Θ denotes all the parameters of the network. $I_{LR}^{(m)}$ is the m^{th} LR image. $I_{HR}^{(m)}$, $I_{\times 2HR}^{(m)}$, and $I_{\times 4HR}^{(m)}$ are the m^{th} goundtruth HR image for the scale of $\times 1$, $\times 2$, and $\times 4$, respectively. ω_1, ω_2, and ω_3 are used to balance the loss between different scales. These three parameters are studied in the training process of the network.

$$\mathcal{L}_{total} = \omega_1 \mathcal{L}_{\Theta}^{(\times 1)} + \omega_2 \mathcal{L}_{\Theta}^{(\times 2)} + \omega_3 \mathcal{L}_{\Theta}^{(\times 4)}. \quad (5)$$

III. EXPERIMENT

A. Datasets & evaluation criteria

In this paper, 800 training images from the DIV2K dataset are used to train the proposed network architecture. Set5, Set14, Urban100, and BSD100 are also used to test the generalization ability of the proposed network. Finally, peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) are used as evaluation criteria.

B. Experimental comparison

In this paper, the proposed method's performance is evaluated on four benchmark datasets, namely Set5, Set14, Urban100, and BSD100. The network compares the reconstruction results at $\times 2$ and $\times 4$ scale with seven current state-of-the-art methods. These methods are SRCNN[1], ESPCN[6], FSRCNN[5], VDSR[2], DRCN[4], and LapSRN[9], DRRN[3]. Finally the results are shown in Table I

From Table II, it can be seen that at $\times 2$ scale, the proposed method achieves not only good values of SSIM, but also the best network parameters for the Set5 benchmark dataset. It is fully demonstrated that the feature sharing module proposed in this paper largely reduces the number of network parameters and improves the network reconstruction performance.

IV. CONCLUSION

In this paper, we propose a multi-scale multi-stage single image super-resolution reconstruction algorithm (MSTN) based on Transformer to improve image super-resolution reconstruction. The network reconstructs SR images with multi-scale resolution simultaneously in a multi-stage manner, and uses the Transformer as a feature sharing module (TFM) to dynamically focus on feature-related information between adjacent stages during the reconstruction process, which makes full use of the high-frequency feature information learned in the T-1 stage to achieve better T stage SR image reconstruction. Finally, we train the network on publicly available datasets and simultaneously test on its four benchmark datasets to demonstrate the generalization ability of the proposed network.

ACKNOWLEDQMENTS

This work is partially supported by the National Natural Science Foundation of China under Grant No. 62161050.

REFERENCES

[1] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *ECCV*, 2014.

[2] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," *2016 CVPR*, pp. 1646–1654, 2016.

[3] Y. Tai, J. Yang, and X. Liu, "Image super-resolution via deep recursive residual network," *2017 CVPR*, pp. 2790–2798, 2017.

[4] J. Kim, J. K. Lee, and K. M. Lee, "Deeply-recursive convolutional network for image super-resolution," *2016 CVPR*, pp. 1637–1645, 2016.

[5] C. Dong, C. C. Loy, and X. Tang, "Accelerating the super-resolution convolutional neural network," in *2016 ECCV*, 2016.

[6] W. Shi, J. Caballero, F. Huszár, 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," *2016 CVPR*, pp. 1874–1883, 2016.

[7] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," *2017 CVPRW*, pp. 1132–1140, 2017.

[8] W. Wang, Y. Luo, J. Ling, Y. Song, and T. Zhou, "C3net: A cross-channel cross-scale and cross-stage network for single image super-resolution," in *2022 IEEE International Conference on Multimedia and Expo (ICME)*, 2022, pp. 1–6.

[9] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Deep laplacian pyramid networks for fast and accurate super-resolution," *2017 CVPR*, pp. 5835–5843, 2017.

[10] K. Zhang, W. Zuo, and L. Zhang, "Learning a single convolutional super-resolution network for multiple degradations," *CVPR*, pp. 3262–3271, 2018.

[11] N. Ahn, B. Kang, and K.-A. Sohn, "Fast, accurate, and, lightweight super-resolution with cascading residual network," *ArXiv*, vol. abs/1803.08664, 2018.

[12] J. Li, F. Fang, K. Mei, and G. Zhang, "Multi-scale residual network for image super-resolution," in *ECCV*, 2018.

[13] R. Lan, L. Sun, Z. Liu, H. Lu, Z. Su, C. Pang, and X. Luo, "Cascading and enhanced residual networks for accurate single-image super-resolution," *IEEE Transactions on Cybernetics*, vol. 51, pp. 115–125, 2021.