Gradually Growing Residual and Self-attention Based Dense Deep Back Projection Network for Large Scale Super-Resolution of Image

Manoj Sharma[1,2](✉), Avinash Upadhyay[1](✉), Ajay Pratap Singh[1](✉),
Megh Makwana[3](✉), Swati Bhugra[2], Brejesh Lall[2], Santanu Chaudhury[2],
Deepak[1], and Anil Saini[1]

[1] CSIR-Central Electronics Engineering Research Institute, Pilani, Rajasthan, India
mksnith@gmail.com, avinres@gmail.com, singhajay518@gmail.com
[2] Department of Electrical Engineering, IIT Delhi, Delhi, India
[3] CCS Computer Pvt. Ltd., Delhi, India
meghmak95@gmail.com

Abstract. Due to the strong capacity of deep learning in handling unstructured data, it has been utilized for the task of single image super-resolution (SISR). These algorithms have shown promising results for small scale super-resolution but are not robust to large scale super-resolution. In addition, these algorithms are computationally complex and require high-end computational devices. Developing large-scale super-resolution framework finds its application in smart-phones as these devices have limited computational power. In this context, we present a novel light-weight architecture-Gradually growing Residual and self-Attention based Dense Deep Back Projection Network (GRAD-DBPN) for large scale image super-resolution (SR). The network is made of cascaded self-Attention based Residual Dense Deep Back Projection Network (ARD-DBPN) blocks to perform super-resolution gradually. Where each block performs 2X super-resolution and fine tuned in an end to end manner. The residual architecture facilitates the faster convergence of network and overcomes the issue of vanishing gradient. Experimental results on different benchmark data-set have been presented to compare the efficacy and effectiveness of the architecture.

Keywords: Large Scale Super-Resolution · Gradual · Residual · Dense · Deep Back Projection Network (DBPN) · Self-attention · Spectral normalization

1 Introduction

Single image super-resolution (SISR) is a challenging and an ill-posed task in computer vision since the objective is to recover the high resolution (HR) image

M. Sharma, A. Upadhyay, A. P. Singh and M. Makwana—Equal Contribution.

B. Deka et al. (Eds.): PReMI 2019, LNCS 11941, pp. 260–268, 2019.
https://doi.org/10.1007/978-3-030-34869-4_29

from its low resolution (LR) counterpart. Difficulty in recovering HR image grows gradually with the increase in super-resolution (SR) ratio since the probability space for the solution increases. Introduction of deep learning based frameworks have revolutionized this field with the distinctive improvement in peak signal-to-noise ratio (PSNR) as compared to the conventional approaches.

In 2014, Dong et al. [2] introduced the first deep-learning based solution for SISR which had only three convolution layers and showed significant improvement over empirical methods. Kim et al. [4] proposed a gradient clipping and skip connections based deeper convolutional network to increase the SR performance. In the aforementioned methods, the LR input was first super-resolved using interpolation techniques that resulted in the addition of spurious noise and artifacts to the LR image. Furthermore, it added an additional computational overhead to the process. To overcome this, Ledig et al. [6] proposed a skip connection-based model SRResNet to prevent the gradient diminishing phenomenon in very deep networks. SRResNet takes LR image as input (with no interpolation step) and up-samples the image within the network. They utilized this network as a generator in generative adversarial network SRGAN. Zhang et al. [14] and Tong et al. [12] proposed a residue and skip connection based dense networks respectively to address the gradient diminishing and low converging speed thus facilitating the flow of information through each layer with fast convergence. Recently, Dense DBPN [3] network has shown state-of-the-art performance for SISR. This network consists of consecutive up-sampling and down-sampling blocks attached in series that helps in generating variation rich HR feature and feeding it back in the LR space to enhance the LR features thereby effectively enhancing the super-resolution process. Although these aforementioned state-of-the-art networks performed instant up-sampling to the desired scale in one step their performance diminished with an increase in the scaling factor. This is due to the fact that at higher scales, there is a huge information gap between the LR and HR image. With instant up-sampling in a single step, the extracted features from the LR image lacks enough information for proper reconstruction of the image. Thus, the models fail to converge adequately. To address this problem, Lai et al. [5] and Zhao et al. [15] showed gradual up-scaling approach that perform large-scale super-resolution in multi-levels. However, in gradual up-scaling the artefacts produced by the lower scaling level of the model gets super-resolved in the consecutive higher scaling levels thereby decreasing the overall performance of the network. Manoj et al. [10] presented a residual gradual upscaling network to perform effective SR on larger scales by using residual architectures and end-to-end training of all levels in the model. This facilitated the fast convergence and reuse of the weights from preceding layers. End-to-end optimization of gradual network removed the artefacts created by the up-sampling network at each level.

Motivated by this paper, we present Gradually growing Residual and self-Attention based Dense Deep Back Projection Network (GRAD-DBPN) to super-resolve LR images for higher magnification scales. The network consists of self-attention based Residual-Dense-Deep Back-Projection Network(ARD-DBPN) block which performs 2X SR at every level. We employed self-attention blocks in each ARD-DBPN blocks to extract robust features. ARD-DBPN blocks also have residual connections between the first and last layer of the block. Here, residual architecture facilitated the fast convergence and reuse of the features from preceding layers. These blocks are then repeated to achieve the required magnification scale. The number of blocks required to get the respective scale can be obtained using the formula $log_2(X)$, where X represents the magnification factor. After reaching the desired scale the concatenated network is then fine-tuned in end-to-end manner by passing the error from the last layer of the last ARD-DBPN block to the first layer of the first ARD-DBPN block.

2 Related Work

In recent years, many SISR algorithms based on Convolutional neural network (CNN) have been proposed. Dong et al. [2] introduced the first primitive CNN based approach consisting of only three convolutional layers. It was later upgraded into a very deep CNN architecture by Kim et al. [4]. In the pursuit to improve SISR, many different approaches such as recursive convolutional network [11], deep residual network [7,11], merged shallow and deep CNN [11], sparse convolutional framework [11] and bi-directional recurrent convolutional network [11] have been proposed. The previously mentioned techniques follow two types of approach for the up-sampling step. For example, [2,4] incorporate a conventional method such as interpolation to increase the scale of input image first and then recreate the HR image using the up-scaled image as input. Wheras, in second type of approach the scale is increased by convolutional layers [6]. LapSRN [5], GUN [15] and IRGUN [10] employ gradual up-scaling technique to reconstruct HR image from LR image gradually.

3 Contribution

The major contributions of our work are: 1. A novel Gradually growing Residual and self-Attention based Dense Deep Back Projection Network for large scale SR. 2. Utilization of self-attention based model along with batch normalization and spectral normalization for effective Large scale SR. 3. Experimental study of the proposed network on different benchmark data-sets.

4 Methodology

GRAD-DBPN: The proposed framework GRAD-DBPN shown in Fig. 2 consists of three ARD-DBPN blocks connected back to back as shown in Fig. 1. Each ARD-DBPN block is responsible for 2X upscaling. These blocks consist of 4 stages, feature extraction, self-attention block, deep-back-projection and reconstruction. They are trained for their respective scales. The number of blocks that are required to achieve the specific SR scale X will be $log_2(X)$. These

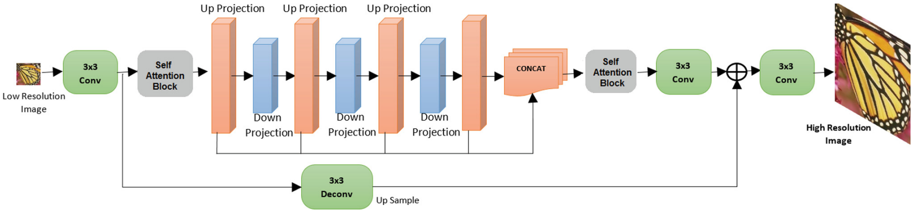

Fig. 1. Block diagram for ARD-DBPN for 2X SR

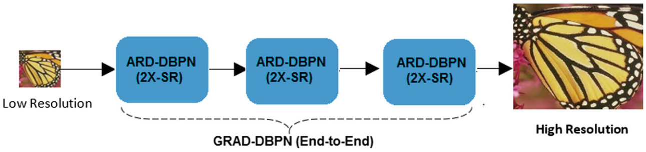

Fig. 2. Block diagram for GRAD-DBPN for 8X SR

ARD-DPBN Blocks are then cascaded to get required magnification factor. The weights from the previous ARD-DBPN block is used in the consecutive blocks instead of random weights except for the first block. Subsequent to training independently, all blocks are cascaded and then fine-tuned in end-to-end manner such that the optimization error from the last block of GRAD-DBPN is utilized for optimizing every ARD-DBPN block simultaneously. ARD-DBPN: An ARD-DBPN block represented in the Fig. 1 consists of following different parts performing respective tasks. 1. Feature Extraction: The first convolution layer L1 with 3X3 kernel size and 64 feature maps extracts the features from the LR image. Next convolution layer L2 with 1X1 kernel size and 32 feature map is used to reduce the dimension of the feature map extracted from the first convolution layer. 2. Self-Attention: The self-attention block [13] is depicted in Fig. 3. These blocks are introduced after the convolutional layer and after the concatenation operation of first and last up-projection network in ARD-DBPN as shown in the Fig. 1. These blocks helps in generating high fidelity natural images using information from all feature location and long range dependencies instead of only depending on spatially local points. 3. Back Projection: The back to back upscaling and downscaling network is used to extract the HR features and projecting it back to LR space to enhance the features. Further, the features of all the upscaling block of the back projection network are concatenated to gather all the enhanced features together. 4. Reconstruction. Another convolution layer L3 with 1X1 kernel size and 64 feature map is used to increase the dimension of the feature map to the dimension of the feature map created by the first layer. This is then passed to another convolution layer L4 of kernel size 3X3 and feature map of 64. Residue from the L1 layer is added here and passed to another convolution layer L6 with kernel size 3X3 and feature size 1 to reconstruct the image. 5. Residue: The features from the L1 layer are upscaled and added to the output of the L4 layer. These combined features are then passed to L5 layer for reconstruction. The passing of feature from the L1 layer to the L5 layer makes the network learn the residue which helps in fast convergence. 6. Spectral Normalization: It is used for stabilizing the training and it facilitates the model to use small computational time by leveraging the power iteration trick resulting in better stability during training [9].

Training: The training of the model is done on RGB image. First, the patches of size 128X128 are extracted from the HR image and termed as 8X-Patches. These patches are used as the ground truth for the end-to-end training of the GRAD-DBPN Model and training the last ARD-DBPN Block. These patches are downscaled to half, quarter and one-eighth of its original size using bi-cubic interpolation and termed as 4X-Patches, 2X-Patches, and 1X-Patches. For 8X SR we used three ARD-DBPN blocks. The first ARD-DBPN block is trained using 1X-Patches as input and 2X-Patches as ground-truth to learn 2X SR. The learned weights from this block are used in the next consecutive block. This block with the learned weights of the previous block is fine-tuned with input as 2X-Patches and ground truth as 4X-Patches. The process is repeated for the third block. After the individual training of all three blocks, they are connected in a cascaded manner and then fine-tuned end-to-end using 1X-Patches as input and 8X-Patches as ground truth. We have used leaky-ReLU as activation function for each convolutional and deconvolutional layer.

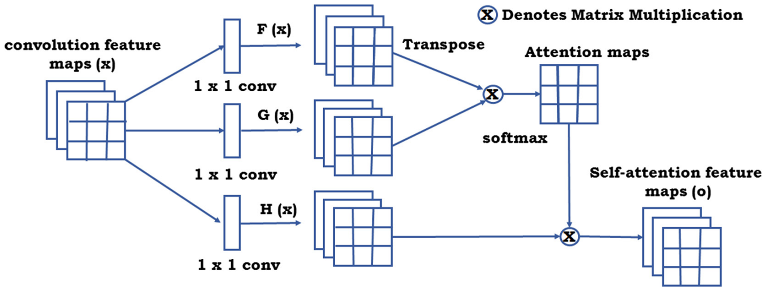
ⓧ Denotes Matrix Multiplication
convolution feature maps (x)
F (x)
1 x 1 conv
G (x)
1 x 1 conv
H (x)
1 x 1 conv
Transpose
Attention maps
softmax
Self-attention feature maps (o)

Fig. 3. Self-attention mechanism [13]

Testing: Testing of the model is done on the complete LR image, instead of patches to avoid the framing effect. Any dimension of LR input can be fed as an input to the model.

Model Specifications: We have used 8 Up-projection and Down-projection block layers. Each convolutional layer of Up-projection and down-projection blocks have 3X3 kernel size and 64 feature maps. LeakyRelu activation function is used. The learning rate of 0.00004 is used for optimization.

Table 1. Comparison of average SSIM and PSNR for various Image SR algorithms at 8X scale for benchmark datasets. Values in red are highest while values in blue are second highest.

Dataset	Scale	Bicubic	-	GUN	-	RDN	-	IRGUN	-	D-DBPN	-	GRAD-DBPN	-
-	-	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Set5	8	24.39	0.657	25.99	0.713	26.10	0.730	26.28	0.740	**26.31**	**0.741**	**26.53**	**0.746**
Set14	8	23.19	0.568	24.23	0.610	24.39	0.614	24.53	0.641	**24.58**	**0.644**	**24.88**	**0.710**
BSD100	8	23.67	0.547	24.42	0.579	24.53	0.584	24.61	0.591	**24.78**	**0.596**	**25.23**	**0.701**
URBAN100	8	20.74	0.515	21.66	0.565	21.70	0.577	21.89	0.594	**21.94**	**0.598**	**22.33**	**0.629**
MANGA109	8	21.47	0.649	23.00	0.717	23.28	0.728	23.43	0.748	**23.46**	**0.749**	**23.85**	**0.792**
DIV2K Validation Dataset	8	23.82	0.549	24.36	0.609	24.47	0.628	24.99	0.652	**24.99**	**0.655**	**25.39**	**0.685**

Table 2. Comparison of average PSNR for various Image SR algorithms at 4X scale for benchmark datasets. Values in red are highest while values in blue are second highest.

Dataset	Scale	Bicubic	VDSR	GUN	RDN	EDSR	IRGUN	D-DBPN	GRAD-DBPN
Set5	4	28.42	31.35	31.50	31.58	32.62	32.65	**32.68**	**33.24**
Set14	4	26.10	28.03	28.04	28.29	28.94	28.98	**29.09**	**29.42**
BSD100	4	25.96	27.29	27.44	27.55	27.79	28.01	**28.24**	**28.74**
URBAN100	4	23.15	25.18	25.24	25.44	26.86	25.48	**25.67**	**26.32**
MANGA109	4	24.92	28.82	28.97	29.09	29.12	29.22	**29.23**	**29.62**
DIV2K Validation Dataset	4	27.32	28.22	28.67	28.92	28.99	29.1	**29.14**	**29.56**

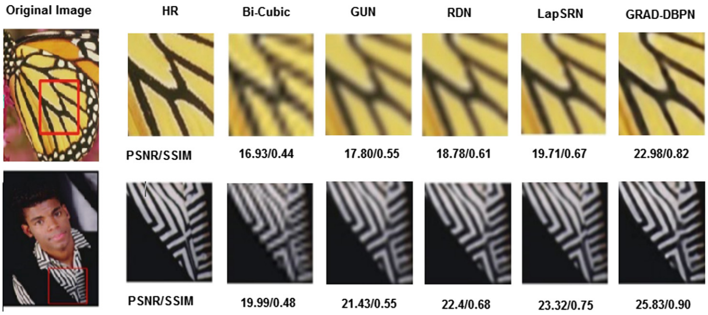

Fig. 4. Comparison at 8X SR on Set5 and BSD100 Dataset

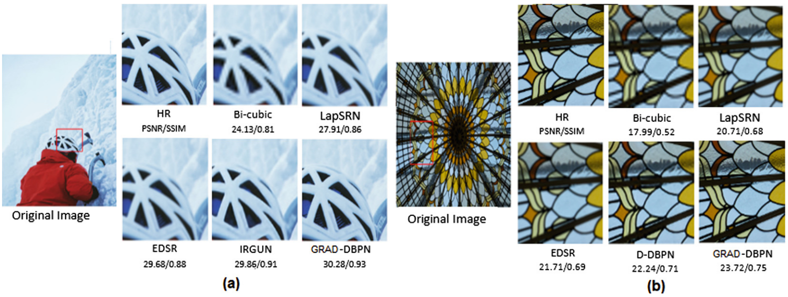

Fig. 5. Comparison at 8X SR on DIV2K Validation Dataset

5 Experiments

We have carried out numerous experiments to show the efficacy of proposed framework. We have achieved better performance from existing modern algorithms with less time complexity. We used popular datasets for all our experiments. Training is done using 50,000 images from the ImageNet dataset. To compare our findings with the existing state-of-the-art SR methods we have used **URBAN100**[10]. **BSD100** [8], **Set5** [2], **Set14** [2], **DIV2K** [1] and **Manga109** [10] datasets. We have used Intel Core i7 processor having clock speed 3.6 GHz and RAM of 128 GB with Nvidia GTX 1080 GPU for all our experiments.

5.1 Comparisons with Other State-Of-The-Art Methods

We have shown the comparison of the performance for proposed framework with other modern algorithms on higher scales(8X) in Table 1. Algorithms such as GUN [15], LapSRN [5], IRGUN [10] and D-DBPN [3] have shown state-of-the-art performance for 8X SR. We have gathered the source codes for these algorithms available publicly and trained them for 8X SR alongside our network. IRGUN and D-DBPN have performed reasonably well for 8X SR in the past. In Table 2 we have shown a comparison of our results with the algorithms which showed the state-of-the-art result on lower resolution scale (4X). These algorithms are VDSR [4], EDSR [7], RDN [11], IRGUN [10], GUN [15] and D-DBPN [3]. D-DBPN currently is the state-of-the-art algorithm. To show the result of our model on this scale we have used only two blocks of ARD-DBPN to get 4X SR. Since our model is deliberately trained and designed for high scale SR, we are not showing any comparison for low-scale SR (2X).

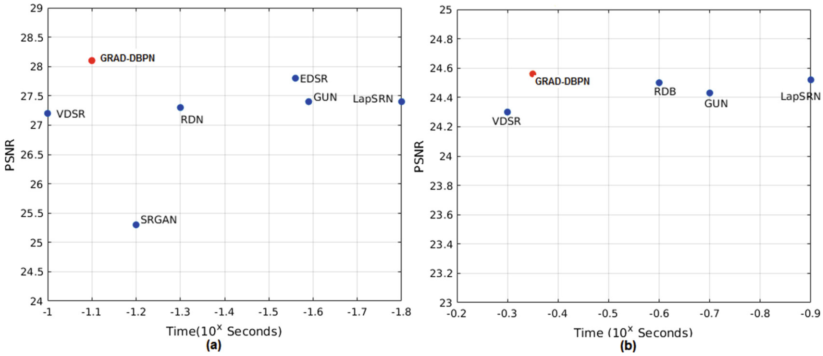

Fig. 6. Run-time performance comparison of various frameworks for (a) 4X SR on BSD100 dataset (b) 8X SR on BSD100 dataset

5.2 Result Analysis

As shown in Table 2, the proposed GRAD-DBPN network outperforms all other architectures in terms of PSNR and SSIM. Our framework have shown an average improvement of 0.36 dB in PSNR and 0.047 in SSIM over the current state-of-the-art framework D-DBPN for 8X. It has also shown a moderate improvement of 0.47 dB in PSNR over DBPN for 4X scale. In the Fig. 6 we have shown the average testing time on 4X and 8X scales for BSD100 Dataset. It is evident that our model gave good trade off between PSNR and time taken for testing in comparision of other frameworks. Visual results of our model is also depicted in Figs. 5 and 4 along with the PSNR and SSIM values for DIV2K and BSD100 datasets. The proposed framework is faster than the present state-of-the-art models without making any compromise over PSNR performance. This makes it suitable for smartphone applications. The model is light-weight and take less than 2 MB of storage space.

6 Conclusion

In this work, we presented a novel Gradually growing Residual and self-Attention based Dense Deep Back Projection Network (GRAD-DBPN) that showed significant improvement in terms of PSNR and SSIM metrics for single image super-resolution (SISR) as compared to the existing algorithms for large magnification ratio. Usage of spectral norm facilitated quicker and improved convergence of the error. Self attention and gradual growing improves the perceptual and objective quality while using less computational resources thus making it a light-weight architecture.

References

1. Agustsson, E., Timofte, R.: 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 (2017)
2. Dong, C., Loy, C.C., He, K., Tang, X.: Image super-resolution using deep convolutional networks. IEEE Trans. Pattern Anal. Mach. Intell. **38**(2), 295–307 (2015)
3. Haris, M., Shakhnarovich, G., Ukita, N.: Deep back-projection networks for super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1664–1673 (2018)
4. Kim, J., Kwon Lee, J., Mu Lee, K.: Accurate image super-resolution using very deep convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1646–1654 (2016)
5. Lai, W.S., Huang, J.B., Ahuja, N., Yang, M.H.: Deep Laplacian pyramid networks for fast and accurate super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 624–632 (2017)
6. Ledig, C., 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, pp. 4681–4690 (2017)

7. Lim, B., Son, S., Kim, H., Nah, S., Mu Lee, K.: 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 (2017)
8. Martin, D., Fowlkes, C., Tal, D., Malik, J., et al.: A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. ICCV Vancouver (2001)
9. Miyato, T., Kataoka, T., Koyama, M., Yoshida, Y.: Spectral normalization for generative adversarial networks. arXiv preprint arXiv:1802.05957 (2018)
10. Sharma, M., Mukhopadhyay, R., Upadhyay, A., Koundinya, S., Shukla, A., Chaudhury, S.: Irgun: improved residue based gradual up-scaling network for single image super resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 834–843 (2018)
11. Timofte, R., Agustsson, E., Van Gool, L., Yang, M.H., Zhang, L.: Ntire 2017 challenge on single image super-resolution: methods and results. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 114–125 (2017)
12. Tong, T., Li, G., Liu, X., Gao, Q.: Image super-resolution using dense skip connections. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4799–4807 (2017)
13. Zhang, H., Goodfellow, I., Metaxas, D., Odena, A.: Self-attention generative adversarial networks. arXiv preprint arXiv:1805.08318 (2018)
14. Zhang, Y., Tian, Y., Kong, Y., Zhong, B., Fu, Y.: Residual dense network for image super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2472–2481 (2018)
15. Zhao, Y., Li, G., Xie, W., Jia, W., Min, H., Liu, X.: GUN: gradual upsampling network for single image super-resolution. IEEE Access **6**, 39363–39374 (2018)