Adversarial Learning for Image Super Resolution Using Auxiliary Texture Feature Attributes

Zhenxing Huang†
Wuhan National Laboratory for Optoelectronics
Huazhong University of Science and Technology
Wuhan, China
e-mail: huangzx@hust.edu.cn
† These authors contributed equally.

Dongdong Liu†
Wuhan National Laboratory for Optoelectronics
Huazhong University of Science and Technology
Wuhan, China
e-mail: ddliu@hust.edu.cn
† These authors contributed equally.

Wei Chen
Wuhan National Laboratory for Optoelectronics
Huazhong University of Science and Technology
Wuhan, China
e-mail: chw@hust.edu.cn

Ping Lu
School of Computer Science & Technology
Huazhong University of Science and Technology
Wuhan, China
e-mail: luping06@hust.edu.cn

Jincai Chen*
Wuhan National Laboratory for Optoelectronics
Huazhong University of Science and Technology
Wuhan, China
*Corresponding author e-mail: jcchen@hust.edu.cn

***Abstract*—Recently, image super resolution (SR) has made great progress in computer vision (CV) domain. Due to fine-grained features by convolution neural network (CNN), the mapping function among low resolution (LR) images and high resolution (HR) counterparts could be learned. Textures, which exerts relevant features measured by texture description algorithms, such as gray level co-occurrence matrix (GLCM), among local image patches, are such vital for image SR to recover high frequency details. However, most previous deep learning based methods directly estimate high resolution images ignoring these similar texture attributes. In this paper, we attempt to seek auxiliary texture feature attributes from image patches and cluster them into group indexes so as to improve the reconstruction quality. For the sake of more latent contextual information, a wide residual block is introduced. What's more, we employ 4 loss functions, including adversarial loss, perceptual loss, textural group-conditional loss as well as total variation (TV) loss, tending to estimate high resolution images with realistic textures during the adversarial training. Extensive experimental results on several benchmark datasets demonstrate the effectiveness of proposed method on improving visual quality.**

Keywords—Adversarial learning, Texture Feature Attribute, Group Index, Textural Group-conditional Loss

I. INTRODUCTION

Image super resolution (SR) is a hot research theme that has gotten much attention in recent years. The purpose of SR is to recover a high-resolution (HR) image from its low resolution (LR) one. Based on the fact that LR image could be down-sampled from several different HR images, the inverse process to recovery high resolution image is considered as a highly ill-posed problem.

Progress has been witnessed that deep learning (DL) based methods gradually become the mainstream in SR domain because of the powerful capability of deep convolution neural networks (CNN) to improve the performance of SR results. Since Dong et al. [1] primarily use 3 convolution layers to establish reconstruction of HR images leading to the extraordinary performance, plenty of CNN frameworks are studied for SR. In order to accelerate the speed of SR, ESPCNN[2] and FSRCNN[3] tend to extract features in the shallow layer and upsample image at the end of network parts through sub-pixel and transposed convolution. Inspired by residual learning[4], several works [5-8] try to modify the residual structure for SR resulting in greatly explosive quality improvement. Particularly, EDSR[6] achieves significant improvement by removing unnecessary Batch normalization (BN) layers. Ahn et al. [9] consider another issue to reduce network parameters and operations so as to build a lightweight model for saving computing resources. To obtain more useful channel-wise features, Zhang et al. [10] introduce the attention mechanism into SR by designing deep residual channel attention blocks. On the other hand, unsupervised learning are also employed to tackle this problem. By introducing perceptual loss and adversarial loss, Ledig et al. [11] propose SRGAN model and obtain more natural images. However, most previous methods ignore similar textural attributes of LR image patches. Generally, auxiliary attributes could lead to a perceived performance improvement, for instance, the facial images with informative attributes (color, hair and etc.) description assist face hallucination.

In this paper, we propose an adversarial network framework, named SRAAGAN, with extra textural feature attributes for the sake of better perceptual visual results. In terms of the textural feature attribute, it is extracted from a sub-image through gray level co-occurrence matrix (GLCM) and these attributes are clustered into different textural group indexes. Then, these auxiliary information enhance the SR performance. What's more, the adversarial learning is introduced to solve this problem. In the adversarial model, both the generative and discriminative networks employ the textural indexes. Furthermore, the group-conditional loss is designed to help the SR performance.

The contributions are divided into three folds: 1) High resolution images are generated via LR ones as well as textural feature attributes, which enhances the pertinency in

978-1-6654-1538-5/21/$31.00 ©2021 IEEE
DOI 10.1109/ACCTCS52002.2021.00035

the sub-images with similar textural group indexes. 2)Wide residual blocks are cascaded and dense connections are employed to acquire more contextual information. 3) During the adversarial training, the novel auxiliary textural group-conditional loss is designed. To our knowledge, this is the first study to introduce the textural feature attribute and textural group index to estimate HR images in SR domain.

II. METHODS

The target to solve SR problem is to learn a complex mapping function between HR and LR images, and LR counterparts are down-sampled via the bicubic interpolation approach. Different from most previous DL-based methods, textural attribute priors are employed as input information. Based on discrepant textural feature attributes from sub-images, we consider to sort out sub-images with similar attribute group indexes to enhance the correlation in reconstruction progress. Since it's a great challenge to define a explicit label for a patch from an image, cluster tag comes into our mind to create a group index for textural feature attribute.

A. Auxiliary Attributes

Inspired by ACGAN[12], input images with label information could generate high resolution images acquiring a superb visual quality. However, it is a challenging problem to give a label for each patch in SR domain. Thus, similar texture feature vectors of patches are considered to be united into one group as an auxiliary tag. As is shown in Fig. 1, auxiliary group indexes are obtained through three steps: 1)Textural feature attributes are extracted for total patches from a limit training dataset; 2) K-means is used to cluster feature attribute stacks so as to obtain the group index predictor; 3)The clustered predictor is utilized to seek the group index for a new patch.

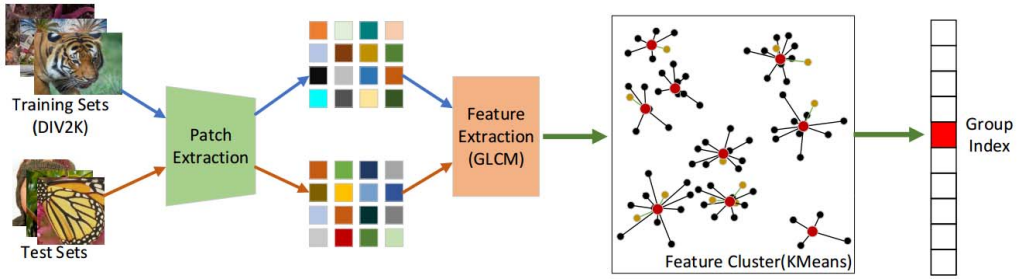
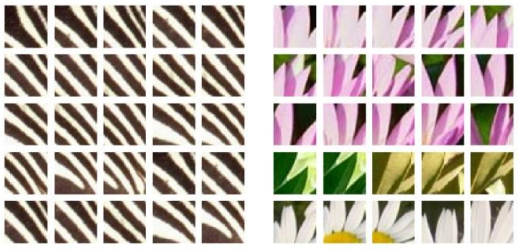


Figure 1: The progress of textural group index estimation.

To trade off the faithful texture generator and the complexity of feature extraction, we empirically set patch size as 32. Secondly, gray level co-occurrence matrix is employed to get textures feature attributes, which consist of correlation, contrast, homogeneity and energy in four directions (0°, 45°, 90° and 135°). In Fig. 2, the proposed predictor shows efficiency in clustering sub-images. Based on observation, we find that sub-images with "zebra stripe" mostly appear in group 15 and results of group 25 consist of images with plant tips.



(a) Sub-images in group 15 (b) Sub-images in group 25

Figure 2. The clustered predictor is applied to distinguish group index of sub-images in Image 91.

B. Main Framework

In Fig. 3, the overview framework roughly consists of the generator network and the discriminator part. For the generator network, an input image with auxiliary textural attributes information will produce estimated HR one. On the other hand, the discriminator network not only discerns the ground truth but also obtains the textural group index. Because the additional textural feature attribute relies on the luminance components, another input information comes from *Y* channel in *YCbCr* color space. In Fig. 4, x stands for input image, y stands for ground-truth image and $\tilde{y}$ stands for estimated image. What's more, G is the generator network and D is the discriminator network. In structure B, c is label of input image. In structure C, P is a group index predictor and x'_c is the group index c' of input image x. Compared with other GANs' structure for SR, our model appends the novel textural group index predictor.

The wide residual block primarily concatenates several feature maps from each branch as input information. A convolution operation with kernel size $1 \times 1 \times 64$ helps to squeeze feature channels. Inspired by WDSR[13], feature channels are expanded to obtain more latent features. What's more, the last convolution layer would compress the feature numbers and a local short path helps to reserve major feature information. *Concat* is used to unite these information from different input blocks. Instead of *ReLU*, *Leaky ReLU* is applied avoid restraining negative signal.

C. Adversarial Learning

Our SRAAGAN model produces the estimated result $\tilde{y}$ from the input LR image x through the generator network G with auxiliary textural group index. Then, the discriminator network D gives both a distribution of sources and a predicted group tag. To accelerate the convergence, the Wassertein distance is employed. Before training the adversarial model, a pre-trained model for generative network is recommended and optimized by mean square error (MSE).

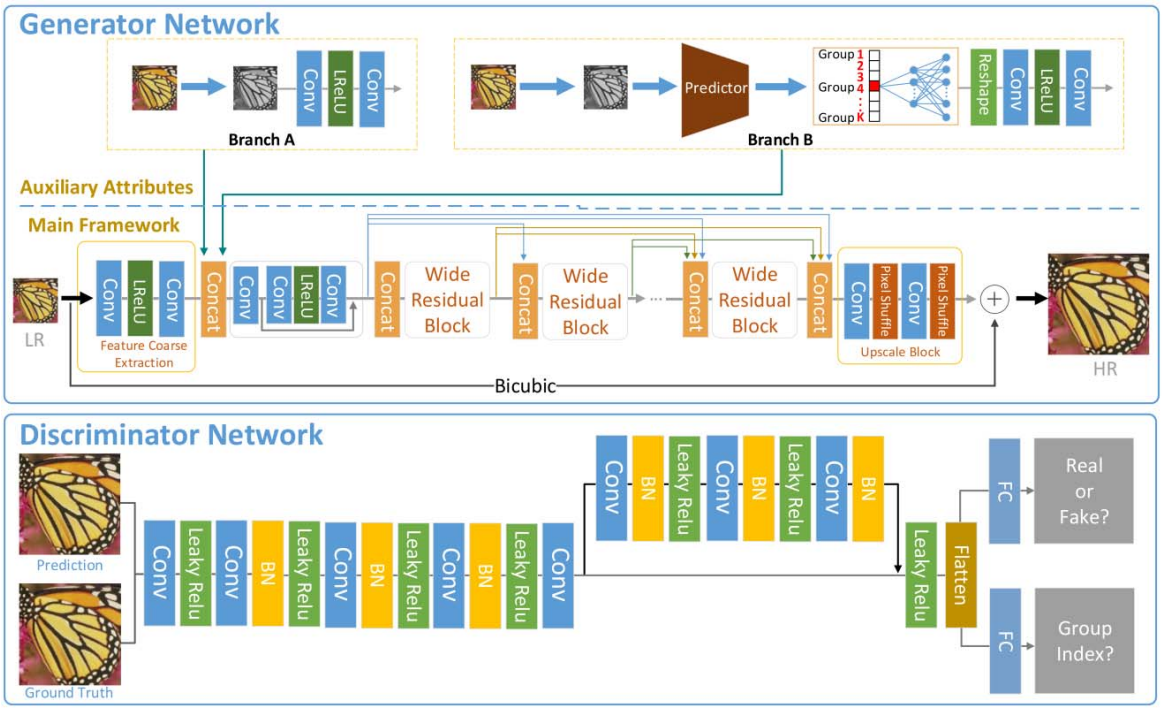


Figure 3: The framework of our proposed model.

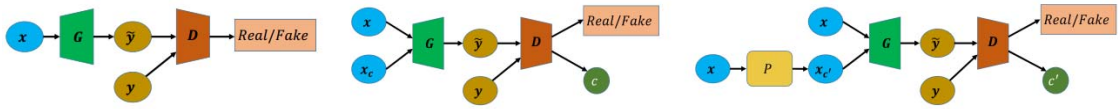


Figure 4: Adversarial learning for SR. (a)Structure A: traditional generator adversarial network for SR (like SRGAN). (b)Structure B: generator adversarial network with auxiliary class label for SR similar to ACGAN. (c)Structure C: our generator adversarial network with auxiliary group index label for SR

The textural group-condition loss is formulated as:

$$L_d = E_x\left[D_s\left(G(x)\right)\right] - E_y\left[D_s(y)\right] \quad (1)$$

Lc=ExDcGx−C+EyDcy−C#1Where C is the one-hot code for the group tag. D_c represents the group label prediction operation in discriminator network. E_x and E_y denote the expectation operation. Besides, the adversarial loss for input image is formulated as:

$$L_d = E_x\left[D_s\left(G(x)\right)\right] - E_y\left[D_s(y)\right] + \beta E_{\hat{x}}\left[\|\nabla_{\hat{x}} D(\hat{x})\|_2 - 1\right] \quad (2)$$

Where D_s represents the true or fake prediction in discriminator network.

Perception loss is also used in our model to chase better visual quality, which is formulated as:

$$L_p = E_x\left[\psi\left(G(x)\right) - \psi(y)\right] \quad (3)$$

Where ψ obtains perception results through Vgg 19[14].

The total variation (TV) restriction is introduced to impose spatial smoothness against noise graphics, which is formulated as:

$$L_{TV} = E_x[\nabla G(x)] \quad (4)$$

Where ∇ is a gradient function to compute gradient of $G(x)$ in both horizontal and vertical directions.

Thus, the ultimate objective loss could be described as:

$$L_U = L_d + \gamma L_c + \lambda L_p + \mu L_{TV} \quad (5)$$

Where γ, λ and μ are balance factors to trade off different cost functions.

III. Experiments

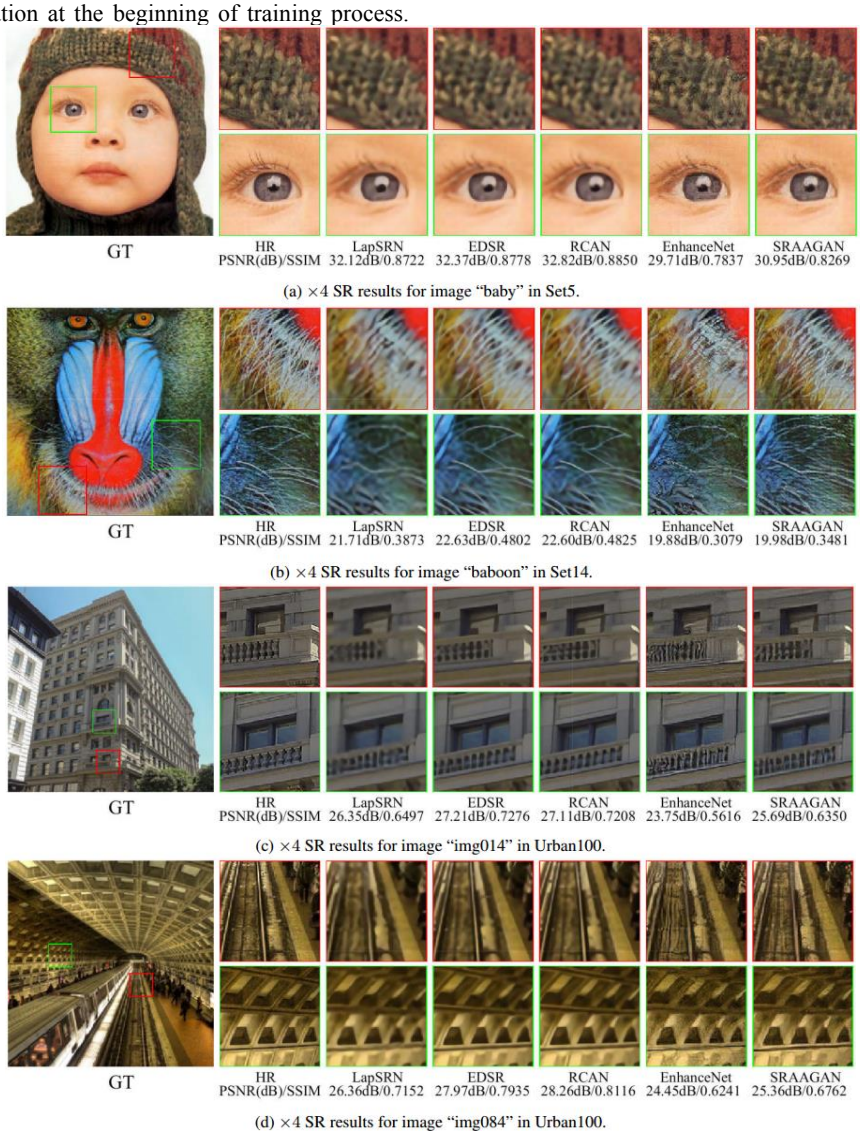
A. Data Sources

There exists several datasets for image SR, such as 91 images and BSD dataset. Recently, the new public DIV2K training dataset with sufficient images are widely used. In terms of test, we apply several standard benchmark datasets, including Set 5, Set 14, BSD 100 and Urban 100.

B. Implementations Details

The training process is made up of two steps. Before training SRAAGAN, we acquire a pre-trained model through Adam optimizer[15] by setting $\beta 1 = 0.9$; $\beta 2 = 0.999$ and the initial learning rate is $1e^{-4}$. The batch size of pre-trained model is set as 64. The sub-image is cropped with patch size

32 × 32. Next, we train the adversarial model through RMSprop optimizer with weights decay. The batch as 0.0001 and decreases as a factor 0.99 in each 50 epochs. The pre-trained model helps the discriminator part to focus more on texture discrimination at the beginning of training process. Two settings of generator are used: the light model contains 8 wide residual blocks and the deeper model uses 20 WRB blocks. We implement our models with the Tensorflow framework and train them with NVIDIA Titan XP GPUs.



(a) ×4 SR results for image "baby" in Set5.

(b) ×4 SR results for image "baboon" in Set14.

(c) ×4 SR results for image "img014" in Urban100.

(d) ×4 SR results for image "img084" in Urban100.

Figure 5: ×4 super resolution example results for different methods

TABLE I. TEST RESULTS (PSNR(DB)/ SSIM) FPR DIFFERENT METHODS ON PUBLIC DATASETS WITH A SCALE FACTOR ×4.

Methods		Set5 PSNR/SSIM	Set14 PSNR/SSIM	BSD100 PSNR/SSIM	Urban100 PSNR/SSIM
CNN-based	LapSRN	31.54/0.885	28.19/0.772	27.32/0.727	25.21/0.756
	EDSR	32.46/0.897	28.80/0.788	27.71/0.742	26.64/0.803
	DBPN	32.47/0.898	28.82/0.786	27.72/0.740	27.08/0.795
	RCAN	32.63/0.900	28.87/0.789	27.77/0.744	26.82/0.809
GAN-based	SRGAN	29.34/0.834	25.83/0.694	25.19/0.641	-/-
	EnhanceNet	28.46/0.809	25.51/0.678	24.95/0.627	23.54/0.694
	SRAAGAN(Ours)	29.70/0.847	26.12/0.707	25.25/0.649	24.47/0.731

C. *Results and Ablation Analysis*

We evaluate proposed model on several public benchmark datasets with several CNN-oriented methods and perceptual-driven approaches. CNN-based methods, including LapSRN[16], EDSR[6], DBPN[17] and RCAN[10] are employed. On the other hand, we also consider perceptual-driven approaches including SRGAN[11] and EnhanceNet[18]. As common image quality metrics, PSNR and SSIM are adopted to evaluate constructed HR images with its homologous ones.

As is observed from Fig. 5, the proposed SRAAGAN generates more sharpness and details. For instance, SRAAGAN produces natural skin textures than other methods in image "baboon" from Set14. Compared with EnhanceNet, the results of our generator is more natural without unpleasing noise. What's more, the superiority of SRAAGAN is proved in Fig. 6. In TABLE. I, CNN-based methods generally perform better in PSNR and SSIM. Our model gains certain advantages compared with other three GAN-based methods. In Fig. 7, our SRAAGAN outperforms compared with the well known EnhanceNet in terms of lower PI. The total variation (TV) restriction is added to impose spatial smoothness against noise graphics in Fig. 8. Compared with non-TV loss, the output is more natural with TV loss. A light model with 8 wide residual blocks trained for $1e^5$ epochs, is used to compare performances in different group number and fusion ways. As for the total group number, K is set as 128, which will improve reconstruction quality compared with 32, 64 as well as 256 when we apply a light model to validate the result (in Fig. 9a). As shown in Fig. 9b, it well demonstrates that multi-sources information does benefit the performance of reconstruction results. Besides, we treat the model without branch information as baseline. Then we compare the baseline model with different branch settings (A, B or A+B). Results prove that textural feature information fusion would avail PSNR.

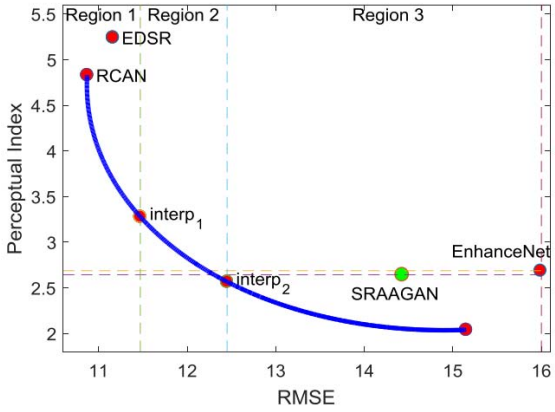


Figure 7. Perceptual index comparison for different methods.

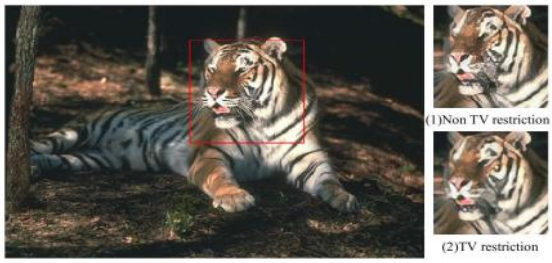
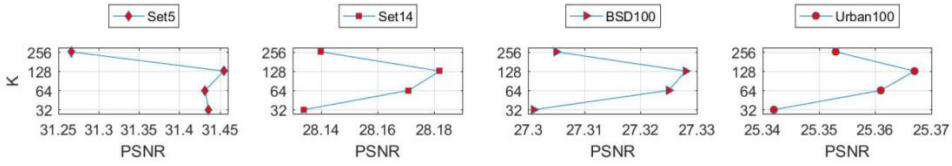


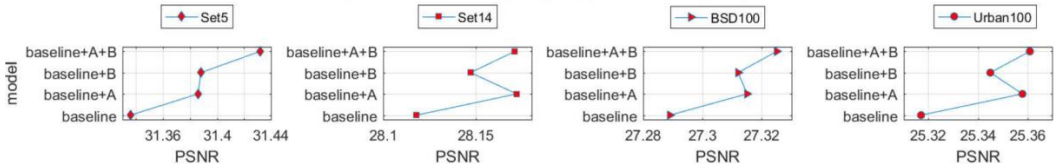
Figure 8.× 4 super resolution results of image '108082' in BSD 100.

(a)GT (b)SRGAN (c)EnhanceNet (d)SRAAGAN

Figure 6: ×4 super resolution example results for GAN-based methods.



(a) ×4 SR performance of our light model by four group tag predictors in group number 32,64,128 and 256.



(b) ×4 SR performance of our light model in multi-sources fusion ways when K is set as 64.

Figure 9. Performance comparison for different parameter settings.

IV. Limitation and Conclusion

Since our proposed method works on image patches, the test results need to conduct image stitching, which limits the performance of quality evaluation, for instance PSNR and SSIM. Although the visual quality of our output HR images is better compared with CNN-based methods, there should exist a trade-off between perceptual quality and distortion measures. Besides, our perceptual index scores may be not the best one, the visual quality is delightedly curved as our model gains an advantage over EnhanceNet. In the future, we tend to pursue a better way to acquire auxiliary attributes. As for attributes features, SIFT may be taken into account for its scale-invariant. Furthermore, CNN-based approaches will be considered to extract similar feature for analogous content as far as possible.

We proposed a SRAAGAN model with auxiliary attributes that achieves a promising visual quality compared with several SR methods. In our generator part, a WRB block helps to enhance the effectiveness of convolution operations. For the sake of better information fusion, the dense connections are employed and multi-sources data is concatenated as input information. As for adversarial training, 4 loss functions restrain the network to produce high resolution results avoiding introducing superfluous noise. Moreover, the auxiliary attributes and group indexes guide the generator to recover more realistic textures. Ablation studies prove that the auxiliary attributes avail the performance for SR outputs. Experiment results show the extra textural feature attributes and group index strengthens the capability of network framework. In the future, we will continue seeking for more appropriate auxiliary attributes beyond current visual quality in an effective way.

Acknowledgement

This work was supported by National Natural Science Foundation of China (61672246, 61272068). We gratefully acknowledge the support from NVIDIA Corporation for providing us the Titan XP GPU used in this research.

References

[1] 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, 2015.

[2] W. Shi et al., "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, 2016, pp. 1874-1883.

[3] C. Dong, C. C. Loy, and X. Tang, "Accelerating the super-resolution convolutional neural network," in European conference on computer vision, 2016: Springer, pp. 391-407.

[4] 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, 2016, pp. 770-778.

[5] J. Kim, J. Kwon Lee, and K. Mu Lee, "Accurate image super-resolution using very deep convolutional networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 1646-1654.

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

[7] 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, 2017, pp. 4799-4807.

[8] Y. Tai, J. Yang, and X. Liu, "Image super-resolution via deep recursive residual network," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 3147-3155.

[9] N. Ahn, B. Kang, and K.-A. Sohn, "Fast, accurate, and lightweight super-resolution with cascading residual network," in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 252-268.

[10] 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), 2018, pp. 286-301.

[11] C. Ledig 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, 2017, pp. 4681-4690.

[12] A. Odena, C. Olah, and J. Shlens, "Conditional image synthesis with auxiliary classifier gans," in International conference on machine learning, 2017: PMLR, pp. 2642-2651.

[13] J. Yu et al., "Wide activation for efficient and accurate image super-resolution," arXiv preprint arXiv:1808.08718, 2018.

[14] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.

[15] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.

[16] 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, 2017, pp. 624-632.

[17] M. Haris, G. Shakhnarovich, and N. Ukita, "Deep back-projection networks for super-resolution," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 1664-1673.

[18] M. S. Sajjadi, B. Scholkopf, and M. Hirsch, "Enhancenet: Single image super-resolution through automated texture synthesis," in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 4491-4500.