FBRNN:feedback recurrent neural network for extreme image super-resolution

Junyeop Lee[1] Jaihyun Park[1] Kanghyu Lee[1] Jeongki Min[1] Gwantae Kim[1] Bokyeung Lee[1]Bonhwa Ku[1]

David K. Han[2] Hanseok Ko[1]

[1]Korea University, Seoul, Korea [2]Army Research Laboratory, Adelphi, MD, USA

Abstract

Single image extreme Super Resolution (SR) is a difficult task as scale factor in the order of 10X or greater is typically attempted. For instance, in the case of 16x upscale of an image, a single pixel from a low resolution image gets expanded to a 16x16 image patch. Such attempts often result fuzzy quality and loss in details in reconstructed images. To handle these difficulties, we propose a network architecture composed of a series of connected blocks in recurrent and feedback fashions for enhanced SR reconstruction. By use of recurrent network, an SR image is refined over a sequence of enhancement stages in coarse to fine manner. Additionally, each stage involves back projection of SR image to LR images for continuously being refined during the sequence. According to the preliminary results of NTIRE 2020 Perceptual Extreme SR challenge, our team (KU_ISPLB) secured 6th place by PSNR and 7th place by SSIM among all participants.

1. Introduction

Single image super-resolution (SISR) takes a low-resolution image and estimates its high-resolution image. An earlier method, such as bicubic interpolation, tries to fill in missing information between pixels by interpolation, thus it does not require training data [3, 7]. Although these methods preserve gross image structures, the interpolation schemes do not guarantee in recovering fine details in HR (High Resolution) images, and often produce fuzzy or blurred images. While it can be said that these methods exploit information from surrounding pixels, the methods have no means of recovering information from correlations among image patches and their semantics. Learning based methods have shown to be effective in exploiting these correlations when given a large set of labeled data. As such, deep learning methods[9, 10, 16] have demonstrated successes on restoring blurred parts into higher contrast, essentially recovering fine image details. Among the learning based models, one of the most widely used models is SRCNN[2]. It delivers super resolution in an end-to-end fashion by using a convolution neural network (CNN), and many of the later models are based on its architecture. While its end-to-end structure is simple, due to its relatively

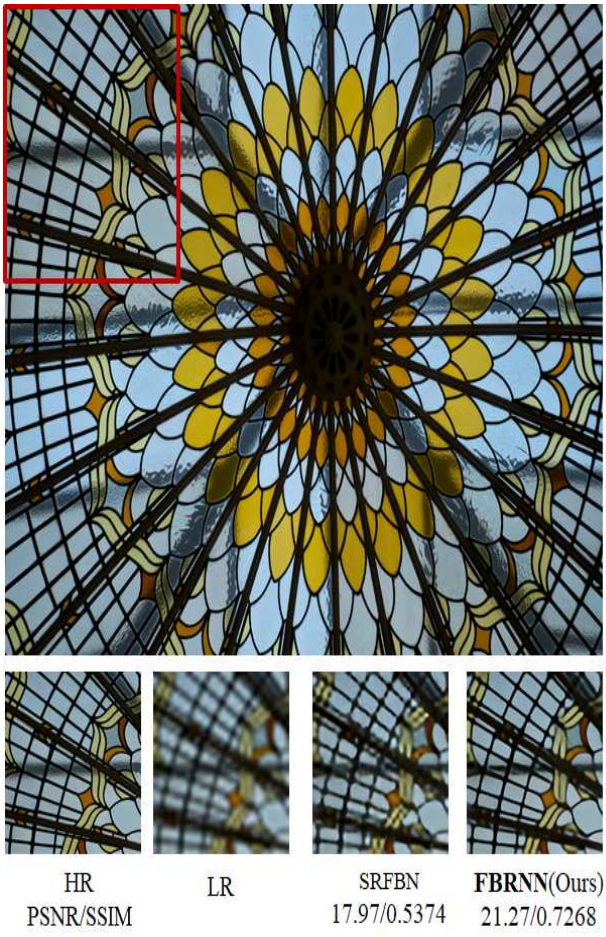

Figure 1. The result of x16 super resolution of proposed model on 0821 from DIV2K validation set compared with baseline model

shallow depth, SRCNN does not fully exploit low level image features for recovering fine details. Recently, many learning based methods focusing on effective recovery of high-frequency details have been proposed by employing deeper network layers to capture low level features in an end-to-end manner.

Previously, the learning-based model[1, 21, 23] and other methods generally performed resolution restoration on an x4 scale. However, with availability of greater computational capacity, super resolution on higher scales has been possible with increase in data set size[4]. Each time the image scale is increased by a factor n, the image output size increases to n squared. Accordingly, for nx super-resolution reconstruction, the model needs to project one pixel of a low image to a nxn pixel image patch. It also means that the load of the Ground Truth images to be reflected in the learning process becomes significantly higher by a factor n squared. Thus with a limited dataset, insuring high SR performance becomes a major challenge.

In NTIRE 2020, an extreme super resolution of scale factor 16 for a given training / test dataset was defined as the task. The training dataset consists of 8K dataset containing high resolution images with dimension up to 5000x5000 or more and low resolution images up to 500x500 dimension. When a high resolution gets reduced in scale by 16 times, loss of small details in HR image from down sampling process poses as the main difficulty. Therefore, the key focus of SR is to restore fine details lost in LR images to be similar to the detail of GT images.

In this work, we develop a feedback recurrent neural network that employs a recurrent structure coupled with a feedback for enhancing low resolution images gradually to solve NTIRE 2020 Image SR challenge. First, we abstract and preserve salient features of Low Resolution (LR) image input and sequentially transmit them to the next network module through a residual dense network(RDN)[24] via an upsampling network. Second, a reconstructed SR image at the current stage is passed through a down-projection network and merged with a previously enhanced LR image via a 1X1 CONV layer to produce an improved LR image for the next SR enhancement sequence.

In summary, our main contribution can be expressed in three folds.

- We propose a feedback recurrent network (FBRNN) architecture for enabling large scale super resolution that effectively extracts features from the original LR images progressively in a coarse to fine fashion by employing the structural framework of a recurrent network.

- Inspired by DBPN[6] and SRFBN[12], we employ a back projection approach, to gradually increase correlated information previously learned from the training data to the LR images sequentially for improved SR reconstruction.

- For further improvement in LR image, we propose fusing of a current LR image and its subsequent LR image in an adaptive fashion.

The rest of this paper is organized as follows. Section 2 shows related works about recurrent and dense networks for image SR. The proposed feed-back recurrent neural network is described in Section 3. Section 4 provides details of the training strategies and result descriptions of our experiments. Lastly, we provide conclusive remarks in Section 5.

2. Related Work

In this section, we review some of the most widely used SR methods including the state of the art methods. They are broadly divided into feed-forward methods, recurrent methods, and feed-back networks. Their brief descriptions and performances are presented.

2.1. Learning based image super resolution

Research on learning based image super resolution models has been active with the end-to-end model being the most direct and intuitive model among them. In these implementations, learning is driven by a Mean Squared Error(MSE) based loss function computed from reconstructed SR and HR Ground Truth images in supervised learning framework. One of the earlier models in this approach is SRCNN which is composed of relatively small number of CNN layers to develop mapping from LR images to HR images. EDSR[13], developed by Lim et al employed residual blocks, skip connections, and pixel shuffle, and demonstrated robust state-of-the-art performance from multiple datasets. Some studies focused on convergence of loss functions when mapping from LR images to HR images. In EEDS[20], convergence is accelerated by using a shallow network while in many other studies the convergence problem was solved by applying a bias using a bicubic interpolation of LR resolution images. In addition, Generative Adversarial Network (GAN) based on the perception driven learning methods has been proposed for SR[19, 8, 18, 17].

2.2. Recurrent method image super resolution

The ultimate goal of image super-resolution is to obtain robust performance in large SR scale factors. To do so, it is necessary to exploit features of an LR image as much as possible. In an end-to-end training method, the prior information of the LR image can be well utilized iteratively or recurrently in the steps of generating an SR image. RDN is a network that exploits hierarchical characteristics from an original LR image, and obtains reasonable performance by

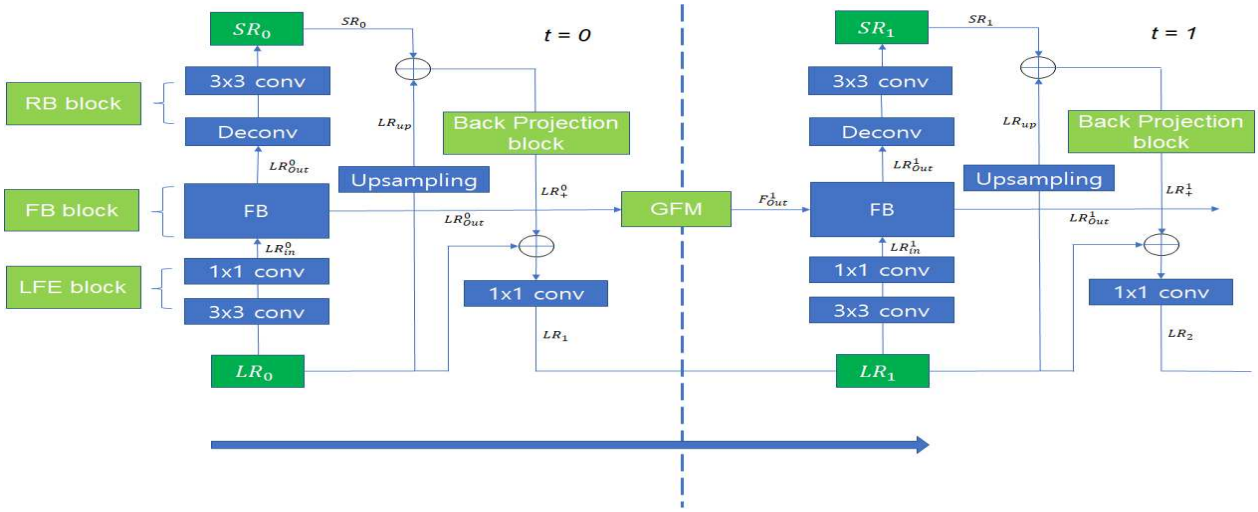

Figure 2. The overall architecture of our proposed feedback recurrent neural network(FBRNN)

mixing global dense features and shallow features in an LR space. Similarly, SRFBN proposed a top-down feedback connection in the LR space, and enabled a fast and powerful reconstruction with only a small number of parameters. SRFBN employs an error-based loss function from HR space to LR space in a curriculum based learning type approach in a coarse to fine fashion by utilizing a recurrent architecture by repeatedly transferring high level information at every step. Our proposed method is inspired in part by SRFBN.

2.3. Back projection network

In addition to the end-to-end method of exploiting the LR input image feature, a method for finally reflecting information about the output image in the network back to the input image has been proposed. The recently announced DPBN utilizes features of HR Space by repeatedly upsampling and downsampling from the feedback format. Also, according to the DPBN, characteristics of the back projection network can ensure robust performance in large scale factors.The feedback network, rather than the one-way mapping feedforward network, ensures good performance at large scale factors that are difficult to converge. HBPN[14] is another latest study that utilizes the back projection block. It captures various spatial correlations using the SR-HG module at all scales and performs softmax based weighted reconstruction.

3. Proposed method

The key idea of the proposed method is repeated use and refinement of lower resolution images in reconstruction of SR in a recurrent structure with back projection of SR to LR for sequential refinement of LR. The dimensional mismatch in the back projection of reconstructed SR in each stage is done by down sampling. At each sequence, embedded features associated with SR reconstructed at each stage is compared to embedded features of HR Ground Truth image for a loss function.

3.1. Network design

As shown in Fig.2, our proposed FBRNN has t steps, and in the subnetwork FB, it has B iterations of residual dense block(RDB)s. First of all, we will denote $conv(s, n)$ and $deconv(s, n)$ as a convolution layer and a deconvolution layer, respectively, where s is size of the filter and n is number of filters. Our subnetwork consists of four parts. In the first LR feature extraction(LFE) block, features of low image are extracted. As [11] does, we adopt the FB block and GFM module in our integrated network with the feed-back(FB) block transmitting information from the low dimensional space. In reconstruction block(RB), SR is constructed through four $deconv$ layers and one 3x3 $conv$ layer, and a global skip connection through bicubic upsampling. The down-projection block is performed through the same four $deconv$ layers step.

3.2. Down-projection network

The down-projection network is an important part of our proposed model. Two main parts of the high scale super resolution scheme are accurate projection of characteristics of the SR space and the training of the model. In the case of the feed-forward network, it is hard to say that the SR image feature is repeatedly involved in training because the SR image appears only once in the training process as end-to-end training when computing the training loss.

$$L(\theta) = \sum_{t=1}^{T} \left\| (I_{HR}^{t} - I_{SR}^{t}) \right\|_1 \quad (1)$$

Our approach aims to change the same input LR image to pass through the network in an iterative way where the weight is tied at high scale super resolution, although the existing method reflects the difference from the HR ground truth in the loss. Therefore, we develop a feedback network in a recurrent manner by adding a down-projection block to the existing network so that an SR image in each iteration is back-projected and merged with the current enhanced LR image as follows.

$$LR_{+}^{0} = f_{BPB}(SR_0, LR_{up}) \tag{2}$$

$$LR_1 = C_0(LR_{+}, LR_0) \tag{3}$$

Additionally, since the down-projection network needs to down-project a high-dimensional image again, it is constructed in the same way as the previous $deconv$ 4 times. In other words, the down-projection block is inverse transformation of the upscale network.

3.3. Sub networks

LRF block basically extracts features from an LR image and LFB is composed of $conv(3, 4n)$ and $conv(s, n)$, and n is the number of predefined channels.

$$LR_{in}^{0} = f_{LFB}(LR^0) \tag{4}$$

LR_{in}^{0} enters the feedback block composed of RDB. We adopt FB sub network as RDB blocks are connected to make up FB Block. B is a predefined number and each block shares the weights. After passing through the FB, the output enters the RB or passes through the GFM (gated feedback module) and all the output of RDB block are weighted to the next FB module. The RB block is composed of 4 $deconv$ layers and 1 3x3 $conv$ layer, and it is suitable to improve performance by passing the deconv layers sequentially 4 times.

$$LR_{out}^{0} = f_{RB}(LR_{out}^{0}) \tag{5}$$

This LR_{in}^{0} enters the feedback block composed of RDB.

Prior to passing through the LFB, the same LR image passes through the $bicubic-upsampling$ block, which acts as a global skip connection.

$$LR_{up} = Bi\_up(LR^0) \tag{6}$$

4. Experimental Results

In this section, we first introduce the dataset with implementation details and conduct a performance evaluation of the proposed model and the state-of-the art models for comparison.

Datasets and metrics. We use DIV8K[5] provided by NTIRE 2020 as our training dataset with no extra dataset. DIV 8K dataset consists of images with resolutions up to 8K, a training set of 1500, and a validation set of 100. PSNR and SSIM are used as performance evaluation metrics. For fair comparison, the same metrics are used for the baseline models (e.g. SRFBN and DBPN). The validation set and test set of DIV8K could not be used for experimental verification because the HR images were not made available. Therefore, we verify the experimental performance using the DIV2K validation set that has already been released.

Ensemble technique. Due to computational limitations, patch size is set by 25x25 and augmentation is performed with an ensemble technique of image patch rotation and symmetry. The dataset is augmented in 8 folds.

4.1. Experiment analysis

We use PRELU[15] for all $conv$ layers and $deconv$ layers, 32 for the batch size, and 7 for the number of blocks. Adam optimizer is used in the training process, with stride of 16, padding of 2, and the kernel size of 20 is set for a scale up of 16. In t = 3, that is, 3 iteration model, the number of parameters is 1.9M, the average runtime is 1.50s, and the total runtime including storage time is 70s. This is the highest PSNR among the models with the number of parameters (green line) below 10M as shown in Fig 3. In the figure, the black dots represent GAN-loss models whereas blue dots represent L1/L2 optimized loss models.

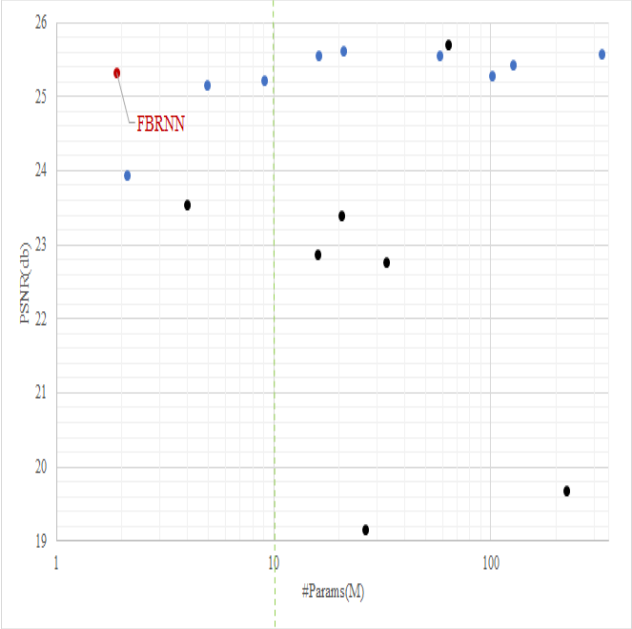

Figure 3. PSNR (db) vs number of parameters (M) of the proposed model compared with other extreme models in the challenge

4.2. Qualitative result

To compare performances between state-of-the-art models and the proposed model, we conduct extensive experiments using datasets such as Manga109, Urban100 and

x4 PSNR/SSIM Result on multiple dataset						
Dataset	bicubic	SRCNN[7]	EDSR[13]	D-DBPN[1]	SRFBN+[12]	Ours
B100	25.96/0.6675	26.90/0.7284	27.71/0.7420	27.72/0.7400	27.77/0.7419	27.85/0.7485
Urban100	23.14/0.6577	24.52/0.7221	26.64/0.8033	26.38/0.7946	26.73/0.8043	27.22/0.8155
Manga109	24.89/0.7866	27.58/0.8555	31.02/0.9148	30.91/0.9137	31.15/0.9160	31.32/0.9117

Table 1. The result of x4 super resolution on our proposed model and other methods on various dataset

x16 PSNR/SSIM/LPIPS Result on challenge				
Dataset	bicubic upsampling	winner-NTIRE20	winner-AIM19	Ours
DIV8K	24.22/0.6017/0.683	23.38/0.5504/0.348	25.63/0.6394/0.554	25.33/0.6299/0.582

Table 2. The result of challenge x16 super resolution on our proposed model.

x16 PSNR/SSIM Result on multiple dataset			
Dataset	D-DBPN[1]	SRFBN+[12]	Ours
DIV2K	27.16/0.6863	28.02/0.6992	28.58/0.7135
Manga109	18.15/0.5530	18.19/0.5725	18.01/0.5843
Urban100	18.86/0.4411	19.07/0.4574	19.32/0.4755
BSDS100	22.04/0.4849	22.36/0.5086	22.58/0.5098

Table 3. The result of x16 super resolution on our proposed model and baseline methods on various dataset

BSDS100 as well as DIV2K validation set. Fig.4 shows the results of 16 times super resolution for the DIV2K validation set, and it is visually apparent that the proposed model achieves superior detail in the SR image. Fig.5 and Fig.6 show qualitative results in Manga109, Urban100, and BSDS100, and these results also show that the proposed model delivers robust performance. The baseline-1 and baseline-2 presented in Figures 3 and 4 refer to DDBPN and SRFBN, respectively. DDBPN is an improved version of DBPN and we used the latest version for accurate comparison. While it is well known that GAN based SR techniques usually deliver better detailed and realistic images, the proposed method is shown to achieve good details and realism, and also demonstrates high performances in terms of quantitative metrics as will be described in the following section.

4.3. Quantitative result

On experimental setting of zooming factor 16, the datasets used in the quantitative result used full images, but the DIV2K datasets are randomly cropped to 576x576 images to reduce experimental computation. Because of the smaller LR size of the dataset, SR reconstruction was found more difficult. Except in Manga109, the proposed method achieves superior performance compared to the baselines. Our model performance on the B100, Urban100, and Manga109 datasets of zooming factor 4 is shown generally better than the baseline methods. Performance measures of ours, winner of AIM19 and NTIRE20 are presented in NTIRE2020[22].

As can be seen in Table 2 and Figure 3, GAN based methods generally show low PSNR and SSIM, but show high performance on LPIPS. GAN based methods are trained to produce realistic images with high contrast and details. Thus they produce more real like images compared to L1/L2 methods. However, they don't necessarily perform well in terms of quantitative metrics. L1/L2 methods, on the other hand, are trained to produce high pixel wise accuracy. Therefore, these methods tend to perform well in terms of PSNR and SSIM, however, they often produce images with fuzzy qualities and reduced details in extreme SR. Our method is an l1/l2 optimized method such as AIM19 presented in Table 2, and should be distinguished from the GAN methods aimed at restoring high frequency. In terms of PSNR and SSIM, our model performed slightly less than that of the aim19 winner or the best values shown in ntire20. However, our model required the least number of parameters and delivered reasonably high qualitative results compared to the other L1/L2 based methods.

5. Conclusion

Inspired by the recent successful recurrent and back-projection based learning approaches, we proposed a recurrent back projection learning method to solve SISR problem with extreme high scale super resolution. To tackle the high scale problem, we developed a novel module which combines a down-projection block and an upscale module to create an improved LR image. Compared to the state-of-the-art approaches,we secured better performance in several types of datasets, and the output result images were shown clearly sharper than the images of other prominent methods.

6. Acknowledgement

This material is based upon work supported by the Air Force Office of Scientific Research under award number FA2386-19-1-4001.

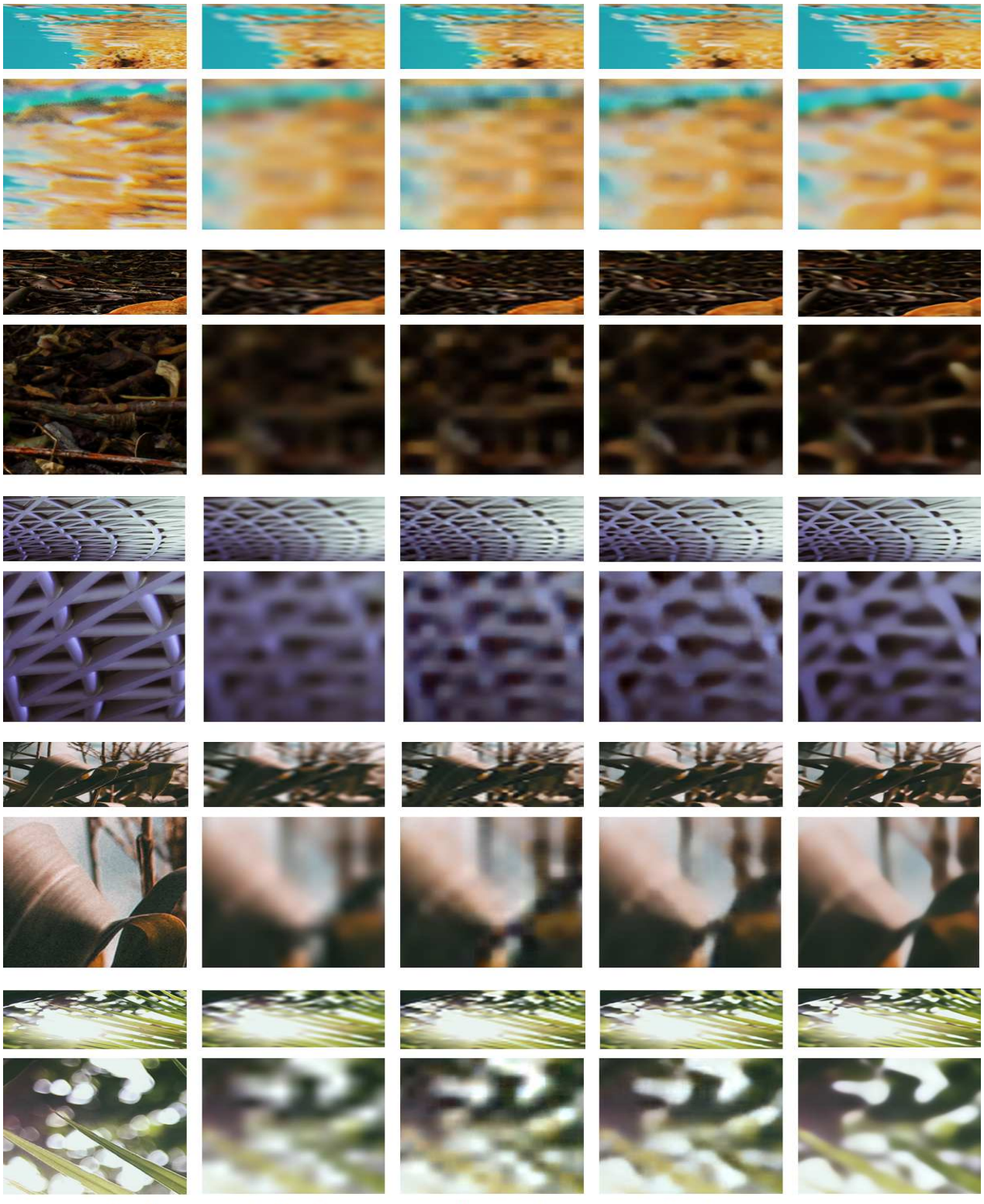

Figure 4. The qualitative Image result on our proposed model and baseline methods on DIV2K validation dataset 0817, 0858, 0892, 0893, 0899.

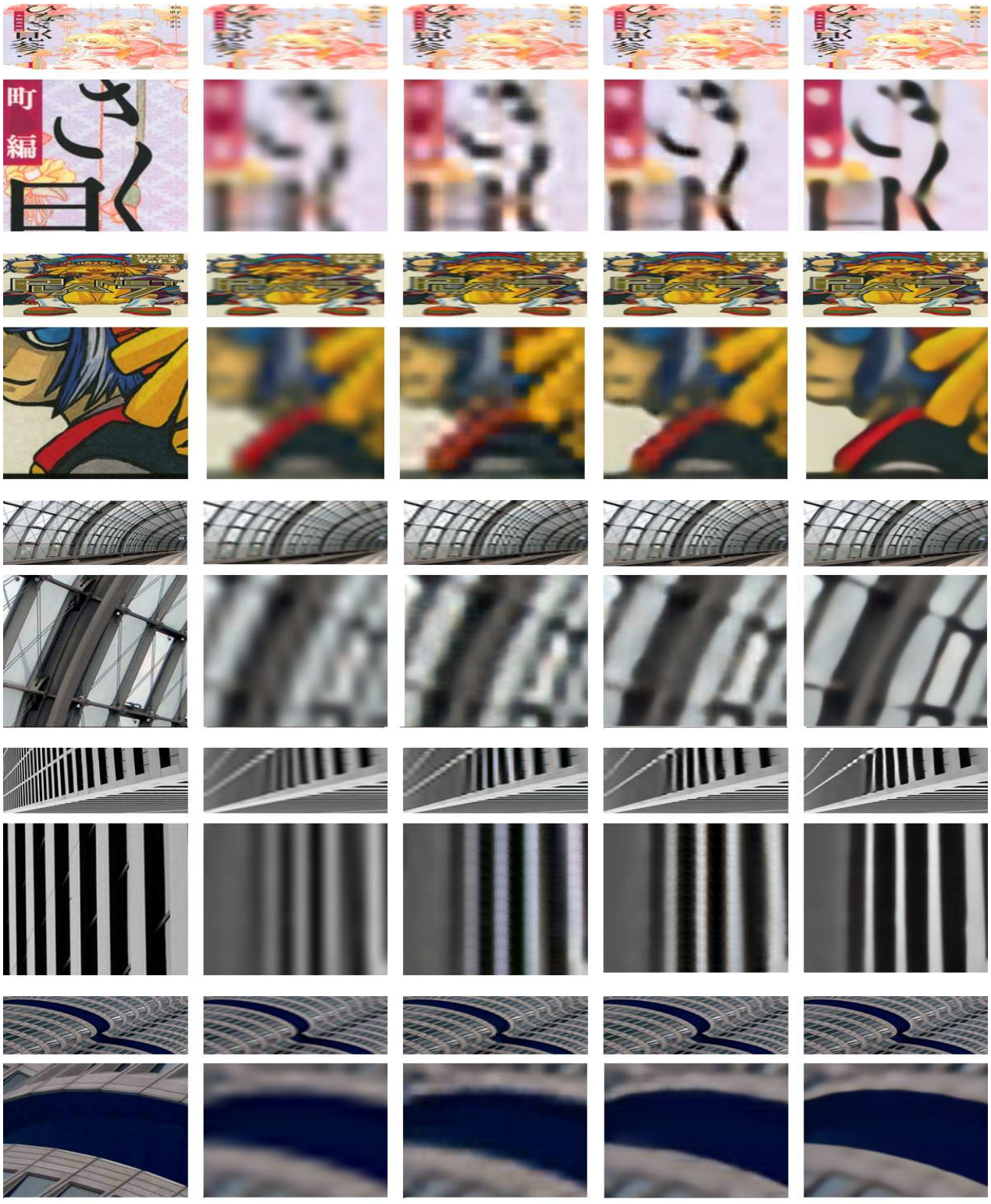

Figure 5. The qualitative Image result on our proposed model and baseline methods on Manga109, Urban dataset

References

[1] Zhen Cui, Hong Chang, Shiguang Shan, Bineng Zhong, and Xilin Chen. Deep network cascade for image super-resolution. In *European Conference on Computer Vision*, pages 49–64. Springer, 2014.

[2] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307, 2015.

[3] Shengkui Gao and Viktor Gruev. Bilinear and bicubic interpolation methods for division of focal plane polarimeters. *Optics express*, 19(27):26161–26173, 2011.

[4] Shuhang Gu, Martin Danelljan, Radu Timofte, Muhammad Haris, Kazutoshi Akita, Greg Shakhnarovic, Norimichi Ukita, Pablo Navarrete Michelini, Wenbin Chen, Hanwen Liu, et al. Aim 2019 challenge on image extreme super-resolution: Methods and results. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 3556–3564. IEEE, 2019.

[5] Shuhang Gu, Andreas Lugmayr, Martin Danelljan, Manuel Fritsche, Julien Lamour, and Radu Timofte. Div8k: Diverse 8k resolution image dataset. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 3512–3516. IEEE, 2019.

[6] Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. Deep back-projection networks for super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1664–1673, 2018.

[7] Jung Woo Hwang and Hwang Soo Lee. Adaptive image interpolation based on local gradient features. *IEEE signal processing letters*, 11(3):359–362, 2004.

[8] Kui Jiang, Zhongyuan Wang, Peng Yi, Guangcheng Wang, Tao Lu, and Junjun Jiang. Edge-enhanced gan for remote sensing image superresolution. *IEEE Transactions on Geoscience and Remote Sensing*, 57(8):5799–5812, 2019.

[9] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1646–1654, 2016.

[10] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Deeply-recursive convolutional network for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1637–1645, 2016.

[11] Qilei Li, Zhen Li, Lu Lu, Gwanggil Jeon, Kai Liu, and Xiaomin Yang. Gated multiple feedback network for image super-resolution. *arXiv preprint arXiv:1907.04253*, 2019.

[12] Zhen Li, Jinglei Yang, Zheng Liu, Xiaomin Yang, Gwanggil Jeon, and Wei Wu. Feedback network for image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3867–3876, 2019.

[13] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 136–144, 2017.

[14] Zhi-Song Liu, Li-Wen Wang, Chu-Tak Li, and Wan-Chi Siu. Hierarchical back projection network for image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[15] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010.

[16] Jaihyun Park, David K Han, and Hanseok Ko. Fusion of heterogeneous adversarial networks for single image dehazing. *IEEE Transactions on Image Processing*, 29:4721–4732, 2020.

[17] Assaf Shocher, Nadav Cohen, and Michal Irani. "zero-shot" super-resolution using deep internal learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3118–3126, 2018.

[18] Casper Kaae Sønderby, Jose Caballero, Lucas Theis, Wenzhe Shi, and Ferenc Huszár. Amortised map inference for image super-resolution. *arXiv preprint arXiv:1610.04490*, 2016.

[19] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018.

[20] Yifan Wang, Lijun Wang, Hongyu Wang, and Peihua Li. End-to-end image super-resolution via deep and shallow convolutional networks. *IEEE Access*, 7:31959–31970, 2019.

[21] Wenming Yang, Xuechen Zhang, Yapeng Tian, Wei Wang, Jing-Hao Xue, and Qingmin Liao. Deep learning for single image super-resolution: A brief review. *IEEE Transactions on Multimedia*, 21(12):3106–3121, 2019.

[22] Kai Zhang, Shuhang Gu, Radu Timofte, et al. Ntire 2020 challenge on perceptual extreme super-resolution: Methods and results. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2020.

[23] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 286–301, 2018.

[24] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2472–2481, 2018.