# Single Image Super Resolution Using Residual Learning and Channel Attention Network

INSPIRATION: FEILONG CAO AND HUAN LIU, "SINGLE IMAGE SUPER-RESOLUTION VIA MULTI-SCALE RESIDUAL CHANNEL ATTENTION NETWORK", PII:S0925-2312(19)30772-6; REFERENCE: NEUCOM 20849

GROUP MEMBERS:

| Name | Roll no. |
|---|---|
| Hritam Basak | 001710801018 |
| Alik Roy | 001710801038 |
| Jeet Bandhu Lahiri | 001710801040 |
| Sayantan Bose | 001710801078 |
| Soumyadeep Patra | 001710801082 |

# Single Image Super Resolution Using Residual Learning and Channel Attention Network

Influenced by: *Single Image Super-resolution via Multi-scale Residual Channel Attention Network* [1]


Hritam Basak
*Roll: 001710801018*
*BEE-IV*
*Department of Electrical Engineering*
Jadavpur University

Alik Roy
*Roll : 001710801038*
*BEE-IV*
*Department of Electrical Engineering*
Jadavpur University

Jeet Bandhu Lahiri
*Roll : 001710801040*
*BEE-IV*
*Department of Electrical Engineering*
Jadavpur University

Sayantan Bose
*Roll : 001710801078*
*BEE-IV*
*Department of Electrical Engineering*
Jadavpur University

Soumyadeep Patra
*Roll : 001710801082*
*BEE-IV*
*Department of Electrical Engineering*
Jadavpur University


## I. Introduction

The aim of our project is to address the problems of reconstructing an accurate high-resolution (HR) image given its low-resolution (LR) counterpart, commonly called as super-resolution (SR) in the terms of computer vision. A high-resolution image has a high pixel density, which means it contains more information about the original scene. For better performance in pattern recognition and image analysis, high resolution is typical in computer vision applications. In medical imaging, high resolution is critical for diagnosis. Many applications, such as surveillance, forensics, and satellite imaging, necessitate zooming of a particular area of interest in the image, making high resolution essential.

High-resolution pictures, on the other hand, are not always available. This is because the setup for high resolution imaging is costly, and it might not always be possible due to the sensor and optics manufacturing technology's inherent limitations. These issues can be solved by using relatively inexpensive image processing algorithms, giving rise to the idea of super-resolution. It has the benefit of being less expensive, and current low-resolution imaging systems can still be used.

The concept behind super-resolution is that a low-resolution (noisy) series of images of a scene can be combined to create a high-resolution image or image sequence. Thus, it attempts to reconstruct the original scene image with high resolution given a set of observed images at lower resolution. The general method treats low-resolution images as the product of resampling a high-resolution image. The aim is to recover the high-resolution image, which, when resampled using the input images and imaging model, yields the low-resolution observed images. Thus, for super-resolution, the accuracy of the imaging model is critical, and incorrect modelling, such as of motion, will actually degrade the image further. The images observed may have come from one or more cameras, or they could be frames from a video series. These images must be aligned with one another in a standard reference frame. This is the registration procedure. In the aligned composite image, the super-resolution technique can then be applied to a region of interest. Accurate alignment, or registration, and the formulation of an effective forward image model are essential for successful super-resolution.

Although a significant progress has been made in this direction, a problem of SR is that for any LR input, there exists multiple solutions. Tackling such problem requires methods that can learn nonlinear mapping between LR and HR image pairs. There had been many traditional methods for this problem, but deep convolutional neural network (CNN) based methods have gained a significant edge over the others in place. In general, as the breadth of a CNN's receptive field deepens, the network's receptive field expands, allowing it to capture not just local but also global features. However, the complexity of training deeper networks increases to some degree at the same time.

To address this challenge, identity mappings in residual networks (ResNet) rendered deep network training relatively simple, and subsequent SR research has been heavily influenced by this concept. A significant number of CNN-based approaches have recently been modified on a regular basis. However, most of the SR models listed above are committed to deeper and more complex network architectures, almost completely ignoring resource and time taken, and they are complex in nature, have a large number of parameters, and are more difficult to train. Also, most CNN-based models treat channel-wise feature maps in the same way, limiting their ability to handle different levels of information among

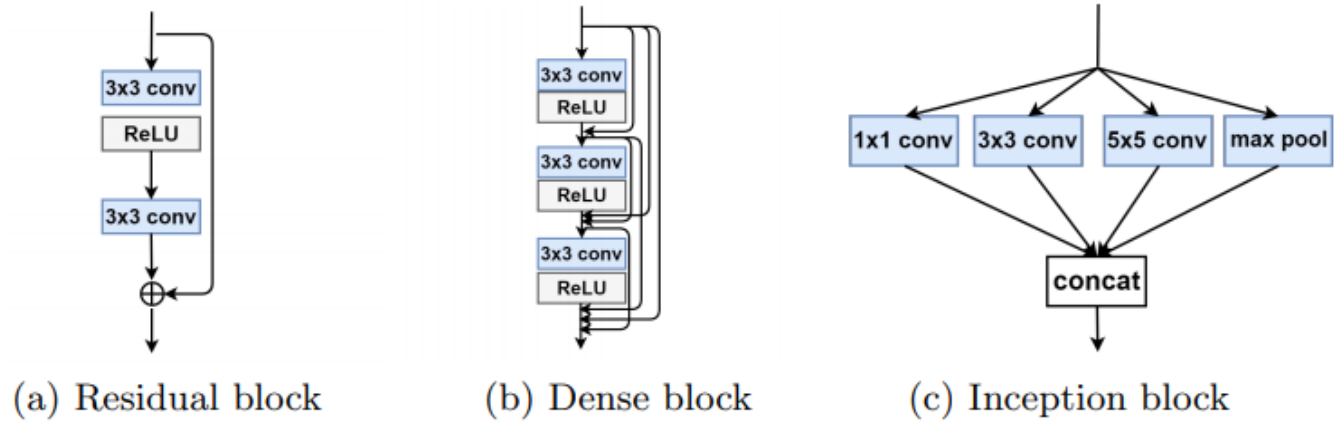


Fig. 1. Different widely used networks, widely used for feature extraction.

discriminative channels. Regardless in which direction these models are applied, they typically reduce the representational capacity of CNNs to some degree.

To effectively resolve these practical issues, this project tries to replicate the novel SR model called the multi-scale residual channel attention network (MSRCAN) [1], which, when compared to existing CNN-based SR methods, can achieve better efficiency without losing too many resources. Instead of directly replicating the aforementioned paper, we have tried to implement some modifications in this project and experimented with some different combinations. Specifically, a multi-scale residual channel attention block (MSRCAB) is exploited. Using the MSRCAB in the proposed method can not only plenarily detect the image features with different convolutional kernels, but also adaptively learn more informative channel-wise features. The results achieved from this project are compared with the same from the original paper, along with some state-of-the-art methods. Surprisingly, the proposed method produces almost similar results to that of the original paper, validating the reproducibility of the research project.

*The contributions of this project:*

1) A residual channel attention block is designed to effectively extract image features of different scales and adaptively recalibrate features by modelling the interdependencies among discriminative channel-wise features.
2) Residual in Residual (RIR) block is used to construct very deep trainable network. The long and short skip connections in RIR block help the redundant information bypass and force the main network learn effective information.
3) A channel attention mechanism is also used.

### A. Background Study

*1) Single Image Super Resolution(SISR):* The SISR problem can be divided into three major stages roughly: interpolation-based algorithms, reconstruction-based algorithms, and learning-based algorithms. Early approaches use interpolation techniques based on sampling theory like linear or bicubic. Those methods run fast, but cannot rebuild the detailed, realistic textures. Improved works aim to establish complex mapping functions between LR and HR images. Those methods rely on techniques ranging from neighbour embedding to sparse coding. Recent works tend to build an end-to-end CNNs model to learn mapping functions from LR to HR images by using large training datasets. Since Dong et al. proposed the SRCNN [2] model, various CNNs architectures have been used on SISR problem. Previous work often used pre-processed LR image as input, which was up scaled to HR space via an up sampling operator as bicubic. However, this method has been proved [3] that it will add computational complexity and produce visible artifacts. To avoid this, new methods are proposed, such as Fast Super-Resolution Convolutional Neural Networks (FSRCNN [4]) and Efficient Sub-pixel Convolutional Networks (ESPCN [3]). All of the models mentioned above are shallow networks (less than 5 layers). Kim et al. [9] first introduced the residual architecture for training much deeper network (20 layers) and achieved great performance. After that, many SR models have been proposed, including DRCN [5], DRNN [7], LapSRN [6], SRResNet [8], etc. Unfortunately, these models become more and more deeper and extremely difficult to train.

*2) Attention Mechanism:* Generally, attention can be viewed as a guidance to bias the allocation of available processing resources towards the most informative components of an input. Recently, tentative works have been proposed to apply attention into deep neural networks, ranging from localization and understanding in images to sequence-based networks . It's usually combined with a gating function (e.g., sigmoid) to rescale the feature maps. Wang et al. [11] proposed residual attention network for image classification with a trunk-and-mask attention mechanism. Hu et al. [10] proposed squeeze-and-excitation (SE) block to model channel-wise relationships to obtain significant performance improvement for image classification. However, few works have been proposed

to investigate the effect of attention for low-level vision tasks (e.g., image SR).

*3) Feature Extraction Block:* Nowadays, many feature extraction blocks have been proposed. The main idea of the inception block is to find out how an optimal local sparse structure works in a convolutional network. However, these different scale features simply concatenate together, which leads to the underutilization of local features. In 2016, Kim et al. [9] proposed a residual learning framework to ease the training of networks so that they could achieve more competitive results. After that, Huang et al. introduced the dense block. Residual block and dense block use a single size of convolutional kernel and the computational complexity of dense blocks increases at a higher growth rate. In order to solve these drawbacks, we propose a multi-scale residual block.

## II. Proposed Method

### A. Network Architecture

From Figure 2, we can see that our RCAN mainly consists four parts: initial feature extraction, residual in residual (RIR) deep feature extraction, upscale module, and reconstruction part. Let $I_{LR}$ be the LR input and $I_{SR}$ the HR output of the RCAN.

**Initial feature extraction stage**: We are using only one convolutional layer to extract the inital feature $F_0$ from $I_{LR}$.

$$F_0 = H_{SF}\left(I_{LR}\right), \tag{1}$$

where $H_{SF}\left(\cdot\right)$ denotes the convolution operation. $F_0$is the input to the RIR stage.

**RIR stage**: Residual in Residual (RIR) module is used for for deep feature extraction with $F_0$as input. So we have

$$F_{DF} = H_{RIR}\left(F_0\right), \tag{2}$$

where $H_{RIR}\left(\cdot\right)$ denotes the residual in residual staructure which contains the residual groups. RIR achieves a large depth and provides very large receptive field size.

**Upscale stage**: The deep feature output from the RIR module is now upscaled by an upscale module.

$$F_{UP} = H_{UP}\left(F_{DF}\right), \tag{3}$$

where $H_{UP}\left(\cdot\right)$ and $F_{UP}$denote the upscale module and upscaled features repectively.

There are several choices for upscale modules like transposed convolution layer, nearest-neighbor upsampling + convolution and ESPCN. Taking computational complexity and reconstruction performance into account, transposed convolution layer is adopted to enlarge the feature maps here.

Transposed convolution layer can be regarded as the inverse of convolution. In the convolution operation, an image when convolved with a filter with stride *k*, the output size is 1/*k* times the input. Thus, in a transversed convolution operation, the output will become *k* times the input. Thus, we can magnify the target image with a desired upscaling factor *r* by setting the stride $k = r$.

**Reconstruction stage**: The upscaled feature is reconstructed using a single convolution layer.

$$I_{SR} = H_{REC}\left(F_{UP}\right) = H_{RCAN}\left(I_{LR}\right), \tag{4}$$

where $H_{REC}\left(\cdot\right)$ and $H_{RCAN}\left(\cdot\right)$ denote the reconstruction layer and the overall function of RCAN respectively.

Then the RCAN is optimized with the help of a loss function. Given a training set $\left\{I_{LR}^i, I_{HR}^i\right\}_{i=1}^N$, which contains $N$ LR inputs and HR outputs. The objective of training RCAN is to minimize the loss function

$$L\left(\Theta\right) = \frac{1}{N}\sum_{i=1}^{N} H_{RCAN}\left(I_{LR}^i\right) - {I_{HR}^i}^{\,2}, \tag{5}$$

where $\Theta$ denotes the parameter set of the proposed network. The loss function is optimized by using stochastic gradient descent.

### B. Residual in Residual (RIR) Structure

Our proposed RIR structure contains G residual groups (RG) and long skip connection (LSC). Each RG further contains B residual channel attention blocks (RCAB) with short skip connection (SSC). This RIR structure allows us to train very deep CNN (over 400 layers) for image super-resolution with higher performance.

Inspired by previous works in SRRestNet [26] and EDSR [25], we use RG as the basic module for deep networks. A RG in the *g*-th group can be expressed as:

$$F_g = H_g(F_{g-1}) = H_g\left(H_{g-1}\left(\cdots H_1\left(F_0\right)\cdots\right)\right), \tag{6}$$

where $H_g$ denotes the *g*-th RG. It has been demonstrated that stacked residual blocks can be used to construct deep CNN, but in image SR, a very deep network with many number of RGs would fail to achieve better performance. To solve this problem, we are using a long skip connection (LSC) in the structure to stabilize the training of the neural network. LSC increases performance with residual learning via

$$F_{DF} = F_0 + W_{LSC}F_G = F_0 + W_{LSC}H_g\left(H_{g-1}\left(\cdots H_1\left(F_0\right)\cdots\right)\right), \tag{7}$$

where $W_{LSC}$ is the weight of the convolution layer at the tail of RIR. We have omitted the bias term for simplicity. LSC increases the flow of information across RGs and makes it possible for RIR to learning residual information at a coarse level.

As discussed in the introduction, there are lots of information in the LR input data and we want to recover all the useful information. This low-frequency information can be bypassed through an identity-based skip connection. For performance in residual learning, we stack $B$ residual channel attention blocks (RCABs) in each RG. The *b*-th RCAB in th *g*-th RG can be expressed as:

$$F_{g,b} = H_{g,b}\left(F_{g,b-1}\right) = H_{g,b}\left(H_{g,b-1}\left(\cdots H_{g,1}\left(F_{g-1}\right)\cdots\right)\right), \tag{8}$$

where $F_{g,b-1}$ and $F_{g,b}$ are the input and output of the *b*-th RCAB in the *g*-th RG. The overall function of the RCAB is

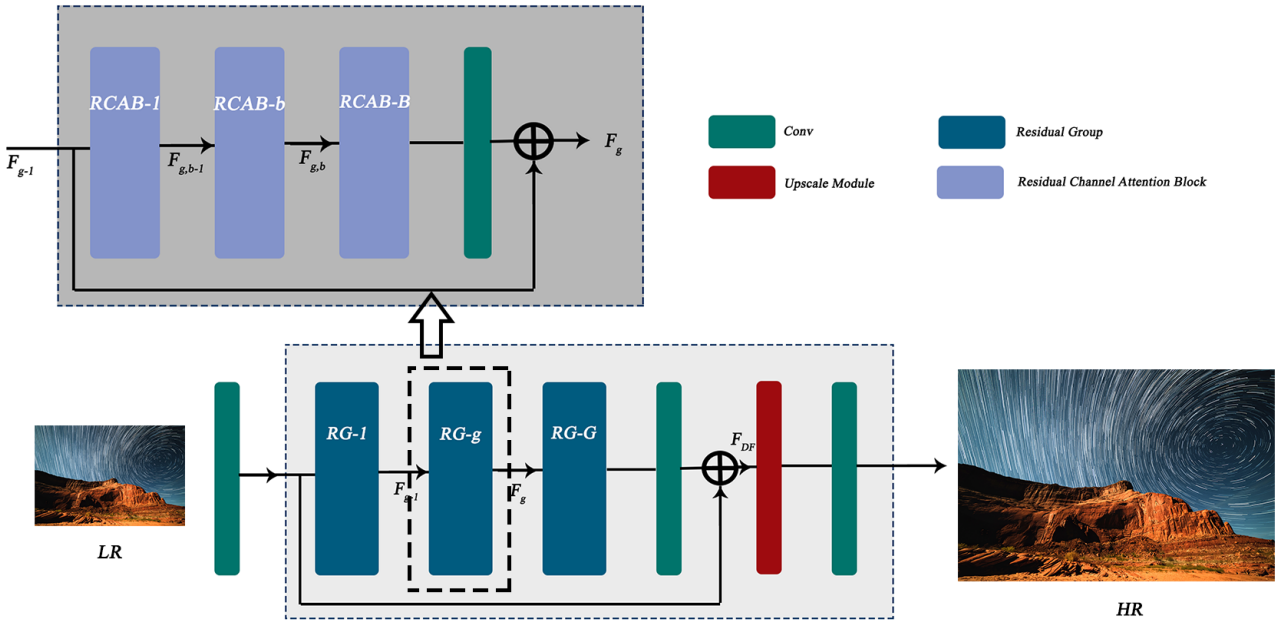


Fig. 2. Network architecture of the residual channel attention network (RCAN)

denoted by $H_{g,b}$. A short skip connection (SSC) is introduced to increase the sensitivity towards more informative features. This can be shown as:

$$F_g = F_{g-1} + W_g F_{g,B} = F_{g-1} + W_g H_{g,B}\left(H_{g,B-1}\left(\cdots H_{g,1}\left(F_{g-1}\right)\cdots\right)\right), \quad (9)$$

where $W_g$ is the weight set to the convolution layer at the end of the $g$-th RG. The SSC helps the main parts of the network to learn residual information. So, with LSC and SSC, low-frequency information can be bypassed in the training process.

### C. Channel Attention (CA)

Earlier CNN-based SR methods treat channel-wise features without any processing, which is not flexible for real cases and results in negligence of inter-dependencies among channels. So, to extract more informative features, we exploit the inter dependencies among feature channels leading to a channel attention (CA) mechanism. (Figure 3). Now we will explain how to explicitly model the dependencies of the channels.

We know that information in the LR space has abundant low-frequency and valuable high-frequency components. The low frequency parts are more complanate while high-frequency components while high-frequency components would include edges, textures and other details. Also, each filter in convolution layer operates with a local receptive field. Consequently, the output after convolution is unable to exploit contextual information outside of the local region.

Thus, we take the channel-wise global spatial information into a channel descriptor using global average pooling. Let, $X = [x_1, \cdots, x_c, \cdots, x_C]$ be the input which has $C$channels with dimension of $H \times W$. The channel-wise statistic $z \in \mathbb{R}^C$ can be obtained by shrinking $X$. Then the $c$-th element of $z$ is determined by

$$z_c = H_{GP}\left(x_c\right) = \frac{1}{H \times W}\sum_{i=1}^{H}\sum_{j=1}^{W} x_c\left(i,j\right), \quad (10)$$

where $x_c\left(i,j\right)$ is the the value at position $(i,j)$ of $c$-th feature $x_c$. $H_{GP}$ is the global average pooling function. Such channel statistic can be thought of a collection of the local descriptors, whose statistics contribute to express the whole image.

Now we use a gating mechanism to fully capture the channel-wise dependencies from the aggregated information by global average pooling. The gating mechanism learns the the nonlinear interactions between channels. Here we are using the gating mechanism with sigmoid function.

$$s = f\left(W_U \delta\left(W_D z\right)\right), \quad (11)$$

where $f\left(\cdot\right)$ and $\delta\left(\cdot\right)$ denote the sigmoid function and ReLU function, respectively. $W_D$ is the weight of a Conv layer, which reduces the number of channels with reduction ratio $r$. After activation by ReLU, the low-dimension signal is upscaled with ratio $r$ by an upscaling layer, whose weight set is $W_U$. Then the sigmoid function is used to generate the final statistics $s$, which is used to rescale the input $x_c$

$$\widehat{x}_c = s_c \cdot x_c, \quad (12)$$

where $s_c$ and $x_c$ are the scaling factor and feature map in the $c$-th channel. With channel attention, the residual component in the RCAB is adaptively rescaled.

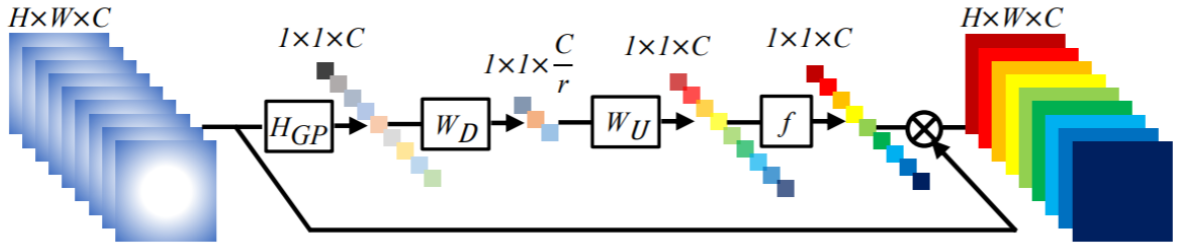


Fig. 3. Channel attention (CA). $\otimes$ denotes element-wise product

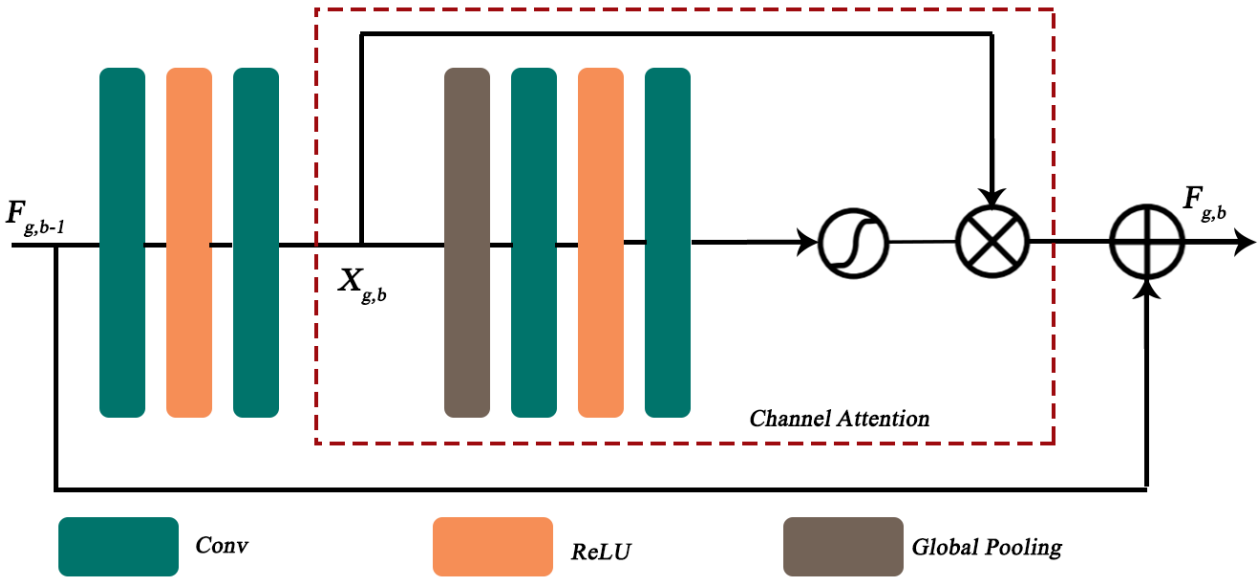


Fig. 4. Residual channel attention block (RCAB)

### D. Residual Channel Attention Block (RCAB)

We know that residual residual groups and long skip connection enables the network to focus on informative parts of the LR features. Channel attention extracts the channel statistics to further enhance the discriminating ability of the network.

So, we integrate CA into RB and propose residual channel attention block (RCAB) ( see Figure 4). For the b-th RB in $g$-th RG, we have

$$F_{g,b} = F_{g,b-1} + R_{g,b}\left(X_{g,b}\right) \cdot X_{g,b}, \tag{13}$$

where $F_{g,b}$ and $F_{g,b-1}$ are the input and output of RCAB, which learns the residual $X_{g,b}$ from the input. $R_{g,b}$ denotes the CA function. The residual component is obtained by two stacked convolutional layers

$$X_{g,b} = W^2_{g,b}\delta\left(W^1_{g,b}F_{g,b-1}\right), \tag{14}$$

where $W^1_{g,b}$ and $W^2_{g,b}$ are the weight sets of the two stacked Conv layers in RCAB.

Based on RCAB and RIR structure, we construct a very deep RCAN for highly accurate image SR and achieve good performance.

## III. Experiments

In order to replicate the paper and the methods used, we have done several experiments regarding the model training, dataset, implementation, result evaluation, ablation studies, etc. We have compared the results obtained from our project with that of the original paper, along with the results obtained from state-of-the-art method, to justify the effectiveness of the project.

### A. Dataset Description

The model was trained on DIV2K dataset [12] that contains 1000 image, split in 8:1:1 ratio for training, testing and validation. The dataset contains SR images along with their LR counterparts for X2, X3, and X4 scales. Post training, the model was tested on four publicly available benchmark datasets named as Set5 [15], Set14 [16], BSDS100 [14],

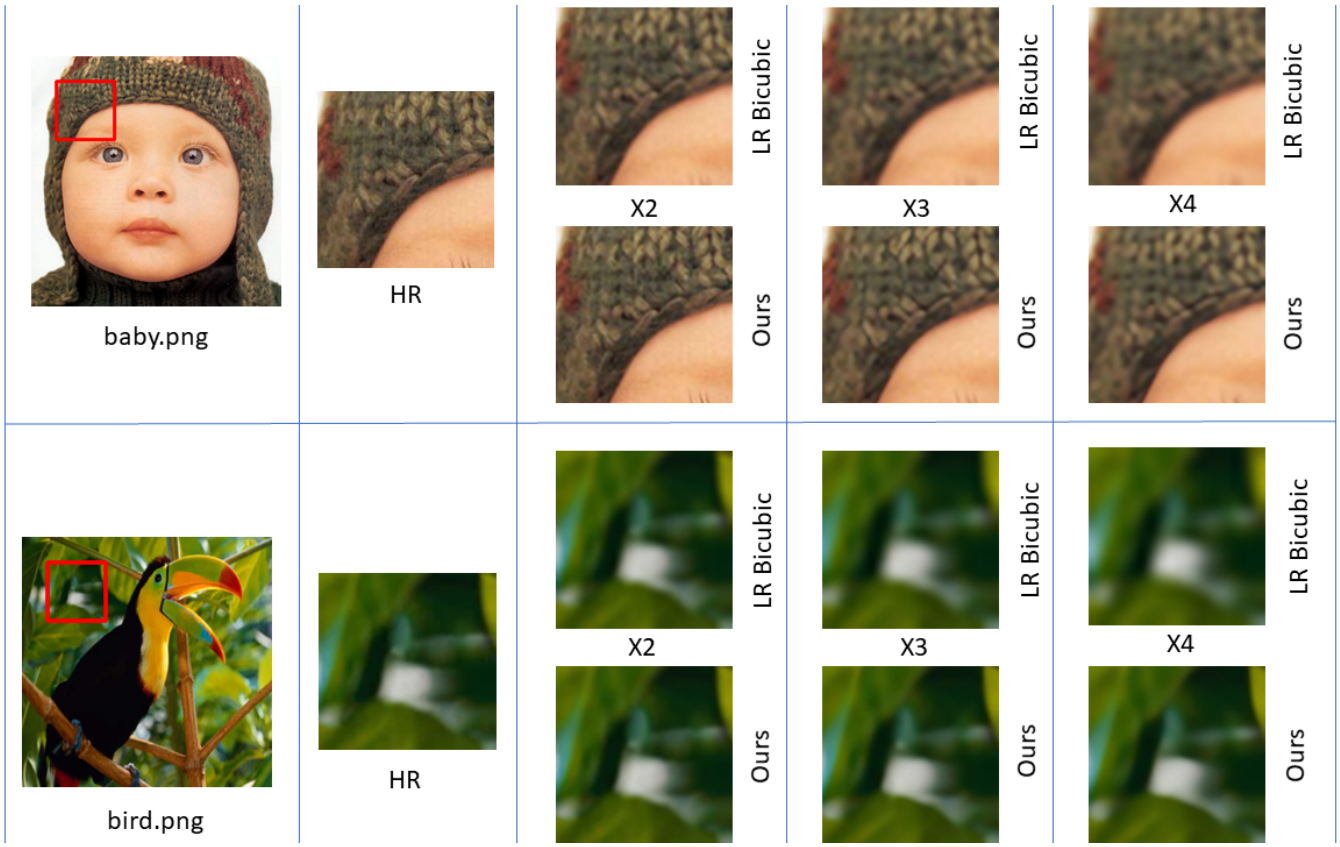


Fig. 5. Visual results on Set5 dataset

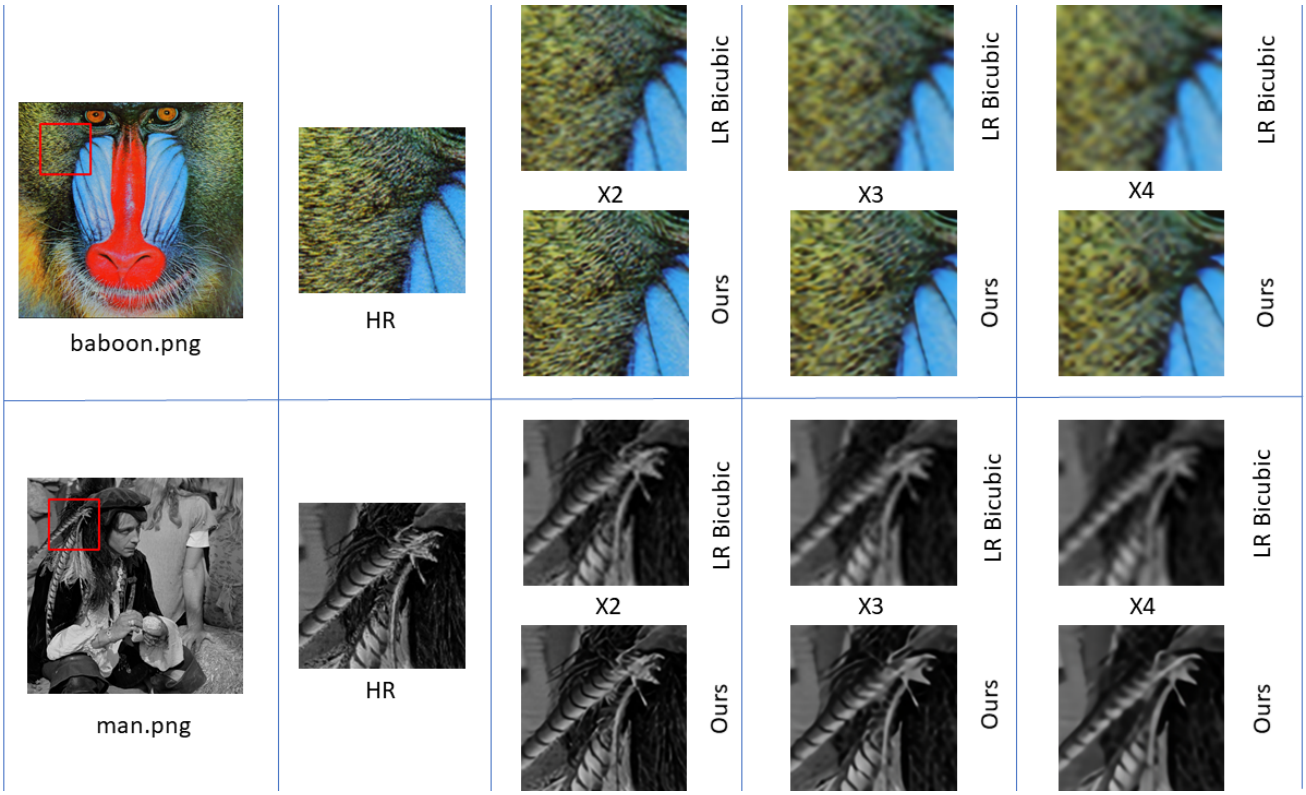


Fig. 6. Visual results on Set14 dataset

and Urban100 [13] containing 5, 14, 100 and 100 images respectively for each classes. Bicubic degradation was used to produce the LR counterparts of the HR images in these datasets.

### B. Performance Evaluation

In order to evaluate the performance of the model, we have used three standard evaluation metrics namely Peak Signal to Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM) and Mean Squared Error (MSE). These are discussed as follows:

*1) Mean Square Error (MSE):* The Mean square error, as the name, suggests helps in assessing the cumulative error between the original image and the compressed image, whereas the peak error is measured by the PSNR scale. Likewise while measuring the Mean square error, the lower the MSE value the lower will be the error quotient. MSE can be calculated between two images $I$ and $K$ as follows:

$$MSE(I,K) = \frac{1}{m \times n} \sum_{i=1}^{m} \sum_{j=1}^{n} [I(i,j) - K(i,j)]^2 \quad (15)$$

where the images $I$, $K$ has dimensions of $m \times n$

*2) Peak Signal to Noise Ratio (PSNR):* Peak signal-to-noise ratio (PSNR) is the ratio between the maximum possible power of an image and the power of corrupting noise that affects the quality of its representation. To estimate the PSNR of an image, it is necessary to compare that image to an ideal clean image with the maximum possible power. PSNR is most commonly used to measure the quality of reconstruction of lossy compression codecs (e.g., for image compression). The signal in this case is the original data, and the noise is the error introduced by compression. When comparing compression codecs, PSNR is an approximation to human perception of reconstruction quality. PSNR is most easily defined by using MSE (already discussed before) as follows:

$$PSNR = 10 \cdot log_{10}\left(\frac{MAX_I^2}{MSE}\right) = 20 \cdot log_{10}(MAX_I) - 10 \cdot log_{10}(MSE) \quad (16)$$

where $MAX_I$ represents the maximum pixel intensity value of image $I$.

*3) Structural Similarity Index Measure (SSIM):* The structural similarity index measure (SSIM) is a method for predicting the perceived quality of digital television and cinematic pictures, as well as other kinds of digital images and videos. SSIM is used for measuring the similarity between two images. The SSIM index is a full reference metric; in other words, the measurement or prediction of image quality is based on an initial uncompressed or distortion-free image as reference. SSIM is a perception-based model that considers image degradation as perceived change in structural information, while also incorporating important perceptual phenomena, including both luminance masking and contrast masking terms.

## IV. Results and Discussions

In this section, we will discuss the quantitative and qualitative performance measurement of the model on the benchmark datasets.

### A. Results on Set5 dataset

Set5 dataset contains 5 high resolution images. We calculated PSNR, SSIM and MSE between the HR images and the output images from the model in three different scales to evaluate the performance of the model. Visual results on Set5 dataset are shown in Figure 5. Table I shows the average PSNR, SSIM and MSE value on Set5 dataset at three different scales.

TABLE I
THE QUANTITATIVE RESULT ON SET5 DATASET AT DIFFERENT SCALES.

| Dataset | Scale | Average PSNR | Average SSIM | Average MSE |
|---|---|---|---|---|
| Set 5 | X2 | 36.028446 | 0.9450757 | 0.000303297 |
| | X3 | 31.32214 | 0.892935 | 0.000834603 |
| | X4 | 30.532715 | 0.8696495 | 0.00098598 |

### B. Results on Set14 dataset

Set14 dataset contains 14 high resolution images. We calculated PSNR, SSIM and MSE between the HR images and the output images from the model in three different scales to evaluate the performance of the model. Visual results on Set14 dataset are shown in Figure 6. Table II shows the average PSNR, SSIM and MSE value on Set14 dataset at three different scales.

TABLE II
THE QUANTITATIVE RESULT ON SET14 DATASET AT DIFFERENT SCALES.

| Dataset | Scale | Average PSNR | Average SSIM | Average MSE |
|---|---|---|---|---|
| Set 14 | X2 | 31.86751 | 0.8984707 | 0.000928365 |
| | X3 | 26.481241 | 0.7844373 | 0.002778149 |
| | X4 | 26.139751 | 0.736015 | 0.003415856 |

### C. Results on BSDS100 dataset

BSDS100 dataset contains 100 high resolution images. We calculated PSNR, SSIM and MSE between the HR images and the output images from the model in three different scales to evaluate the performance of the model. Visual results on BSDS100 dataset are shown in Figure 7. Table III shows the average PSNR, SSIM and MSE value on BSDS100 dataset at three different scales.

TABLE III
THE QUANTITATIVE RESULT ON BSDS100 DATASET AT DIFFERENT SCALES.

| Dataset | Scale | Average PSNR | Average SSIM | Average MSE |
|---|---|---|---|---|
| BSDS 100 | X2 | 31.022408 | 0.8932048 | 0.001192211 |
| | X3 | 27.990114 | 0.7937682 | 0.002287413 |
| | X4 | 26.464645 | 0.7207724 | 0.003146333 |

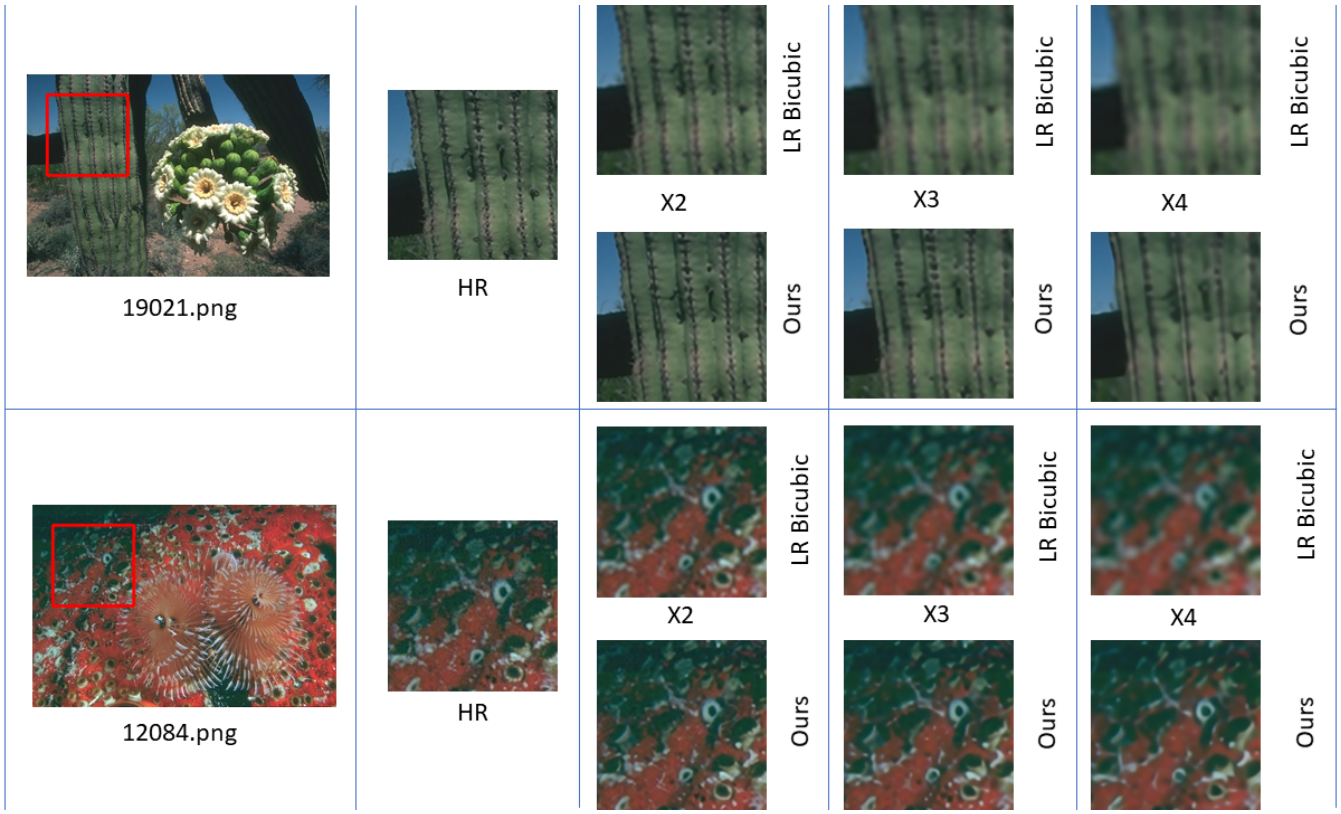


Fig. 7. Visual results on BSDS100 dataset

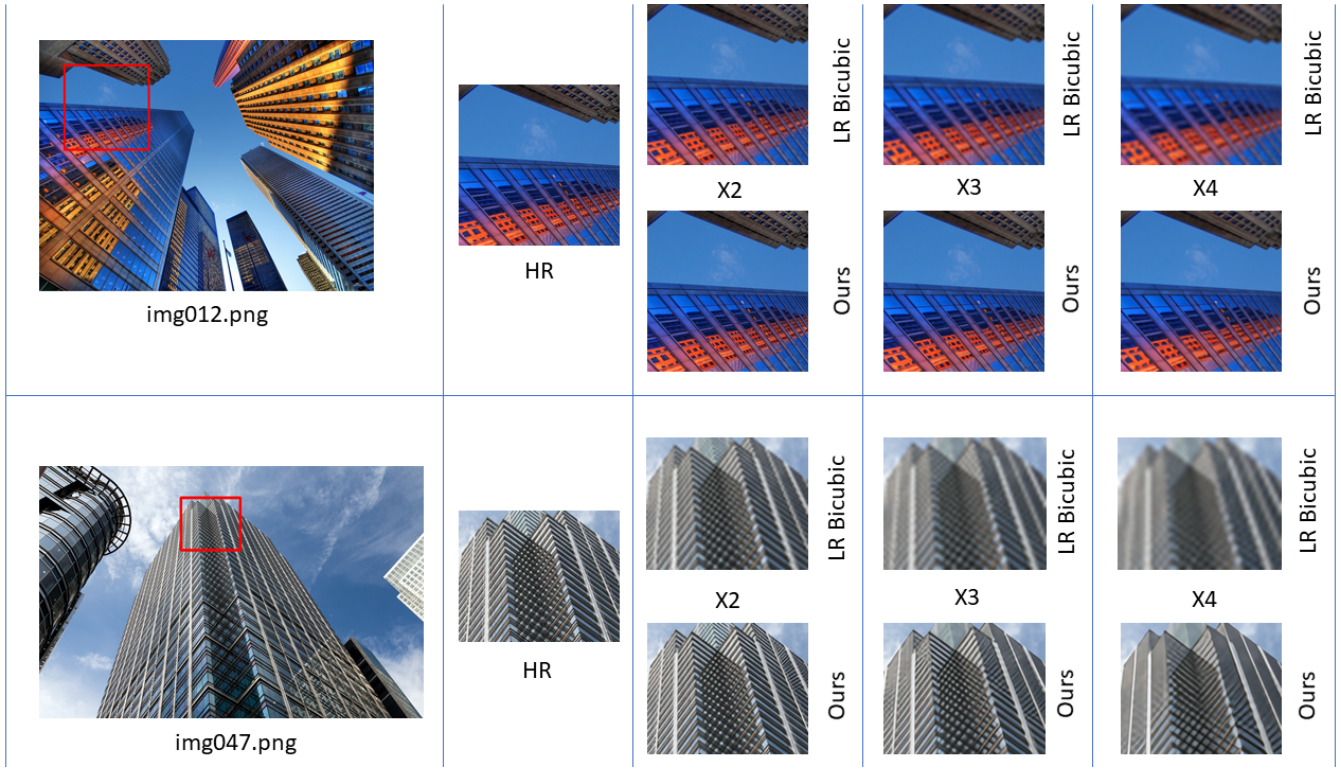


Fig. 8. Visual results on Urban100 dataset

TABLE IV
QUANTITATIVE COMPARISON OF PSNR/SSIM MEASURE OF THE RESULT OBTAINED FROM THIS PROJECT WITH DIFFERENT EXISTING METHODS ALONG WITH THE ORIGINAL PAPER.

| Scale | Method | Set5 | Set14 | BSD100 | Urban100 |
|---|---|---|---|---|---|
| X2 | SRCNN | 36.66/0.9542 | 32.42/0.9063 | 31.36/0.8879 | 29.50/0.8946 |
| | ESPCN | 37.00/0.9559 | 32.75/0.9098 | 31.51/0.8939 | 29.87/0.9065 |
| | FSRCNN | 37.06/0.9554 | 32.76/0.9078 | 31.53/0.8912 | 29.88/0.9024 |
| | VDSR | 37.53/0.9587 | 33.03/0.9124 | 31.90/0.8960 | 30.76/0.9140 |
| | DRCN | 37.63/0.9584 | 33.06/0.9108 | 31.85/0.8947 | 30.76/0.9147 |
| | LapSRN | 37.52/0.9591 | 33.08/0.9109 | 31.80/0.8949 | 30.41/0.9112 |
| | DRRN | 37.74/0.9591 | 33.23/0.9136 | 32.05/0.8973 | 31.23/0.9188 |
| | MemNet | 37.78/0.9597 | 33.28/0.9142 | 32.08/0.8978 | 31.31/0.9195 |
| | EDSR | 38.11/0.9601 | 33.92/0.9195 | 32.32/0.9013 | 32.93/0.9351 |
| | MSRCAN (Original paper) | 37.90/0.9601 | 33.37/0.9158 | 32.10/0.8989 | 31.72/0.9242 |
| | **Ours** | **36.03/0.9451** | **31.87/0.8985** | **31.02/0.8932** | **31.42/0.9276** |
| X3 | SRCNN | 32.75/0.9090 | 29.30/0.8215 | 28.41/0.7863 | 26.24/0.7989 |
| | ESPCN | 33.02/0.9135 | 29.49/0.8271 | 28.50/0.7937 | 26.41/0.8161 |
| | FSRCNN | 33.18/0.9140 | 29.37/0.8240 | 28.53 0.7910 | 26.43/0.8080 |
| | VDSR | 33.66/0.9213 | 29.77/0.8314 | 28.82/0.7976 | 27.14/0.8279 |
| | DRCN | 33.85/0.9215 | 29.89/0.8317 | 28.81/0.7954 | 27.16/0.8311 |
| | LapSRN | 33.82/0.9227 | 29.79/0.8320 | 28.82/0.7973 | 27.07/0.8272 |
| | DRRN | 34.03/0.9244 | 29.96/0.8349 | 28.95/0.8004 | 27.53/0.8378 |
| | MemNet | 34.09/0.9248 | 30.00/0.8350 | 28.96/0.8001 | 27.56/0.8376 |
| | EDSR | 34.65/0.9282 | 30.52/0.8462 | 29.25/0.8093 | 28.79/0.8655 |
| | MSRCAN (Original paper) | 34.21/0.9255 | 30.09/0.8376 | 28.99/0.8020 | 27.72/0.8397 |
| | **Ours** | **31.32/0.8929** | **26.48/0.7844** | **27.99/0.7938** | **25.93/0.7546** |
| X4 | SRCNN | 30.48/0.8628 | 27.49/0.7503 | 26.90/0.7101 | 24.52/0.7221 |
| | ESPCN | 30.66/0.8646 | 27.71/0.7562 | 26.98/0.7124 | 24.60/0.7360 |
| | FSRCNN | 30.72/0.8660 | 27.61/0.7550 | 26.98/0.7150 | 24.62/0.7280 |
| | VDSR | 31.35/0.8830 | 28.02/0.7680 | 27.29/0.0726 | 25.18/0.7540 |
| | DRCN | 32.56/0.8810 | 28.15/0.7627 | 27.24/0.7150 | 25.25/0.7530 |
| | LapSRN | 31.54/0.8855 | 28.19/0.7720 | 27.32/0.7280 | 25.21/0.7553 |
| | DRRN | 31.68/0.8888 | 28.21/0.7721 | 27.38/0.7284 | 25.44/0.7638 |
| | MemNet | 31.74/0.8893 | 28.26/0.7723 | 27.40/0.7281 | 25.50/0.7630 |
| | EDSR | 32.46/0.8968 | 28.80/0.7876 | 27.71/0.7420 | 26.64/0.8033 |
| | MSRCAN (Original paper) | 31.89/0.8907 | 28.33/0.7753 | 27.45/0.7308 | 25.75/0.7733 |
| | **Ours** | **30.53/0.8696** | **26.14/0.7360** | **26.46/0.7208** | **23.91/0.7546** |

### *D. Results on Urban100 dataset*

Urban100 dataset contains 100 high resolution outdoor scene images. We calculated PSNR, SSIM and MSE between the HR images and the output images from the model in three different scales to evaluate the performance of the model. Visual results on Urban100 dataset are shown in Figure 8. Table V shows the average PSNR, SSIM and MSE value on Urban100 dataset at three different scales.

TABLE V
THE QUANTITATIVE RESULT ON URBAN100 DATASET AT DIFFERENT SCALES.

| Dataset | Scale | Average PSNR | Average SSIM | Average MSE |
|---|---|---|---|---|
| Urban 100 | X2 | 31.423206 | 0.9276064 | 0.001179445 |
| | X3 | 25.93362 | 0.8275486 | 0.003984207 |
| | X4 | 23.907404 | 0.7546184 | 0.006032668 |

### *E. Comparison with different other methods*

In this section, we have compared the result of this project with the original paper. Also, we have shown quantitative comparison with different state-of-the-art methods from literature that have effectively performed well in the past. The quantitative results of these papers have been directly obtained from these papers. The results have been compared with

these following papers: SRCNN [19], ESPCN [20], FSRCNN [19], VDSR [18], DRCN [21], LapSRN [22], DRRN [23], MemNet [24], EDSR [25], MSRCAN (original paper) [1]. The comparison is table is shown in Table IV. The bold texts indicate the results from our project whereas the red texts indicate the results from the original paper. MSE could not be compared because it was missing in most of the other papers.

## V. Conclusion

From Table IV it is visible that the result produced from the project is quite similar to the original result from the paper. Though, the result or the original paper is superior, probably because of the usage of multi-scale RCAB module. Besides, the model parameters have been set by the authors through extensive research and fine-tuning, which was not possible in this project because of computational resources and time constraints. All in all, we would like to thank our mentor Dr. Amitava Chatterjee, our department and our classmates for their constant support and supervision, whenever required.

## References


[1] Cao F, Liu H. Single image super-resolution via multi-scale residual channel attention network. Neurocomputing. 2019 Sep 17;358:424-36.

[2] Dong, C., Loy, C.C., He, K., Tang, X.: Learning a deep convolutional network for image super-resolution. In: European conference on computer vision, Springer (2014) 184–199

[3] Shi, Wenzhe and Caballero, Jose and Husz´ar, Ferenc and Totz, Johannes and Aitken, Andrew P and Bishop, Rob and Rueckert, Daniel and Wang, Zehan.: 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) 1874–1883

[4] Dong, C., Loy, C.C., Tang, X.: Accelerating the super-resolution convolutional neural network. In: European Conference on Computer Vision, Springer (2016) 391–407

[5] Kim, J., Kwon Lee, J., Mu Lee, K.: Deeply-recursive convolutional network for image super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) 1637–1645

[6] Lai, W.S., Huang, J.B., Ahuja, N., Yang, M.H.: Deep laplacian pyramid networks for fast and accurate superresolution. In: IEEE Conference on Computer Vision and Pattern Recognition. Volume 2. (2017) 5

[7] Tai, Y., Yang, J., Liu, X.: Image super-resolution via deep recursive residual network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Volume 1. (2017) 5

[8] Ledig, C., Theis, L., Husz´ar, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al.: Photo-realistic single image superresolution using a generative adversarial network. In: The IEEE conference on computer vision and pattern recognition. (2017)

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

[10] Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. arXiv preprint arXiv:1709.01507 (2017)

[11] Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., Wang, X., Tang, X.: Residual attention network for image classification. In: CVPR. (2017)

[12] Agustsson E, Timofte R. Ntire 2017 challenge on single image super-resolution: Dataset and study. InProceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops 2017 (pp. 126-135).

[13] Huang JB, Singh A, Ahuja N. Single image super-resolution from transformed self-exemplars. InProceedings of the IEEE conference on computer vision and pattern recognition 2015 (pp. 5197-5206).

[14] Martin D, Fowlkes C, Tal D, Malik J. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. InProceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001 2001 Jul 7 (Vol. 2, pp. 416-423). IEEE.

[15] Bevilacqua M, Roumy A, Guillemot C, Alberi-Morel ML. Low-complexity single-image super-resolution based on nonnegative neighbor embedding.

[16] Zeyde R, Elad M, Protter M. On single image scale-up using sparse-representations. InInternational conference on curves and surfaces 2010 Jun 24 (pp. 711-730). Springer, Berlin, Heidelberg.

[17] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," IEEE Trans. Patt. Anal. Mach. Intell., 38(2) 295-307, 2016.

[18] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in Proc. IEEE Conf. Comput. Vis. Patt. Recogn. (CVPR), Jun. 2016, pp. 1646-1654.

[19] C. Dong, C. C. Loy, and X. Tang, "Accelerating the super-resolution convolutional neural network," in Proc. Eur. Conf. Comput. Vis. (ECCV), Oct. 2016, pp. 391-407.

[20] W. Shi, J. Caballero, F. Huszar, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient subpixel convolutional neural network," in Proc. IEEE Conf. Comput. Vis. Patt. Recogn. (CVPR), Jun. 2016, pp. 1874-1883.

[21] J. Kim, and J. K. Lee, and K. M. Lee, "Deeply-recursive convolutional network for image super-resolution," in Proc. IEEE Conf. Comput. Vis. Patt. Recogn. (CVPR), Jun. 2016, pp. 1637-1645.

[22] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Deep Laplacian pyramid networks for fast and accurate super-resolution," in Proc. IEEE Conf. Comput. Vis. Patt. Recogn. (CVPR), Jul. 2017, pp. 624-632.

[23] Y. Tai, J. Yang, and X. Liu, "Image super-resolution via deep recursive residual network," in Proc. IEEE Conf. Comput. Vis. Patt. Recogn. (CVPR), Jul. 2017, pp. 3147- 3155.

[24] Y. Tai, J. Yang, X. Liu and C. Xu, "MemNet: A persistent memory network for image restoration," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 4549-4557.

[25] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," in Proc. IEEE Conf. Comput. Vis. Patt. Recogn. (CVPR), Jul. 2017, pp. 136-144.

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