

Ultra Sharp : Single Image Super Resolution using Residual Dense Network

Karthick Gunasekaran
University of Massachusetts,Amherst
CICS, Amherst, MA 01002
kgunasekaran@umass.edu

Abstract

For years Single Image Super resolution(SISR) is an interesting and ill posed problem in Computer Vision. The traditional Super Resolution(SR) imaging approaches involve Interpolation, Reconstruction and Learning based methods. Interpolation methods are fast and uncomplicated to compute but they are not so accurate and reliable. Reconstruction based methods are better compared with Interpolation methods but are time consuming and quality degrades as the scaling increases. Even though, Learning based methods like Markov random chain are far better than all the previous they are unable to match the performance of deep learning models for SISR. In this project, Residual Dense Networks will be used to tackle the problem of SISR. This architecture makes full use of hierachial features from original low-resolution (LR) images to achieve higher performance. The network structure will consists of four main blocks. The core of the architecture is the residual dense block(RDB) where the local features are extracted and made use of via dense convolutional layers. In this work, investigation of each block was performed and effect of each modules was be studied and analyzed. Analyses by use various loss metric was also carried out in this project. Also a comparison was made with various state of the art models. The modules in the model were be built from scratch and were trained and tested. The training and testing was be carried out for various scaling factors and the performance was be evaluated.

1. Introduction

Single Image Super Resolution(SISR) is the process in which High resolution(HR) images are recovered from Low resolution(LR) images. It is one of the important problem in Computer vision and has a wide range of applications from security and surveillance purpose to medical imaging scenarios. The problem can be dealt with in a variety of ways from classical computer vision techniques to using deep learning to solve it. The classical super resolution

techniques include interpolation based methods [8] such as nearest neighbours, bilinear, bicubic interpolation and reconstruction based methods. However those methods are unable to perform as good as the learning based methods. Given the increasing amount of data being available learning based methods are able to achieve better results and the prediction time is comparatively less. Various deep learning based methods have been proposed for solving the Image Super resolution problem.

The first basic Convolutional Neural Network based method was proposed by dong et al. [2] in 2014 which performed much better than traditional methods. From then various different approaches and architecture were proposed differing in terms of network architecture, loss functions and learning principles. kim et al.[9] in the following year proposed an very deep convolution network inspired from VGG-net. By this way they were able capture the contextual information over large regions of the images. However, the models proposed should have different architecture for different scales .Inspired by ResNet architecture, lim et al. [11] proposed a residual learning based approach where in high resolution images are produced from different scaling factors by using a single model. Tai et al. [13] proposed MemNet which tackles the long term dependency problem in deep models. MemNet uses memory blocks to explicitly mine persistent memory through an adaptive learning process. Memory blocks help in adaptive controlling how much should be remembered from previous states. Tong et al. [14] proposed a dense skip connections based approach where in feature maps of each layer is added to the feature maps in every other layer by effectively combining low-level features with the high level features and eliminating vanishing gradient problem. Deconvolution layers are added to the network to learn the upsampling filters and to speedup reconstruction. However there are problems with these approaches. These methods fails to fully make use of the hierarchical features for reconstruction. Also the methods don't extract multi level features from the images and in some methods upsampling is performed as pre-processing from Low resolution which greatly misses many of the in-

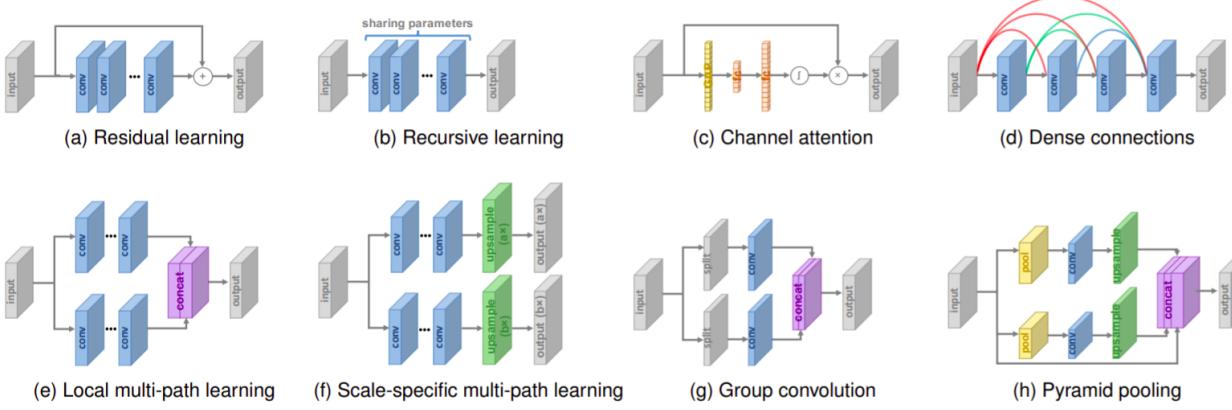


Figure 1: Various Design strategies [16]

formation and increase the time complexity as well. To overcome these problems a new residual dense network is proposed. This network fully makes use of hierarchical features with residual dense block (RDB) forming the core of the architecture. The network structure will consists of four main blocks shallow feature extraction network, residual dense block, dense feature fusion block and up sampling block. Shallow feature extraction module will extract shallow features by using Convolutional layers. Residual block has dense connected layers, local feature fusion(LFF) and uses continuous memory (CM) mechanism. Dense feature fusion block uses Global feature fusion and Global residual learning. The Global feature fusion module adaptively fuses hierarchical features from all RDBs in Low resolution space. The layers in RDB consists of densely connected layers and local feature fusion. Contiguous memory mechanism is also implemented in the RDB.

The main contributions of this work include

- A new architecture which makes use of hierarchical features called Residual Dense Network is presented for single image super resolution.
- Residual Dense block is proposed which consists of local dense connections and local feature fusion and uses contiguous memory mechanism.
- At last all the features which were obtained from RDB and inputs are fused adaptively in low resolution space and combined with features extracted initially to perform residual learning.

2. Background/Related Work

Super resolution is an ill-posed problem, there are many different problems such as designing up-sampling block, deciding the architecture to be used etc. In this section we will analyse the most prominent up-sampling strategies and

architectures used by researchers over time for solving this problem.

2.1. Upsampling Architectural Techniques

In Super Resolution(SR), four different upsampling strategies are used in general to scale the low resolution image to map the high resolution image. The four model frameworks include pre-upsampling, post-upsampling, progressive upsampling and iterative up-down upsampling.

2.1.1 Pre-upsampling Architecture

In this architecture the upscaling is initially carried out using classical upsampling techniques such as bicubic interpolation. This helped the model to reduce complexity since the CNN can just refine the up-sampled image to obtain HR image. The complex upsampling task was carried out by the classical techniques. This setup was initially adopted by [2] for SR-CNN and was widely used earlier during initial development of SR architectures using the CNNs. Figure 2(a) shows pre-upsampling architecture. However this approach has been replaced by learning based approaches in recent times. The disadvantage of this is that there is no learning taking place when the LR images are up-scaled to HR images.

2.1.2 Post-upsampling Architecture

In post-upsampling architecture, multiple learnable layers are added at the end of the SR architecture where upsampling is performed. This kind of a framework was initially proposed by Dong et al. [3] where he performs most of mappings in LR space. Figure 2(b) shows an sample post upsampling architecture. Since most of the features are extracted in LR space the complexity is low and also the computational cost is low. This increase the performance of the network and widely used in current SR research.

2.1.3 Progressive upsampling Architecture

Progressive upsampling was suggested to overcome the disadvantages of Pre and Post upsampling. In pre-upsampling the upscaling is done in a single step which inhibits learning while in post upscaling a new network is needed which cannot perform multiscale super resolution. Figure 2(c) shows an example of progressive upsampling architecture. Lai et al proposed Laplacian pyramid network [10] to overcome these difficulty. This model progressively reconstructs high resolution images. This model however has its own problems such as it involves complicated architectures and design. They also required advanced training strategies and guidance.

2.1.4 Iterative upsampling Architecture

In iterative upsampling architecture, the mutual dependency of LR and HR image is made use. This considered as a reconstruction problem where the LR image is iteratively converted in HR image and compared with the HR image to compute the loss. However, this approach bring in complication to the design since different training and testing framework have to be present. This is the most recently proposed architecture which under study in recent years. Figure 2(d) shows an sample iterative upsampling architecture.

2.2. Design

In SR architectures various design strategies are being carried to improve the efficiency. Commonly used network designs are explored and discussed in this section.

2.2.1 Residual Learning:

Inspired from ResNet [5] architecture residual learning is applied at different levels. Figure 1(a) shows an sample residual design. There are two ways mainly employed for residual learning which are Global Residual learning and Local Residual learning. These are employed by making direct connection from input to the output block.

2.2.2 Recursive Learning:

In recursive learning same module is applied multiple times in a recursive manner which helps in achieving wider and larger receptive field. This also helps in extracting higher level features as well. However there are problems like vanishing and exploding gradient occuring in this. A Sample of which is shown in 1(b).

2.2.3 Multi-Path Learning:

As shown in 1(e) and (f) multi-path learning involves involves learning features in multiple paths where in each

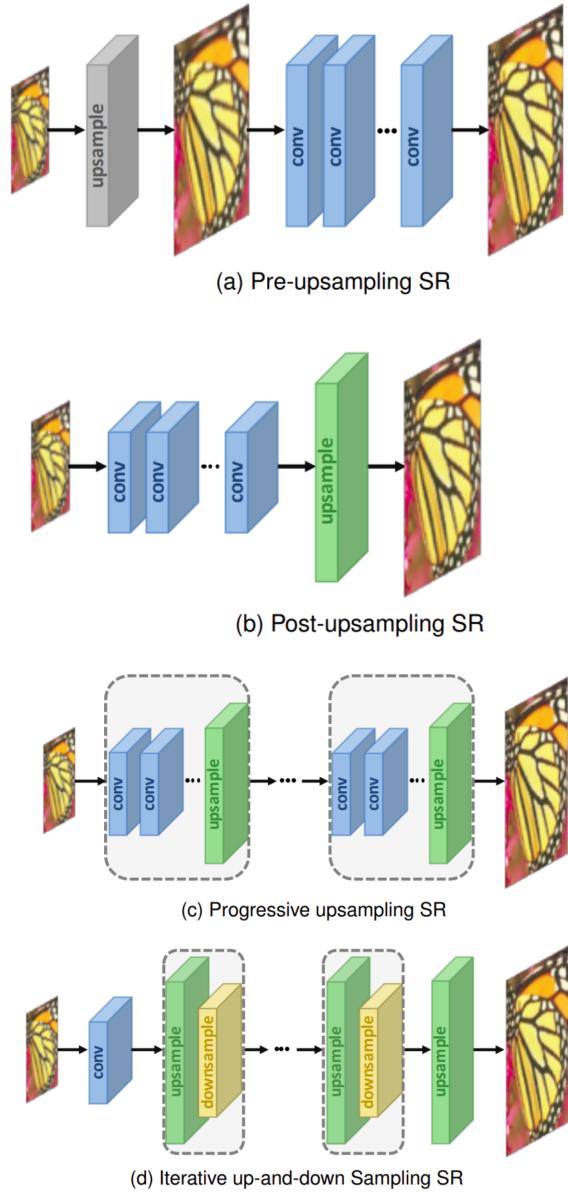


Figure 2: Different Up-sampling architectures [16]

path different operations are performed. For SR purpose they are divided into Global, Local and Scale-specific multi-path learning. In scale specific learning there are multiple different paths for different scale of images.

2.2.4 Dense layer connection:

The dense layer connections were inspired from the DenseNet proposed by Huang et al. [7]. In each layer the connections are made with all the other previous layers by having the feature maps of the preceding layers as input

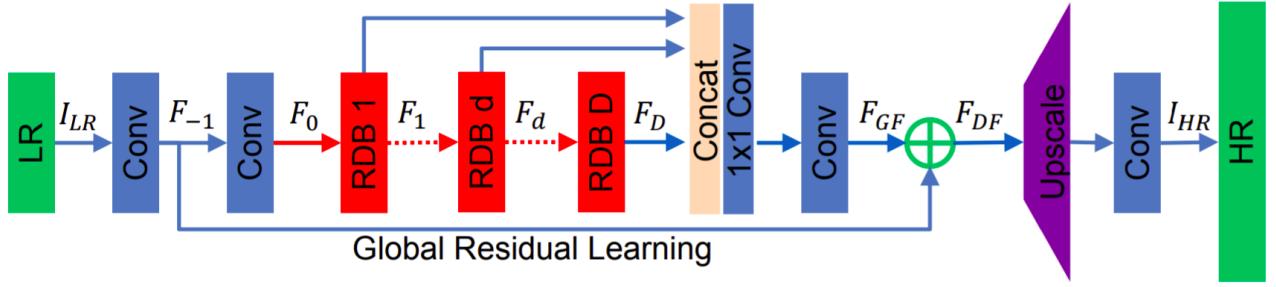


Figure 3: Architecture of Residual dense network [17]

here. This connection helps in better signal propagation and feature reuse. Figure 1(d) shows dense connections.

2.2.5 Channel Attention:

In channel attention mechanism the inputs are fused into channel descriptor where global average pooling is performed. Then the descriptors are directly used by two fully connected layers to produce scaling factors for each channel. Channel attention mechanism is shown in Figure 1(c).

3. Approach

3.1. Network Architectures:

The main residual dense network architecture consists of four main components as shown in the Figure 3. They main blocks are

- Shallow Feature Extraction block
- Residual Dense Block
- Dense Feature Fusion block
- Upsampling block

3.1.1 Shallow Feature Extraction(SFE) block

Shallow feature extraction block consists of convolutional layers where in the features are extracted in low resolutions(LR) space. The features extracted are fed into each of the residual dense block layers and also finally used in residual learning for global feature fusion. This layer is able to achieve a good high level representation of the images. It consists of two convolutional layers where the output of first layer is fed into the second convolutional layer.

Let Img_{LR} be the input low resolution image and $Conv$ be the convolutional operation in feature extraction layer.

$$F_{-1} = Conv_1(Img_{LR}) \quad (1)$$

where F_{-1} are the features extracted from input low resolution image

$$F_0 = Conv_2(F_{-1}) \quad (2)$$

Here, F_0 be the features extracted from this block which will be fed into residual dense block.

3.1.2 Residual Dense Block

The Residual Dense Block(RDB) performs the Local residual learning and Local feature fusion which helps in extracting the local dense features and providing better context locally.

Multiple residual dense blocks are implemented in a feed forward manner taking the output of previous block as the input to current block. Each block consists of three feed forward convolutional layers followed by non linear layer and finally all the layers within a block are fused and residual input is applied. Figure 4 clearly shows the architecture of Residual Dense Block(RDB).Each residual block makes use of the output from the previous block as shown in the below equation.

Let F_{d-1} be the features from the previous layer.

$$F_d = Conv_{RDB,1}(F_{d-1}) \quad (3)$$

Further residual dense blocks are processed in a feed forward manner as follows using the output from previous blocks,

$$F_d = Conv_{RDB,d}(F_{d-1}) \quad (4)$$

$$F_d = Conv_{RDB,d}(Conv_{RDB,d-1}(\dots)) \quad (5)$$

Within each residual dense block, local feature fusion is performed where all the outputs of the previous layers are fused or concatenated adaptively together as shown in the figure 4 to be fed as input to next layer within RDB. This means the output of all preceding layers of current RDB and output of previous RDB block are given as input to current layer within RDB. Continuous memory mechanism is

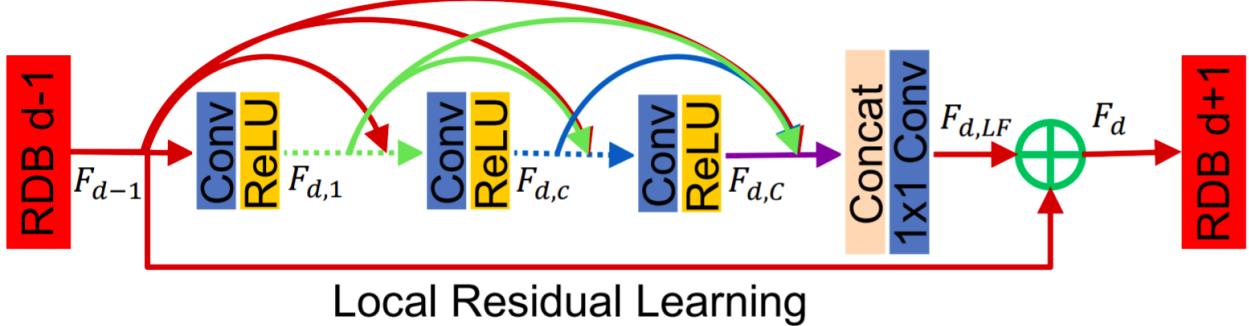


Figure 4: Architecture of Residual dense block [17]

achieved since all the preceding RDB output is feed into each layer with current RDB. This helps preserve the feed forward tendency and extracts dense features. The output and input for each layer in the RDB is of the equation,

$$F_{d,l} = \text{ReLU}(\text{Conv}([F_{d-1}, F_{d,1}, \dots, F_{d,l-1}])) \quad (6)$$

where $F_{d,l}$ and $F_{d-1}, \dots, F_{d,l-1}$ are the outputs and inputs of the current residual dense block and $\text{ReLU}[4]$ is the non-linearity after the convolutional layer.

Finally the concatenation of the outputs of RDB layers is carried out and $1*1$ Convolutions are performed such that the feature maps size is reduced.

$$F_d = \text{Conv}_{1*1}([F_{d-1}, F_{d,1}, \dots, F_{d,C}]) \quad (7)$$

After this the local residual learning is performed in which the the input to the RDB block is added to the output of the current RDB block. This helps in maintaining the information flow through out the network.

$$F_d = F_d + F_{d-1} \quad (8)$$

Local residual Learning helps the network learn and remember local structure better resulting in better performance.

3.1.3 Dense Feature Fusion(DFF) block

In DFF, global feature fusion and global residual learning is performed. This helps in achieving the hierarchical features from the LR images in HR space. In this block all the outputs from RDB layers are concatenated and residual learning is performed from LR image input.

Global feature fusion is achieved by fusing or concatenating all the outputs from the RDBs and performing $1*1$ convolution on top of it. A $3*3$ convolutional filter is again applied on the output from $1*1$ filter.

$$F_{GF} = \text{Conv}_{3*3}(\text{Conv}_{1*1}((F_1, \dots, F_D))) \quad (9)$$

Global residual learning is performed in which the output from GFF is added up with the high level features of the LR input image. The output of which is fed into the upscaling layer.

$$F_{DFF} = F_{-1} + F_{GFF} \quad (10)$$

3.1.4 Upscaling block

Upscaling to higher resolution space is carried out by using sub-pixel based convolutional neural network. The feature maps from the DFF which were extracted in low resolution space is upscaled into HR space using an efficient sub pixel based approach as proposed by Shi et al. [12].

4. Implementation

The code for the implementation can be found at the following github link online at <https://github.com/karthickpgunasekaran/SuperResolutionWithRDN>

4.1. Dataset:

Four different datasets were used for the Overall training and testing the network. For training DIV2K dataset was used. The DIV2K dataset consists of 800 training and 100 validation 2K resolution images. Degradation was carried out by using Bicubic downsampling on the HR dataset to be used as inputs. The testing was carried out using SET-5, SET-14 and URBAN-100 datasets. LR Datasets were created for different scales such as 2x,3x and 4x.

Figure 5 shows different scaled versions of the dataset. In the first row is whole original image of the dataset is displayed. To present the difference in resolution of the images in the next rows zoomed versions of images are presented. Second row is the original image which is just zoomed in. Third row contains images which are 2x downsampled by bicubic downsampling. Fourth row and Fifth row shows im-

ages which are 3x and 4x downsampled by bicubic down-sampling.

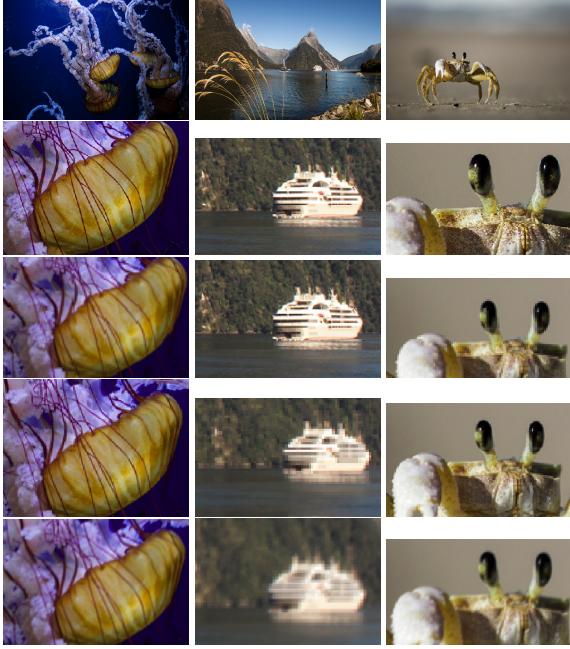


Figure 5: Various image scales Row 1 - Original image, Row 2 - Original zoomed, Row 3 - 2x bicubic downscaling, Row 4 - 3x bicubic downscaling, Row 5 - 4x bicubic downscaling

4.2. Loss Metrics:

The network uses different loss metrics training and evaluation. The training was carried out using L1 loss metric. While Structure Similarity Index Matrix (SSIM) [15] and Peak Signal to Noise Ratio(PSNR) were used as the loss functions for the evaluation. SSIM gives a good measure of how similar two images are in terms of brightness, contrast and structure while PSNR focuses on noise on the image.

$$SSIM(Img_{pred}, Img_{gt}) = \frac{(2\mu_{Img_{pred}}\mu_{Img_{gt}} + C_1)(2\sigma_{Img_{pred}Img_{gt}} + C_2)}{(\mu_{Img_{pred}}^2 + \mu_{Img_{gt}}^2 + C_1)(\sigma_{Img_{pred}}^2 + \sigma_{Img_{gt}}^2 + C_2)}$$

$$PSNR(Img_{pred}, Img_{gt}) = \frac{R^2}{\frac{1}{mn} \sum_{i=0,j=0}^{i=m,j=n} [Img_{pred}(i,j) - Img_{gt}(i,j)]^2}$$

$$MAE(Img_{pred}, Img_{gt}) = \frac{1}{mn} \sum_{i=0,j=0}^{i=m,j=n} \|Img_{pred}(i,j) - Img_{gt}(i,j)\|$$

where Img_{pred} and Img_{gt} are the predicted and ground truth HR images and R is the maximum fluctuation in the input image data.

4.3. Setting:

The batch size of 8 HR and LR images where used for training the model and batch size of 4 was used for testing the model. The weights were initialized by kaiming he method [6]. Adam optimizer [1] was used to perform gradient updates. The regularization was set to 1e-8. The learning rate was initially set to 1e-4 and was reduced every 15 epochs by 2. The model was trained on Google lab for 200 iterations. The model took 14 1/2 hours to train. Since Colab supports only approximately 12 hours of continuous usage model was saved in between and continued.

4.4. Results:

Table 1: Comparison of model performance among different scales

Dataset	2x		3x		4x	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Set-5	30.67	0.91	28.05	0.88	26.19	0.85
Set-14	28.56	0.87	27.78	0.83	24.21	0.79
Urban-100	28.93	0.88	26.3	0.83	24.02	0.77

The model was trained and tested with different degradation scales such as 2x,3x and 4x. A comparison was also made with various other state-of-the-art models across different datasets. The efficiency of few components in the architecture was also studied to understand their impact their impact on the overall architecture.

Table 1 shows the quantitative performance of the proposed model for different scales of LR images. It can be clearly seen that the performance of the model is better on 2x down sampled images compared to others across all the datasets. This is due to the fact that the highly degraded images are less accurately converted to high resolution images. The particular loss in details are bound to happen. The SSIM metric is around 0.90 while the PSNR is around 30 for 2x scaling while the SSIM value was 0.85 for 4x scaling. The performance seemed to be varying for different datasets but similar pattern was seen.

The performance of the proposed model with various state of the art models is shown in Table 2. The comparison was made with SRCNN [2], MemNet [13] and Laplacian SR [10]. SR CNN uses the pre-upsampling layer and uses sparse encoding mechanism. MemNet uses the recursive learning mechanism mentioned in literature survey along with persistent memory mechanism. Laplacian SR uses pyramid mechanism where at each level takes feature maps and predicts high-frequency residual maps. Though

Table 2: Comparison of model performance with different models

Dataset	Proposed		MemNet SR		Laplacian SR		SRCNN	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Set-5	26.19	0.85	25.85	0.83	25.27	0.83	23.74	0.81
Set-14	24.21	0.79	24.1	0.77	24.15	0.77	23.04	0.76
Urban-100	24.02	0.77	23.57	0.78	23.76	0.78	21.96	0.75

Table 3: Model performance after removing functionalities

Dataset	Baseline		Global residual learning removed		Local Dense Connection removed		Local residual learning removed	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Set 5	26.19	0.85	21.55	0.72	13.45	0.45	18.31	0.54
Set 14	24.21	0.79	21.36	0.68	12.84	0.38	18.59	0.51
Urban-100	22.02	0.75	20.78	0.69	13.12	0.41	18.72	0.56

other state-of-the-art networks exists the performance was compared only with selected few since they had more distinct architecture than the proposed one and also due to the limitation of computation resources and time. The comparison was only performed with 4x scale degradation across all the models. The RDN which was used consists of only lesser no of RDB block due to the fact that the training time limitation. However it could be seen that the proposed model outperforms all the other state of the art models. From this its understandable that the dense residual learning based architecture performs better than others.

Table 3 shows the performance of the model with some features removed from the architecture. A comparison was made with Baseline by removing the global residual learning functionality, by removing the dense feature connection and also by removing local residual learning individually. Each of these modules/functionality were removed individually and trained and tested. From the results presented in the table it could be inferred that the Local dense connection was one of the important functionality since removing it the model was not able to convert LR images to HR efficiently which led to very poor SSIM and PSNR value. Local residual learning is also seems to be an important functionality since removing it led to an PSNR/SSIM values of around 18/0.54 while the baseline model was much better around 26/0.85. This clearly shows the performance of the individual modules. Then Global residual learning didn't lead to performance degradation as local dense connection and local residual learning. The Global residual learning has a SSIM/PSNR value of 0.72/21.55 which is comparatively little less than the baseline but not as important as other two modules.

5. Conclusion:

This paper proposes Residual Dense Network architecture for Single Image Super Resolution where novelty

comes from Residual Dense Blocks. Contiguous memory mechanism is achieved in RDB where each block is connected to all the previous blocks in the module. Local Dense feature fusion within each RDB helps with context and preserve local information. Local residual learning helps with the flow of context and feature knowledge within the local regions. All features are extracted in the LR space leading to better performance with lesser computational resources. The proposed model uses both the local features and global features to effectively perform super resolution. Comparison was made with various state of the art models. Also the model was trained and tested for different degradation scales. Performance and importance of each module in the architecture was also evaluated.

References

- [1] J. B. Diederik P. Kingma. Adam: A method for stochastic optimization. *Machine Learning*, 29(6):1153–1160, December 2014.
- [2] C. Dong, C. C. Loy, K. He, and X. Tang. Learning a deep convolutional network for image super-resolution. In D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 184–199, Cham, 2014. Springer International Publishing.
- [3] C. Dong, C. C. Loy, and X. Tang. Accelerating the super-resolution convolutional neural network, 2016.
- [4] X. Glorot, A. Bordes, and Y. Bengio. Deep sparse rectifier neural networks. In G. Gordon, D. Dunson, and M. Dudík, editors, *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 315–323, Fort Lauderdale, FL, USA, 11–13 Apr 2011. PMLR.
- [5] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.
- [6] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *CoRR*, abs/1502.01852, 2015.

- [7] G. Huang, Z. Liu, and K. Q. Weinberger. Densely connected convolutional networks. *CoRR*, abs/1608.06993, 2016.
- [8] R. Keys. Cubic convolution interpolation for digital image processing. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 29(6):1153–1160, December 1981.
- [9] J. Kim, J. K. Lee, and K. M. Lee. Accurate image super-resolution using very deep convolutional networks, 2015.
- [10] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang. Deep laplacian pyramid networks for fast and accurate super-resolution, 2017.
- [11] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee. Enhanced deep residual networks for single image super-resolution, 2017.
- [12] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. *CoRR*, abs/1609.05158, 2016.
- [13] Y. Tai, J. Yang, X. Liu, and C. Xu. Memnet: A persistent memory network for image restoration, 2017.
- [14] T. Tong, G. Li, X. Liu, and Q. Gao. Image super-resolution using dense skip connections. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 4809–4817, Oct 2017.
- [15] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *Trans. Img. Proc.*, 13(4):600–612, Apr. 2004.
- [16] Z. Wang, J. Chen, and S. C. H. Hoi. Deep learning for image super-resolution: A survey. *CoRR*, abs/1902.06068, 2019.
- [17] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu. Residual dense network for image super-resolution. *CoRR*, abs/1802.08797, 2018.