IMAGE SUPER-RESOLUTION MODEL IMPROVEMENT

ABSTRACT

Super-resolution is task of reconstructing images from low-resolution (LR) ones. Deep learning methods have achieved outstanding results recently, especially in the field of transformer-based methods. Swin IR is the first successful method that applies Swin Transformer as the super-resolution backbone directly. It uses self-attention (SA) mechanism for feature extraction. Although Swin Transformer uses the techniques of window partitioning and window shifting to reduce the redundant computation of self-attention, the amount of computation is still large. There is a method called Efficient Long-Range Attention Network (ELAN) that further reduces the amount of computation of transformer-based method. However, we find that there are some classical modules, such as Cutblur, and dense connection, are not used in this model. Hence, we aim at improving the performance (PSNR and SSIM) of ELAN method by modifying its algorithm in this paper.

1. INTRODUCTION

Single Image Super-Resolution (SISR) aims at reconstructing a High-Resolution (HR) image from its Low-Resolution (LR) one.

There are many applications of super-resolution. For example, restoring JPEG image in which image quality is broken by lossy compression. Futhermore, [9] demonstrates the benefit of medical imaging. Moreover, the application field of super-resolution is not limited to SISR. For instance, video restoration [8] can be applied in the entertainment industry. The classical method to solve this problem is the interpolation method such as bilinear or bicubic interpolation. Those methods are efficient. However, the quality of super-resolution is still expected to improve. In this decade, deep learning based SISR methods, such as using Convolution Neural Networks (CNN) have achieved great improvement than the traditional ones. And recently, there are many researchers doing Transformer-based methods which use the self-attention mechanism (SA) [18] borrowed from NLP. Transformed-based super-resolution methods [1, 3, 4] have much more outstanding reconstruction performance than CNN-based methods [5, 6, 7]. However, the usage of computational resources is much larger than CNN-based methods. Especially, we have to use at least 160GB GPU memory space to reproduce the experiment in HAT-L [4]. ELAN [1], an efficient long-range attention network aims at improving the computational efficiency of the Transformer-based method. Despite the tradeoff between reconstruction performance and efficiency, it still performs well like some of the outstanding Transformer-based methods, such as SwinIR [3]. Based on the image super-resolution model – ELAN, this paper proposes some modifications to improve its performance. The first step is to enhance the shallow feature extraction stage because inspired by Inception [19], we think adding a multi-scale convolution kernel can help model extract more information. The second step is applying Cutblur [16] data augmentation method to our paper because training the original model only uses rotation or flip. We would like the model can have better generalization. The third step is using the dense and residual connection in the deep feature extraction stage because the paper [2] which is a classical super-resolution model takes the way and has good reconstruction performance.

The paper is structured as follows. Section 2 presents related works, Section 3 presents methods and steps, Sections 4 and 5 present experiments and their results, and finally, Section 6 concludes and presents future work.

Compared with the original model -- ELAN, we summarize the original contribution of this paper to improve ELAN as follows:

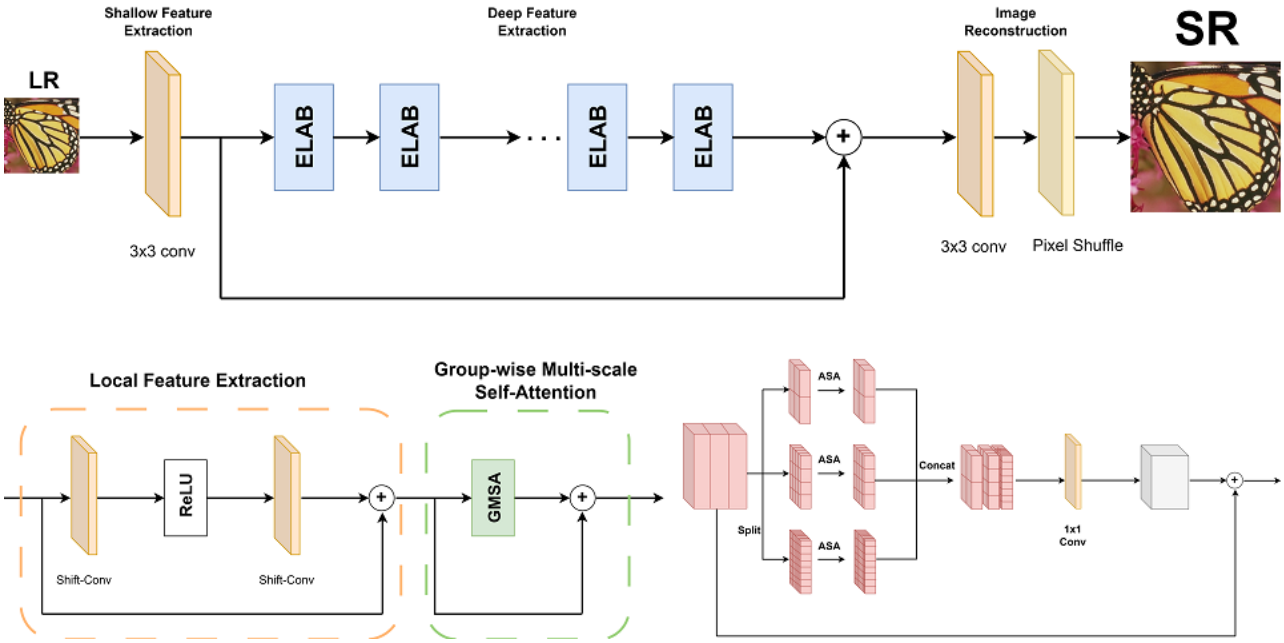


Fig. 1: Illustration of the overall pipeline of the Efficient Long-range Attention Network (ELAN). The architecture of the Efficient Long-range Attention Block (ELAB). Illustration of the computation of Group-wise Multi-scale Self-Attention (GMSA).

First, we modify the shallow feature extraction part as an Inception-like model to use different receptive fields to extract more features.

Second, we integrate the Cutblur data Augmentation method which generates more training images than only flipping or rotating the images to enhance model generalization.

Third, we modify the ELAN model by adding both dense
connection and residual connection to extract more deep features of the images.
Last, we add the bicubic residual connection to make our model have better performance.

2. RELATED WORKS

In this section, we briefly describe the related works of deep learning image super resolution methods.

2.1. CNN-based Methods

The earliest CNN-based super-resolution methods employed simple shallow convolution neural network structures, such as SRCNN (Super-Resolution Convolution Neural Network) [5]. The methods achieved some level of improvement in resolution by learning the mapping from low-resolution to high-resolution images. Then papers began exploring deeper convolution neural network structures to enhance super-resolution performance. For example, VDSR (Very Deep Super-Resolution) [6] introduced deeper network architectures that increased depth and parameters to improve the quality of super-resolved images. To further enhance super-resolution performance, researchers experimented with incremental learning and residual learning techniques. Methods like EDSR (Enhanced Deep Super-Resolution) [7] introduced residual learning, learning residuals to improve the quality of super-resolved results. To build a better model for super-resolution, attention mechanisms have been integrated into CNN-based super-resolution. For instance, RCAN [22] introduces channel attention to super-resolution. To generate more realistic high-resolution images or generating very-high-resolution images, generative adversarial networks (GANs) have been integrated into CNN-based super-resolution, such as SRGAN [23], ESRGAN [21], and etc.

2.2. Transformer-based Methods

After achieving big success in the field of NLP, many researchers are focus on taking advantage of the self-attention mechanism [18] in the field of computer vision. ViT [26] is the first well-known vision transformer that applies the self-attention mechanism directly that outperforms the CNN-based method in Imagenet [24] classification task. However, the amount of computational resources is very high. Swin Transformer [25] uses Windows Multi-head Self-Attention (WMSA) which integrates the advantages of CNN and Transformer. It reduces the amount of computation by just calculating local attention and sharing information between different local areas by shifting windows. Although this method may lose some useful information, it still has good performance on the benchmark like

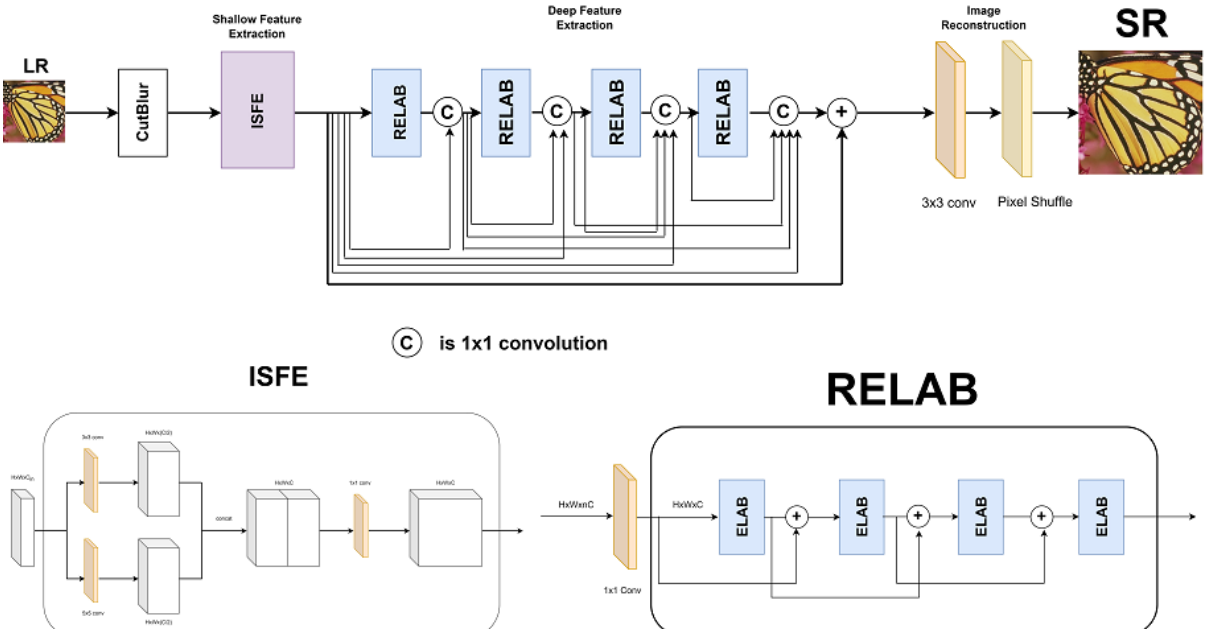


Fig. 2: The description of our proposed model.

Imagenet. Swin-IR [5] takes advantage of Swin Transformer as a deep feature extraction module on image super-resolution and was the SOTA at that time. Later, there are many Swin-Transformer-based super-resolution models introduced, such as HAT [4], and ELAN [1].

2.3. ELAN: Efficient Long-Range Attention Network

ELAN [1] is the main algorithm we use in this research. The architecture is shown in Fig. 1, 2, and 3.

Fig. 1 illustrates the overall network architecture. First, the Low-Resolution image passes the shallow feature extraction module which is a 3×3 convolution layer. Second, it passes the deep feature extraction stage which has multiple number of Efficient Long-Range Attention Block (ELAB). Finally, the image reconstruction stage generates High-Resolution images by both the shallow features and deep features.

Fig. 2 and 3 illustrate the details of ELAB. After local feature extraction, then the model does GMSA (Group-wise Multi-scale Self-Attention) which does local self-attention on different scales.

3. METHODS

In the section, we first describe the overall network architecture in Section 3.1. In the remainder section, we then describe the details of our works which are the ways to enhance ELAN.

3.1. Overall Network Architecture

As mentioned before, this paper is based on the ELAN model, and we enhance it by modifying some architecture. Like ELAN, there consists of three modules: shallow feature extraction, deep feature extraction, and image reconstruction. The overall network architecture of our model is shown in Fig. 1.

3.1.1. Shallow Feature Extraction

Given a Low-resolution (LR) input $I_{LR} \in R^{H\times W\times Cin}$ (H, W and Cin are the image height, width and input channel number, respectively), we use both 3×3 and 5×5 convolution layer $H_{SF}(\cdot)$ which is modified by ourselves to do enhancement to extract more shallow feature $I_{SF} \in R^{H\times W\times C}$ as

$$I_{SF} = H_{SF}(I_{LR}) \quad (1)$$

where C is the feature channel number, and $I_{SF} \in R^{H\times W\times C}$ denotes the output of shallow feature extraction. The convolution layer can help Transformer have more stable and better results in computer vision [20]. The details are shown in Fig. 2.

3.1.2. Deep Feature Extraction

Then, we extract deep feature $I_{DF} \in R^{H\times W\times C}$ from I_{SF} by $H_{DF}(.)$ as

$$I_{DF} = H_{DF}(I_{SF}) \quad (2)$$

where $I_{DF} \in R^{H\times W\times C}$ denotes the output of deep feature extraction.

3.1.3. Image Reconstruction

As the same as ELAN, We use 3×3 convolution and pixel shuffle to reconstruct the image $I_H \in R^{(SxH)\times(SxW)\times Cout}$ as

$$I_H = H_{RC}(I_{SF}+I_{DF}) \quad (3)$$

where H_{RC} (.) is the image reconstruction function, S is the scale factor, and Cout is the number of output channel.

3.1.4. Loss Function

Following ELAN, we optimize the weights of our model by minimizing the pixel-wise L_1 loss *L*:

$$L = \frac{1}{N}\sum_{i=1}^{N}\left|\mathrm{I_{H,i}} - \mathrm{G_{H,i}}\right| \quad (4)$$

where *N* is the number of the ground truth HR images $\{\mathrm{G_{H,i}}\}_{\mathrm{i=1\ to\ }N}$.

Fig. 2 is the modified architecture of ELAN which we propose to improve the performance from the original one. In each subsection, we will describe the details.

3.2. Data Augmentation

Unlike image classification task, the data augmentation methods of image super-resolution for deep learning are just rotate and flip the image. This is because image super-resolution is low-level vision task while image classification is high-level vision task. Some methods, like resize the image, may destroy low-level feature of an image.

There is a paper – Cutblur [16] proposing a hybrid data augmentation method of image super-resolution. However, nobody takes advantage of it in the recent super-resolution model. We will integrate it with ELAN that the training image of ELAN is generated by Cutblur to improve the performance of the model.

3.3. ISFE: Inception Shallow Feature Extraction

To enhance shallow feature extraction, we use an Inception-like convolution layer, which has two branches, 3×3, and 5×5, respectively. The output channel of both branches is half of C. C is the output channel after entering the original shallow feature extraction module. After feature extraction, we concatenate two results into one and enter a 1×1 convolution layer. And the number of output channel is also C. The illustration of this module is shown in Fig. 7.

3.4. Dense Connections to the ELAN Model

The idea is based on this paper [2]. It thinks that dense connections can extract more useful information when training a super-resolution model. Hence, this paper adds dense connections to the deep feature extraction phase of the ELAN [1]. The illustration of this module is shown in Fig. 4.

3.5. RELAB: Residual ELAB

Residual connection [27] is a technique used in deep learning to help train very deep neural networks. They work by adding a shortcut path from the input of a layer to the output of the layer. It can benefit the performance a lot when we do deep learning. Hence, we decide to add residual connections in each dense block. The illustration of this module is shown in Fig. 5, and 6. To keep the number of channel in each place, we add such a mechanism.

4 EXPERIMENT SETUP

4.1. Datasets and Evaluation Metrics

We employ the DIV2K dataset [10] with 800 images to train my model. We test our model on four classical super-resolution benchmarks: Set5 [11], Set14 [12], BSD100 [13], and Urban100 [14] dataset.

PSNR and SSIM are used to be the evaluation metrics, which are calculated on the Y channel after converting RGB to YCbCr format.

The definition of PSNR which maximum possible pixel value is 255 is:

$$\log_{10}\frac{255^2}{\frac{1}{3mn}\sum_{c\in R,G,B}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}[I(\mathrm{i,j,c}) - G(\mathrm{i,j,c})]^2}$$

where m and n are height and width of the image, respectively, *I*(.) represents SR image, and *G*(.) represents ground truth HR image.

The definition of SSIM which maximum possible pixel value is 255 is:

$$[l(x,y)]^{\alpha}[c(x,y)]^{\beta}[s(x,y)]^{\gamma}$$

where *x* and *y* represents SR and ground truth HR image, respectively. *l*(.) represents luminance, *c*(.) represents contrast, and *s*(.) represents structure, respectively.

4.2. Hardware

We use a single Nvidia 3080 which has 12 GB GPU memory to train my model.

4.3. Training Details

We employ ELAN as the original backbone model. Because of the hardware limitation, the patch size is set to be 128×128, the batch size is set to be 32, the number of ELAN blocks (ELABs) is set to be 16, and the number of channels is set to be 48. The loss function is L1, which is the absolute difference between the inference SR image and the ground truth HR image. The GPU memory usage is about 6 GB.

The number of epochs is 120. The learning rate is 0.0002, and be half after 50, 80, 90, 95, and 100 epochs. Model is trained using the ADAM optimizer with $\beta 1 = 0.9$, $\beta 2 = 0.999$, and $\epsilon = 10^{-8}$.

We first train the original model which we call "Baseline" in the next section by our parameter setting. Next, we train our proposed model and compare the

performance with the first one. For fair comparison, the number of ELABs, channels, patch size, batch size, and epochs of our proposed model is the same as the original one. For the original model, the data augmentation method is random rotate 90°, 180°, 270°, or horizontal flip. Our proposed model employs advanced data augmentation methods like Cutblur [18].

5. RESULTS

5.1. Performance

Tables 1, 2, and 3 are the results of our proposed model at different scale. We put bicubic and baseline method together to let readers can compare with other methods conveniently. It is obvious that our model has great improvement after enhancing by our method. When the dataset is Urban100, or when the scale is large, the improvement between our method and baseline is much more relevant.

Table 1(a): PSNR X2 (dB).

Method	Set5	Set14	B100	Urban100
Bicubic	33.64	30.22	29.55	26.66
RAISR[29]	36.15	32.13	-	-
A+[28]	36.54	32.28	31.21	29.20
Baseline	37.89	33.46	32.11	31.34
Ours	**37.90**	33.45	**32.15**	**31.44**
	+0.01	-0.01	+0.04	+0.10

Table 1(b): SSIM X2.

Method	Set5	Set14	B100	Urban100
Bicubic	0.9292	0.8683	0.8425	0.8408
RAISR	0.9510	0.9020	-	-
A+	0.9544	0.9056	0.8863	0.8938
Baseline	0.9603	0.9171	0.8992	0.9303
Ours	**0.9605**	**0.9173**	**0.8997**	**0.9317**
	+0.0002	+0.0002	+0.005	+0.014

Table 2(a): PSNR X3 (dB). Urban100 when scale=3 is not available.

Method	Set5	Set14	B100	Urban100
Bicubic	30.38	27.53	27.20	-
RAISR	32.21	28.86	-	-
A+	32.58	29.13	28.29	-
Baseline	33.93	30.14	28.87	-
Ours	**34.16**	**30.21**	**28.98**	-
	+0.23	+0.07	+0.09	-

Table 2(b): SSIM X3. Urban100 when scale=3 is not available.

Method	Set5	Set14	B100	Urban100
Bicubic	0.8678	0.7737	0.7382	-
RAISR	0.9010	0.8120	-	-
A+	0.9088	0.8188	0.7835	-
Baseline	0.9233	0.8383	0.8001	-
Ours	**0.9252**	**0.8401**	**0.8028**	-
	+0.0019	+0.0018	+0.0027	-

Table 3(a): PSNR X4 (dB).

Method	Set5	Set14	B100	Urban100
Bicubic	28.42	25.99	25.96	23.14
RAISR	29.84	27.00	-	-
A+	30.28	27.32	26.82	24.32
Baseline	31.92	28.41	27.45	25.68
Ours	**32.07**	**28.56**	**27.53**	**25.85**
	+0.15	+0.15	+0.08	+0.17

Table 3(b): SSIM X4.

Method	Set5	Set14	B100	Urban100
Bicubic	0.8101	0.7023	0.6672	0.6573
RAISR	0.8480	0.7380	-	-
A+	0.8603	0.7491	0.7087	0.7183
Baseline	0.8910	0.7769	0.7328	0.7726
Ours	**0.8932**	**0.7806**	**0.7349**	**0.7781**
	+0.0022	+0.0037	+0.021	+0.055

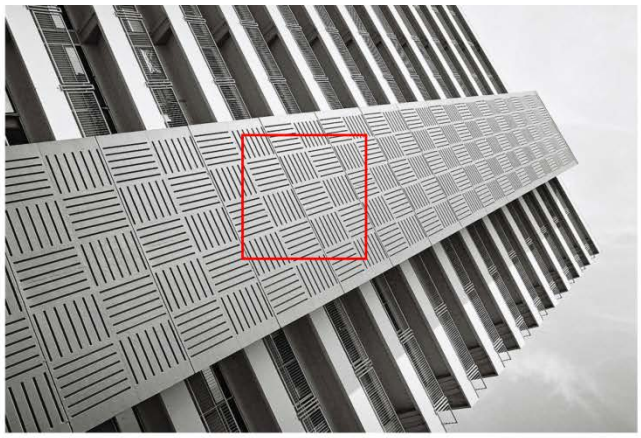
5.2. Qualitative Comparison

We show the results in Fig. 3 and 4.

6. CONCLUSION AND FUTURE WORKS

In this paper, we proposed methods to enhance ELAN for image super-resolution. With Cutblur, Inception Shallow Feature Extraction (ISFE), dense and residual connection, the model improves a lot in comparison with the baseline model, especially when the super-resolution scale is large, or the dataset like Urban100 has regular and dense geometric features.

In our future works, we will still investigate better super-resolution methods.



HR

Baseline Ours

Fig 3: Qualitative comparison of SR models for X4 upscaling. The image is from Urban100.

REFERENCES

[1] Zhang, X., Zeng, H., Guo, S., & Zhang, L. (2022, October). Efficient long-range attention network for image super-resolution. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVII* (pp. 649-667). Cham: Springer Nature Switzerland.

[2] Tong, T., Li, G., Liu, X., & Gao, Q. (2017). Image super-resolution using dense skip connections. In *Proceedings of the IEEE international conference on computer vision* (pp. 4799-4807).

[3] Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., & Timofte, R. (2021). Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 1833-1844).

[4] Chen, X., Wang, X., Zhou, J., & Dong, C. Activating More Pixels in Image Super-Resolution Transformer. arXiv 2022. *arXiv preprint arXiv:2205.04437*.

[5] Dong, C., Loy, C. C., He, K., & Tang, X. (2015). Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, *38*(2), 295-307.

[6] Kim, J., Lee, J. K., & Lee, K. M. (2016). Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1646-1654).

[7] Lim, B., Son, S., Kim, H., Nah, S., & Mu Lee, K. (2017). Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (pp. 136-144).

[8] Kappeler, A., Yoo, S., Dai, Q., & Katsaggelos, A. K. (2016). Video super-resolution with convolutional neural networks. *IEEE transactions on computational imaging*, *2*(2), 109-122.

[9] Mahapatra, D., Bozorgtabar, B., & Garnavi, R. (2019). Image super-resolution using progressive generative adversarial networks for medical image analysis. *Computerized Medical Imaging and Graphics*, *71*, 30-39.

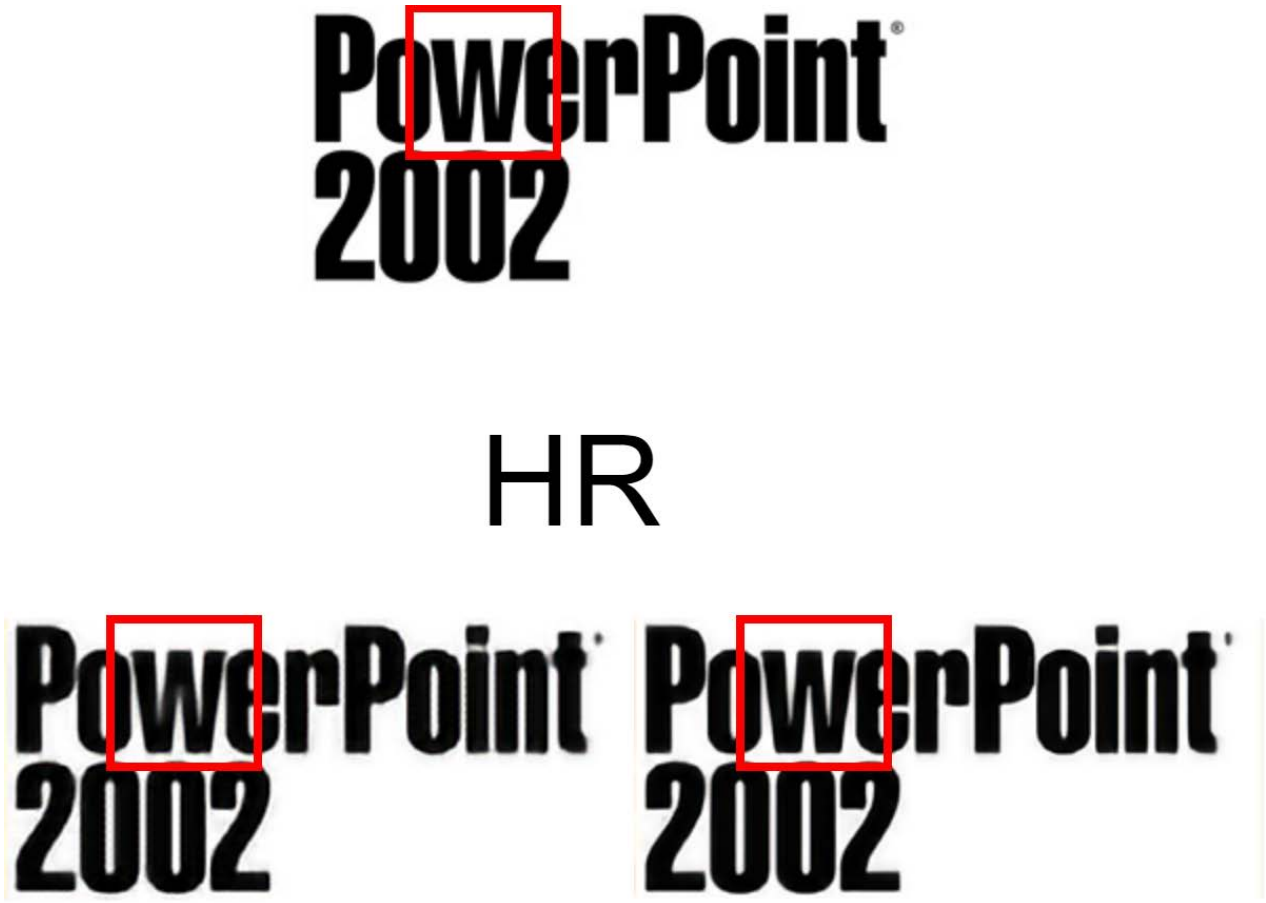


Fig 4: Qualitative comparison of SR models for X4 upscaling. The image is from Set14.

In *Proceedings of the European conference on computer vision (ECCV) workshops* (pp. 0-0).

[10] Agustsson, E., & Timofte, R. (2017). Ntire 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (pp. 126-135).

[11] Bevilacqua, M., Roumy, A., Guillemot, C., & Alberi-Morel, M. L. (2012). Low-complexity single-image super-resolution based on nonnegative neighbor embedding.

[12] Zeyde, R., Elad, M., & Protter, M. (2012). On single image scale-up using sparse-representations. In *Curves and Surfaces: 7th International Conference, Avignon, France, June 24-30, 2010, Revised Selected Papers 7* (pp. 711-730). Springer Berlin Heidelberg

[13] Martin, D., Fowlkes, C., Tal, D., & Malik, J. (2001, July). A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001* (Vol. 2, pp. 416-423). IEEE.

[14] Huang, J. B., Singh, A., & Ahuja, N. (2015). Single image super-resolution from transformed self-exemplars. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5197-5206).

[15] Matsui, Y., Ito, K., Aramaki, Y., Fujimoto, A., Ogawa, T., Yamasaki, T., & Aizawa, K. (2017). Sketch-based manga retrieval using manga109 dataset. *Multimedia Tools and Applications*, *76*, 21811-21838.

[16] Yoo, J., Ahn, N., & Sohn, K. A. (2020). Rethinking data augmentation for image super-resolution: A comprehensive analysis and a new strategy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 8375-8384).

[17] Chen, H., Gu, J., & Zhang, Z. (2021). Attention in attention network for image super-resolution. *arXiv preprint arXiv:2104.09497*.

[18] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

[19] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... & Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1-9).

[20] Xiao, T., Singh, M., Mintun, E., Darrell, T., Dollár, P., & Girshick, R. (2021). Early convolutions help transformers see better. *Advances in Neural Information Processing Systems*, *34*, 30392-30400.

[21] Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., ... & Change Loy, C. (2018). Esrgan: Enhanced super-resolution generative adversarial networks.

[22] Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., & Fu, Y. (2018). Image super-resolution using very deep residual channel attention networks. In Proceedings of the European conference on computer vision (ECCV) (pp. 286-301).

[23] Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., ... & Shi, W. (2017). Photo-realistic single image super-resolution using a generative adversarial network. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4681-4690).

[24] Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009, June). Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition (pp. 248-255). Ieee.

[25] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., ... & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 10012-10022).

[26] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

[27] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).

[28] Timofte, R., De Smet, V., & Van Gool, L. (2015). A+: Adjusted anchored neighborhood regression for fast super-resolution. In Computer Vision--ACCV 2014: 12th Asian Conference on Computer Vision, Singapore, Singapore, November 1-5, 2014, Revised Selected Papers, Part IV 12 (pp. 111-126). Springer International Publishing.

[29] Romano, Y., Isidoro, J., & Milanfar, P. (2016). RAISR: rapid and accurate image super resolution. IEEE Transactions on Computational Imaging, 3(1), 110-125.