Novel Mobile-Oriented Neural Networks for Single Image Super-Resolution

Jiawen Lin

A Thesis

Presented to the University of Dublin, Trinity College

in partial fulfilment of the requirements for the degree of

Master of Science in Computer Science (Intelligent Systems)

Supervisor: Fergal Shevlin

September 2020

Declaration

I, the undersigned, declare that this work has not previously been submitted as an exercise for a degree at this, or any other University, and that unless otherwise stated, is my own work.

Jiawen Lin

September 7, 2020

Permission to Lend and/or Copy

I, the undersigned, agree that Trinity College Library may lend or copy this thesis upon request.

Jiawen Lin

September 7, 2020

Acknowledgments

First of all, I would like to thank my supervisor Fergal Shevlin for supporting me. He allowed me to freely explore areas of my interest, provided valuable feedback and important weekly meetings throughout the whole academic year.

Secondly, I thank my family for their great support and continuous encouragement in my education. I am especially grateful to my mother for giving me emotional support during this difficult outbreak of corona-virus.

Finally, I would extend my gratitude to all the friends I met in Dublin. You are the ones that made this difficult period a treasure.

Jiawen Lin

University of Dublin, Trinity College
September 2020

Novel Mobile-Oriented Neural Networks for Single Image Super-Resolution

Jiawen Lin, Master of Science in Computer Science

University of Dublin, Trinity College, 2020

Supervisor: Fergal Shevlin

Single Image Super-Resolution (SISR) has achieved the most advanced accuracy through deep learning technology. However, how to balance between the efficiency and accuracy of super-resolution remains an open question. This dissertation discusses state-of-the-art SISR algorithms implemented on mobile devices. Several innovative algorithms are compared and discussed. Finally, I propose two efficient light-weighted SISR methods which are suitable for mobile devices. The first is NSAN, which learns hybrid residual features using non-local second-order attention network, based on which the residual HR image can be reconstructed. The second is ARSN, from which the specified residual blocks and skip connections were utilized for residual scaling, global and local residual learning. The proposed methods have different strengths; they both achieve good results in terms of performance, speed and hardware consumption, and have high practical value.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

Due to the increasing volume of the data and the accelerating development of the hardware, single image super-resolution enjoys the prevailing advancement of deep learning and become more and more attractive recently. Image super-resolution reconstruction technology has very broad application prospects, and has tremendous practical value in various fields such as aerial imaging [1], facial image improvement [2],medical image processing [3]. Single image super-resolution reconstruction belongs to typical computer vision problems, aiming to reconstruct low-resolution (LR) images into high-resolution (HR) images.

Achieving a high-resolution picture of a low-resolution picture might be a complicated issue, but the convolutional neural network has had an immense effect on this area, rendering the resulting image more fragile and realistic. In this work, I am mainly dealing with super-resolution tasks of a single image.

Given the successful history of the Convolutional Neural Network in advanced computer vision tasks, C. Dong et al. [4] proposed a CNN-based SR algorithm, specifically SRCNN. Therefore, since then, CNN has attracted more researchers to solve super-resolution tasks[5, 6, 7, 8, 9]. Although the performance has been greatly improved, there are still some problems when applying them to native mobile applications. First, and most importantly, previous research focused on introducing more complex convolutional neural networks to increase performance while ignoring computational expenses.

The significant amount of calculations cause the implementation of this algorithm on mobile devices difficult. Second, when the network complexity grows, the training phase becomes more unpredictable [7, 10], implying further technical abilities are required to prepare the network to boost efficiency. Third, some of the earlier techniques do not make fair use of the training data to generate super-resolution images.

To solve the problems mentioned above and make super-resolution appropriate for mobile devices, I propose a novel attention-inspired network structure. At the beginning, I concentrate on non-local blocks and train the self-attention learning network by catching remote dependencies. Second, I built the network to utilize the training data fully to restore images with super-resolution. It offers an unparalleled view of deep learning data enhancement, and an exceptional benefit for network architecture. I used the suggested scheme in this work to formulate a compact and effective network. Experimental findings on the benchmark data set demonstrate that the best balance between image quality and processing speed can be reached using this approach.

Later I propose a light-weight network with Automatic Residual Scaling (ARSN), which is combined with the FSRCNN[5], VDSR[11], DRCN[12], MemNet[13] and other networks. As the network layers of these methods tend to become deeper and deeper, while the network weights of the method in this paper are relatively light, and the number of layers is much less. At the same time, the method in this paper can directly input low-resolution images without bi-cubic interpolation, which can reduce additional calculations.

1.2 Objectives

For better performance, it is a design trend to deepen or widen the network. However, the result in these methods require huge computing costs and memory consumption thus not very suitable for mobile and embedded vision applications. In the meantime, conventional convolutional neural networks typically follow a cascaded network topology, such as VDSR. In this way, the feature map of the individual layer is submitted to the following layer without distinction.

My objective is to propose our lightweight model that can run on 80 per cent of the current mobile devices(Android platform), which could balance the time and the performance.

1.3 Structures

Chapter 1 briefly reviews the development of the SR field in recent years and the purpose of the research

Chapter 2 briefly introduces and analyzes the latest technology. Among them, the first part introduces the latest progress and model of SR in the traditional PC platform. The second part tests the currently limited algorithms optimized for mobile phones.

Chapter 3 is the design and test of the algorithm I proposed.

Chapter 4 tests and compares the accuracy and operating efficiency of all the algorithms involved in the paper.

Chapter 5 summarizes the main contributions and gives possible ways of future work.

Chapter 2

Related work

2.1 Development of Approaches

Among all the jobs of the computer vision, image super-resolution(SISR) is a backbone problem. The traditional method in our literature is to learn the mapping from a low-resolution image towards a high-resolution image for reconstruction. Modern machine learning approaches have been commonly used in this field, including kernel-based methods [14], PCA-based mothods[15], sparse coding approaches [16], embedding approaches [17], etc. A robust approach can secure complete use of the similarities of images without the need for additional data. [18] uses patch redundancy to create super-resolution photos. Freedman et al. [19] have later developed a localized search tool. Huang et al. [20] extend the algorithm used to direct a patch search utilizing the detected perspective geometry.

The new development of the super-resolution makes excellent use of the efficient representation capabilities of the convolutionalneural networks.

Dong et al. [4] firstly suggested SRCNN to restore high-resolution images. This structure in CNN was described as an extraction layer, followed with a non-linear mapping layer, and then a reconstruction layer corresponding to these phases in the[16].

Kim et al. [11] utilizes a deep residual CNN to improve the performance. This technology uses bicubic interpolation to upload low-resolution images to the appropriate scale; furthermore, they feed the network to produce super-resolution images. After that, super-resolution approaches based on the convolutional neural network have become a must towards better results. These approaches include LapSRN[6], DRRN[8], SRResNet[21], EDSR[7] and RCAN[9].

Nevertheless, the size of the network adds many calculations and increases the response time. Dong et al. utilize a relatively smaller filter size and a relatively deeper network, specifically FSRCNN [5] to solve this problem. This network eliminates the bicubic interpolation layer and embeds the deconvolution layer at the end of FSRCNN. In order to minimize the parameters, DRRN[8] suggested a combination of the remaining skip and recurrence links, which would decrease the running speed. CARN[22] uses several bypass connections and multilevel design to achieve a cascading method on the residual network. Aiming to use multi-scale features, Li et al. .[23] have proposed a model to achieve multi-scale features of different scales. Dai et al. proposed SAN [24] to reduce the calculation time, while He et al. proposed an ordinary differential equation (ODE) inspired structure[25]. They all implemented high-order function extractors to catch high-order statistics but skipped the convolutionallayer. Operations are local, so I have combined non-local operations with high-order statistical extractors to boost our network.

While most CNN-based super-resolution approaches have extensively promoted progress for this area, most advanced models aimlessly increase network depth and parameters, and disregard the fact that convolution is local.

2.2 Attention Model

Attention typically suggests that the human being's optical system focuses on the relevant area[26] and adaptively processes visual information.

Currently, some studies suggested embedding attention mechanisms to improve CNN's performance on various jobs, including image segmentation, multimedia classification [27, 28]. Wang et al. introduced neural network [28] for multimedia classification. This method incorporates non-local processes into remote spatial attention features. Hu et al. SENet [27] introduced a method to achieve the channel-level relationship aiming to improve the image classification performance. The expectation to maximise attention network for semantic segmentation is proposed by Li et al. Utilising the EM algorithm to optimise parameters and reduces the complexity of non-local block operations is suggested by Huang et al. Fu et al. proposes a method for semantic segmentation of crisscross attention [29]. That can effectively capture context patterns from remote dependencies. The Dual Attention Network (DANet)[30] is proposed, consisting mainly of a location attention structure and a channel-attention structure. This utilises the location attention structure to study spatial interdependence. The channel-attention module aims to trace the interdependence of channels. By capturing rich contextual relevance, the results of segmentation are greatly improved. A residual channel attention network (RCAN)[9] for single image super-resolution is proposed by Zhang et al. This method utilises the channel attention to capture channels adaptively by considering the interdependence of information between channels. First, insert non-local blocks into a single super-resolution image. Then they introduced residual non-local attention structure[31] to obtain extra detailed knowledge by retaining relatively low-level features that are proper for super-resolution reconstruction. The design pursues better representation capabilities and delivers high-level image reconstruction results. After that, a non-local residual enhancement group (NLRG)[24] is proposed to achieve spatial context that significantly improves the result of the model.

2.3 Categories for Techniques

[32] divides current methods of modelling neural network architectures into the following groups.

2.3.1 Linear Networks

The linear network has a simple network type, with only one signal direction, no bypass connections or multiple branches. In this network architecture, multiple convolutional layers are superimposed on each other, so the input flows from the initial layer to the node. SRCNN [4], VDSR [11], etc., are usually linear network designs.

2.3.2 Residual Networks

Unlike linear networks, the residual network utilizes network architecture links to skip away to prevent gradient disappearance and make deep networking feasible. First proved its importance in the topic of image classification. Some networks, such as EDSR[7] have recently been using residual learning to increase the efficiency of SR systems. In this process, the residual is learned by the algorithm, which is the high frequency between the input and ground truth. Present residual learning methods are classified toward single-stage or multi-stage, depending on the number of steps utilized in these networks.

2.3.3 Recursive Networks

The recursive structure uses convolutionallayers or recursively linked units that are recursively related. The fundamental purpose behind these designs is to slowly decompose the more complex SR problems into a series of easier-to-resolve, more specific problems.

2.3.4 Progressive Reconstruction

Generally speaking, the CNN algorithm can predict output in one step; however, larger-scale factors may not be capable of this. Many algorithms (such as LapSRN [6]) will predict the performance in multiple stages, i.e. perform two times, then four times, and so on, to deal with broader factors.

2.3.5 Densely Connected

Motivated by the DenseNet's successful image classification, a new algorithm built on tightly connected CNN layers is introduced for performance improvements. The key reason for this design is to integrate the available hierarchical indications along the depth of the network to achieve a high degree of versatility and a better representation of functions.

2.3.6 Multi-branch Networks

Multi-branch networks aim to achieve a complex range of functions on different background scales, in contrast to architectures focused on single-stream (linear) and skip links. Then this additional information is merged to get a better reconstruction of HR. Multi-path signal flow can also be realized by design so that that information can be exchanged in the training process between previous and subsequent stages. Multi-branch design is typical in many other computer vision tasks, too.

2.3.7 Multiple Degradations Network

The hitherto addressed super-resolution network considers bicubic degradation. Nonetheless, this may not be a realistic conclusion in practice, since at the same time, several degradations may occur. A system similar to ZSSR is proposed to tackle this fact.

2.3.8 GAN

The Generative Adversarial Network (GAN) uses an approach to game theory in which the model's two components (namely the generator and the discriminator) attempt to deceive the latter. The generator produces the super-resolution image such that it can not be recognized like a genuine high-resolution image or an artificial super-resolution output by the discriminator. The approach allows for the creation of HR images of better perceptual accuracy.

Chapter 3

Design and Experiment

3.1 Non-local Second-order Attention

3.1.1 Structure Design

I have published this NSAN algorithm in CD-MAKE2020 Conference[33]. As illustrated in the structure in Figure 3.1, this NSAN structure consists primarily of the following parts: the shallow feature extractor, the high-order enhancement group with deep feature extraction(HEG), and the expansion and reconstruction layer. First input I_{LR} and I_{SR} as our NSAN's input value and the output value;Then I applied a convolutional layer after [7, 24] to capture the shallow features from the input

$$F_0 = H_{SF}(I_{LR}) \tag{3.1}$$

where H_{SF} denotes the convolution process. Later the shallow feature F_0 filled in HEG based deep feature extraction thus achieves the deep feature as

$$F_{DF} = H_{HEG}(I_0) \tag{3.2}$$

where H_{HEG} is the HEG based feature extraction structure, which comprised of two RL-NL modules to achieve the long-range information and G residual channel attention groups. In this way, our proposed HEG method could achieve intense depth and could obtain more information. Next, the extracted F_{DF} is going to be upsampled through the upscale module through

$$F_{\uparrow} = H_{\uparrow}(I_{LR}) \tag{3.3}$$

from which $H_{\uparrow}$ and $F_{\uparrow}$ are upsample layer and upsampled feature individually.

There are multiple options in previous works to be done as an upscale procedure, for instance, transposed convolution [5], ESPCN [34].

In recent SR models [5, 24, 7], the integration of upscaling functionality in the last few layers achieves a reasonable trade off between output and computational burden. Then upscaled function is through one layer of convolution

$$I_{SR} = H_R(F_{\uparrow}) = H_{NSAN}(I_{LR}) \tag{3.4}$$

Where HR, $H\uparrow$ and $HNSAN$ are respectively the reconstruction layer, the upsample layer and the NSAN function.

NSAN will then be optimized with a function of losses. Some loss functions were widely used, such as perceptual losses in L2, L1. To verify our NSAN 's effectiveness, I have adopted the L1 loss functions that followed previous works. Because of the training set with low-resolution N images and high-resolution images denoted by $\{IHR, IHR\}^N$, the NSAN aims to optimize the loss function:

$$L(\Theta) = \frac{1}{N}\sum_{i=1}^{N} ||I_{HR} - I_{SR}||_1 \tag{3.5}$$

Where θ represents the NSAN range of parameters. To optimize the loss function, I select Adam algorithm.

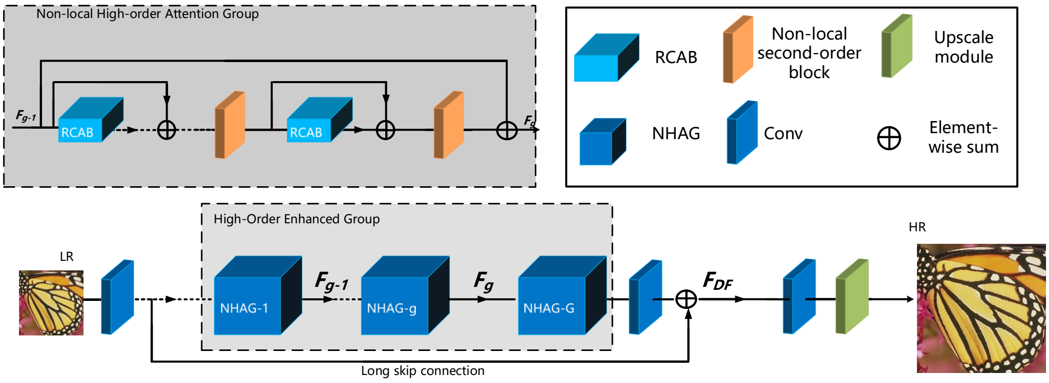


Figure 3.1: The overall structure of the NSAN network design

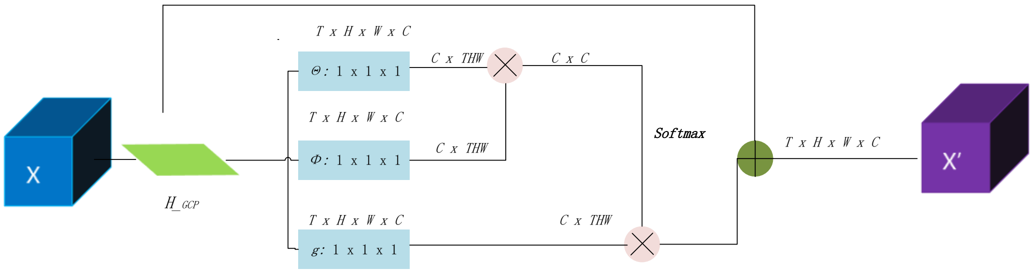


Figure 3.2: The detailed structure of the attention module

3.1.2 Enhanced High-order Group (HEG)

Now I describe edge HEG module (see Figure 3.1), which could be divided into branch enhancement and edge enhancement of the main branch. The main branch is comprised of 2 regional levels [24] and G structure of non-local (RL-NL) residual channel attention group (NRCAG) modules. RL-NL may collect information from distant locations. Each NRCAG also includes simplified residual M channel blocks with local skip connections followed by a non-local channel attention module (NCA). The edge enhancement branch consists of padding module and V. The NRCAG can take full advantage of edge information and use edge information to improve channel features.

The method for stacking residual blocks has been verified as useful. It creates in [7, 23, 24] a deep network. Even so, deeper networking built in this way will cause performance gaps and difficulty training. When problems arise, the gradients disappear and explode in deep network.

Stacking repeated blocks, as we all know, may simply not provide better performance. I have introduced NHAG to solve this problem, and also the large amount of low-frequency LR image information contributes to our deep network training.

Then a HEG is represented in the group g-th as:

$$F_g = H_g(F_{g-1}) \tag{3.6}$$

Where Fg, $Fg - 1$ denotes a g-th HEG output and input. For simplicity the term of the bias is omitted. The function g-th HEG is Hg.

Deep feature is then obtained as:

$$F_{DF} = F_0 F_G \tag{3.7}$$

3.1.3 Non-local Second-order Attention

Most previous CNN-based SR models overlooked interdependence between functions. To use this information to the full, SENet parencite hu2018squeeze introduced CNN to rescale the image SR function at channel level. However, SEnet only uses that First-order statistics are realised through the global average pool, and non-local statistics are ignored which are richer than local statistics, hampering the network's discriminative ability.

Inspired by the above work, I proposed a non-local second-order attention (NSA) module to capture the interdependence of higher-order characteristics by including non-local characteristics (see Figure 3.2). I remodelled the feature map $\mathrm{F} = [\mathrm{f}_1, \cdots, \mathrm{f}_C]$ with C feature maps with size of $H \times W$ to a feature matrix X with $s = WH$ features of C-dimension. Then compute the sample covariance matrix as

$$\Sigma = \mathrm{X}\overline{I}\mathrm{X}^T$$

where $\overline{I} = \frac{1}{s}(\mathrm{I} - \frac{1}{s}1)$, I and 1 are the $s \times s$ identity matrix and manix of all ones, respectively.

Covariance normalisation plays a vital role with more biassed representations. For this function, I first perform a covariance normalisation for the covariance matrix Σ obtained, which is a symmetrical positive semi-definite and thus has an own value decomposition (EIG) as follows.

$$\Sigma = \mathrm{U}\Lambda\mathrm{U}^T$$

where U is an orthogonal matrix and $\Lambda = \mathrm{diag}(\lambda_1, \cdots, \lambda_C)$ is diagonal matrix with eigenvalues in non-increasing order. Then convariance normalization can be converted to the power of eigenvalues:

$$\hat{\mathrm{Y}} = \Sigma^\alpha = \mathrm{U}\Lambda^\alpha\mathrm{U}^T$$

where α is a positive real number, and $\Lambda^{\alpha} = \mathrm{diag}(\lambda_1^{\alpha},\ \cdots,\ \lambda_C^{\alpha})$. When $\alpha = 1$, there is no normalization; when $\alpha < 1$, it nonlinearly shrinks the eigenvalues larger than 1.0 and streches those less than 1.0.

The normalized covariance manix characterizes the correlations of channel-wise features. I then take such normalized covariance matrix as a channel descriptor by global covariance pooling. As illustrated in Fig. 2, let $\hat{\mathrm{Y}} = [\mathrm{y}_1,\ \cdots,\ \mathrm{y}_C]$, the channel-wise statistics $\mathrm{z} \in R^{C\times 1}$ can be obtained by shrinking $\hat{\mathrm{Y}}$. Then the c-th dimension of z is computed as

$$z_c = H_{GCP}(\mathrm{y}_c) = \frac{1}{C}\sum_{i}^{C}\mathrm{y}_c(i)$$

Where $HGCP$ represents the global covariance pooling function. Compared to the widely employed first-order pooling (*e.g.*, regional average pooling), our regional covariance pooling examines the distribution of features and collects higher-than-first-order figures with more unequal representations.

To better leverage the interdependencies of the aggregated information by global covariance pooling, I also add a non-local block to catch long-range trends. Inspired by the Dual Attention Network introduced by Fu et al., which implemented a non-local channel focus block to gain more valuable functionality, but did not include a second-order feature. In our network, I pair the second order extractor with the non-local attention, which will extract the second order function and then generate a non-local attention map.

$$\mathrm{X}' = H_{NSB}(z)$$

where H_{NSB} denote the non-local second-order block.

3.2 ARSN - light-weight Network with Automatic Residual Scaling

Though NSAN has excellent performance on latest mobile phones with powerful GPUs, the hardware of standard smart terminals is not suitable for large-scale deep neural network models proposed. To make up for its shortcomings, I propose a light-weight Network with Automatic Residual Scaling (ARSN). The innovation of this approach is to achieve fewer layers and lighter weights of the method without sacrificing much accuracy. At the same time, this method can directly input low-resolution images without bi-cubic interpolation, which could reduce additional calculations.

3.2.1 Architecture

The basic structure of ARSN is shown in Figure 3.3. The model contains several special residual blocks for feature extraction. The number of residual blocks can be increased or decreased according to the actual situation. Compared with many different deep learning-based algorithms, this structure has fewer layers and parameters. Besides,the specified residual blocks and skip connections in this network were utilized for residual scaling, global and local residual learning. This results on test datasets prove that this model achieves balanced performance on both reconstruction quality and operating speed. The proposed network achieves good results in terms of performance, speed and hardware consumption, and has high practical value.

3.2.2 Residual module

As shown in Figure 3.4, each residual block contains two convolutional layers, plus each convolutional layer contains 64 convolution kernels including a kernel size of 3*3 to maintain the ratio of the output map.

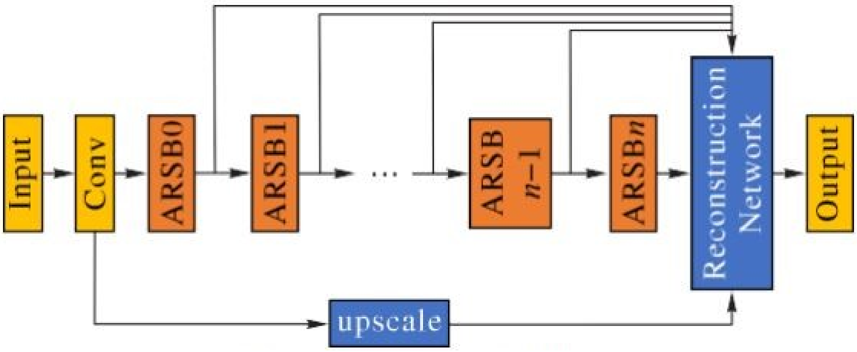


Figure 3.3: Structure of ARSN.

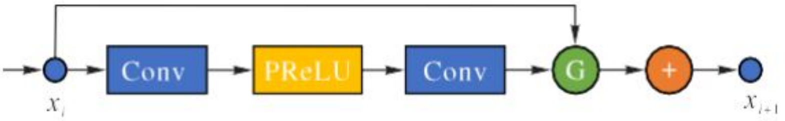


Figure 3.4: Structure of a specified residual block.

Unlike the residual blocks in the standard residual network and SRresNet, the residual blocks in the design proposed in this dissertation will delete the useless batch normalization layer. Szegedy et al. found that adding feature maps would make the model training unstable. Therefore, according to the methods of Szegedy et al. and Lim et al., the model proposed in this paper adds residual gating following the second convolutional layer of the residual block.

3.2.3 Structure of reconstruction network

Global residual learning is utilized in the reconstruction of the network, which may cause the main branches of the network to learn or predict the details of the image. ARSN also uses subpixel shuffle to enlarge low-resolution images. Experiments have found that global residuals can converge faster and produce better quality than simple deconvolution layers or bicubic interpolation. The main reconstruction process is shown in Figure 3.5

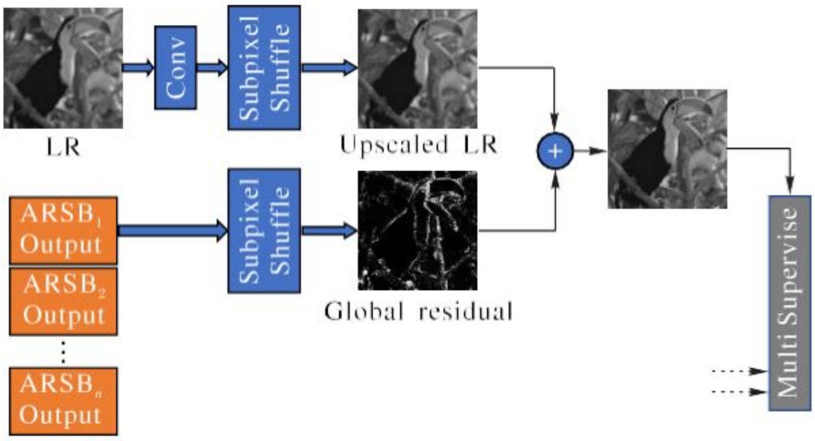


Figure 3.5: Structure of a specified residual block.

3.3 Comparing Approaches

Here I extensively compare the following most-advanced super-resolution Convolutional Neural Networks together with my methods to benchmark single image super-resolution.

I also compare the following models by network size, input and output model, learning information, technical variations, operating speed, and peak signal-to-noise ratio.

3.3.1 SRCNN

The SRCNN[4](shown in Figure3.6) is the first successful model to achieve super-resolution using convolutional layers only. This work can be seen as a ground-breaking work focused on deep learning SR, which has inspired several subsequent attempts in this direction. The layout of the SRCNN is plain and simple. This consists only of convolutional layers where a ReLU unit accompanies the nonlinearity of the individual layer (except the last layer).

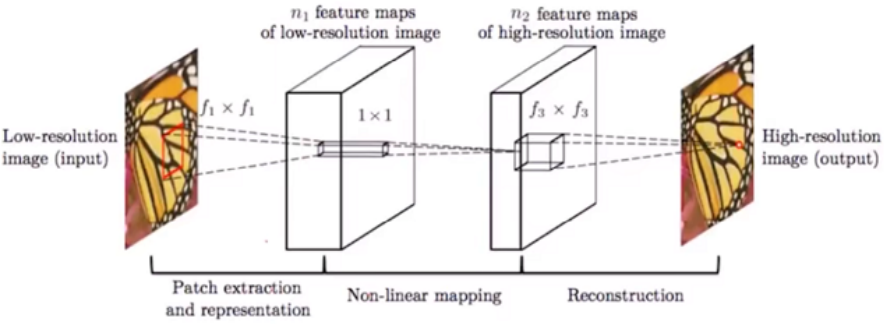

n_1 feature maps of low-resolution image
n_2 feature maps of high-resolution image
Low-resolution image (input)
$f_1 \times f_1$
1×1
$f_3 \times f_3$
High-resolution image (output)
Patch extraction and representation
Non-linear mapping
Reconstruction

Figure 3.6: Structure of SRCNN

3.3.2 FSRCNN

FSRCNN[5] (shown in Figure3.7) has improved speed and accuracy as compared to SRCNN [4]. It aims to achieve a real-time calculation rate (24 fps) compared to SRCNN (1.3 fps).

Compared with SRCNN, the model does not need to preprocess the input for interpolation, directly conduct model training, and then perform up-sampling in the last deconvolution; then use 1x1 convolution in the middle of the model layer to perform dimensionality increase and dimensionality reduction , To further reduce the amount of model parameters.

The PReLU used by the nonlinear activation function in the middle of the model is used to avoid the dead zone problem of ReLU during the training process, and the loss function is still the MSE used.

3.3.3 VDSR

SRCNN[4]has three problems that need to be improved: 1. It depends on the content of the small image area; 2. The training convergence is too slow; 3. The network is only effective for a certain ratio.

The VDSR[11](Figure3.8) model mainly has the following contributions: 1. It increases the receptive field and has advantages in processing large images, from 13*13

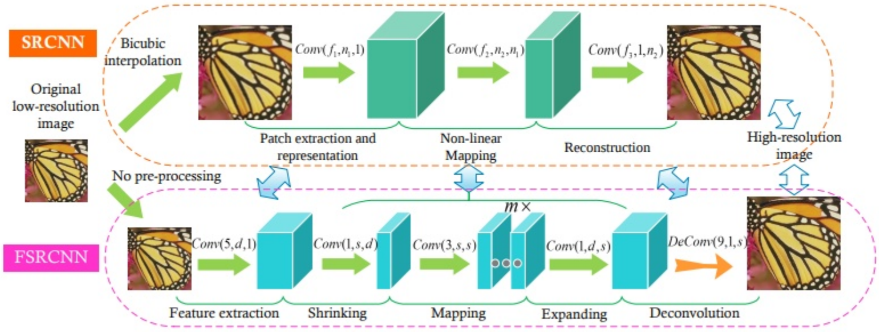


Figure 3.7: Structure of FSRCNN

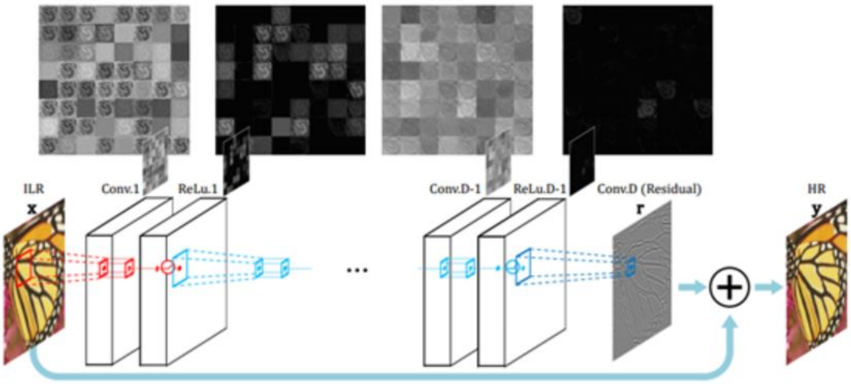


Figure 3.8: Structure of VDSR

of SRCNN to 41*41. 2. Using residual images for training, the convergence speed becomes faster. Because the residual image is more sparse and easier to converge (another understanding is the low-frequency information of the LR carrier, this information is still trained to the hr image, but the low-frequency information of the HR image and the LR image are similar, which takes a lot of time to train) . 3. Considering multiple scales, a convolutional network can handle multi-scale problems.

3.3.4 LAPSRN

Deep Laplacian Pyramid Super Resolution Network (LapSRN)[6] uses a structure with pyramids.The LapSRN network is made up of three component types, namely a convo-

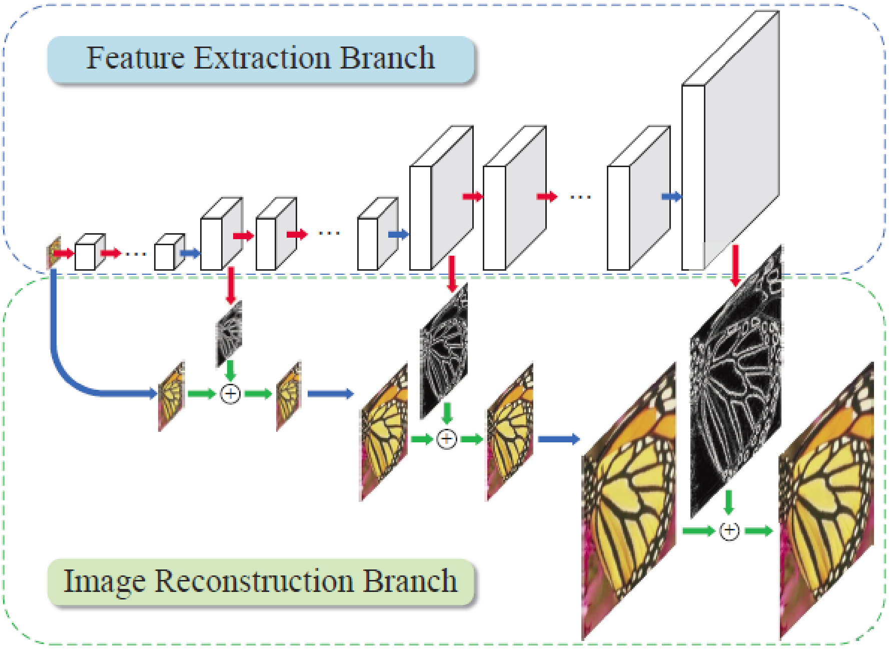


Figure 3.9: Structure of LAPSRN

lutional layer, a ReLU leakage, and a deconvolutional layer. Following the CNN rule, the convolutional layer is placed at the end of the sub-layer before the leaky ReLU (allowing a negative 0.2 slope) and before the deconvolutional layer to increase the residual image size to the corresponding proportion.

LapSRN uses a ’1 loss function differential variant named Charbonnier that can accommodate outliers. In each subnet, losses are recognized, similar to a multi-loss structure. In comparison, the philtre sizes of the layer convolution and deconvolution are 3 ranges and 4 ranges, each with 64 channels.

The LapSRN software uses 3 different 2x, 4x, and 8x SR models. We also suggested a single model that could learn together, called LapSRN Multi-Scale (MS). Hold multiple SR rates. Ironically, a single MSLapSRN model’s performance is higher than that of the three different models. One reason for this effect is that specific interscale features are used by a single model to help produce more accurate results.

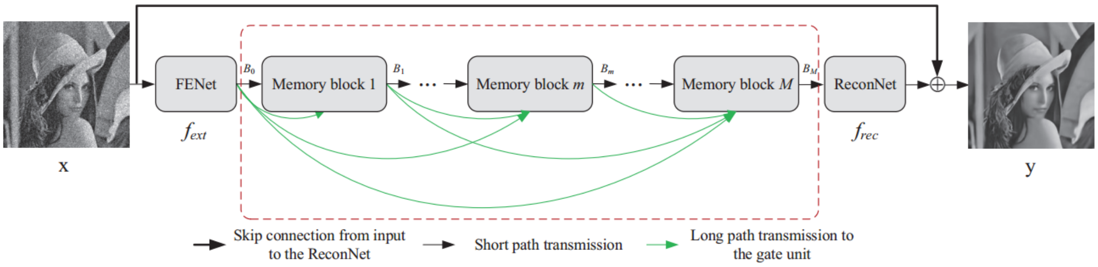


Figure 2. Basic MemNet architecture. The red dashed box represents multiple stacked memory blocks.

Figure 3.10: Structure of MemNet

3.3.5 MemNet

Tai et el.[13] presents a novel, persistent memory network for image super-solution (MemNet)(Figure 3.10).Traditional neural networks are basically one-way propagation, so in the lower layer, the received signal is very weak. This one-way propagation network, such as VDSR, DRCN, etc., is called short-term memory network. RED, ResNET, the neurons in the network are not only affected by the direct predecessor, but also by the heroes of the additional designated predecessor neurons. This is called a restricted long-term memory network.

The long-term memory model has the following 3 special features:

1. The memory unit uses the gate unit to establish long-term memory. In each memory unit, the gate unit adaptively controls the weights of different blocks in the final output and controls which ones to keep Unit, what information is stored.

2. Deep network (80 layers), dense connection structure (as can be seen from the figure above), signal compensation mechanism (the neurons in the back are directly connected by the neurons in the front), which maximizes the information in different Flow between memory cells.

3. The structure is proven to have a strong learning ability, and one model handles multiple tasks (the model is used for image restoration, denoising, and super-resolution)

3.3.6 EDSR

Enhanced Deep Ultra High Resolution (EDSR)[7] updated ResNet's originally proposed architecture for image classification to be used in SR tasks. Specifically, they showed substantial improvement by removing the batch normalization layer (from each residual block) and activating ReLU (outside of the residual block). Like VDSR, their single-scale method to work on multiple scales has also been extended. Their proposed multi-scale depth architecture for SR (MDSR) reduces the number of parameters by most shared parameters. To learn about scale-related representations, only near the input and output blocks are applied in parallel scale-specific layers.

The proposed deep architecture uses 1 loss for the training. Data enhancement (rotation and flip) is used to create ”self-integration,” i.e. the transformed input is passed through the network, reverse-transformed and averaged together to create a single output. The author points out that this system of self-integration does not need to learn several separate models but can bring comparable advantages to conventional models based on integration.

3.3.7 SRMD

Super-resolution multidegradation network (SRMD)[35] shoots cascaded low-resolution images and their maps of degradation. First, a cascaded, three-part filter size convolution layer is applied to the extracted features, and then a series of normalization layers for Conv, ReLU and Batch. In addition, the convolution operation is used to extract the HR sub-images, and the last step is to convert multiple HR sub-images into a single final HR output. SRMD learns the HR image directly, and not the image residual. The author also introduced a variant called SRMDNF that learns from degradation without noise. In the SRMDNF network the connection is removed in the convolutional layer from the first noise level mapping; however, the rest of the architecture is similar to SRMD. The author trained a separate model for each upsampling scale, compared to multi-scale training. The number of convolutional layers is fixed at 12, with 128 feature maps for each layer. The initial learning is set at 10-3, then lowered to 10-5. The standard for lowering the learning rate between two consecutive times is based on the era of error changes.

3.3.8 DBPN

DBPN (Deep Back-Projection Network)[36] is the winner of the PIRM Super Resolution Competition 2018. The innovation is that the modules combined with upsampling and downsampling use stacking in residual mode which can take advantage of dependence on upsampling and downsampling. It makes use of iterative upper and lower layers of sampling to provide an error feedback mechanism for the projection error of each stage. Create interconnected up-sampling and down-sampling stages, each stage represents a different form of image degradation and components with a high resolution. By extending this idea to allow for cascading features in the up- and down-sampling phase (dense DBPN), the results can be further improved.

3.3.9 RDN

It is proposed that the Residual Dense Block (RDB)[37] extract rich local characteristics through densely connected layers of convolution.

RDB also allows direct connections to all current RDB layers from the previous RDB state , resulting in a Continuous Memory (CM) mechanism. In RDB, the local feature fusion is then used to adjust and learn more successful features from past and current local features, and to improve the training of a broader network.

International function fusion is used in a holistic way, jointly and adaptatively to learn different hierarchical features.

3.3.10 RCAN

Residual Channel Attention Network (RCAN) [9] is a recently proposed deep CNN architecture for single image super-resolution.

The first novelty of this structure allows multiple paths from initial to final information level. The second contribution allows the network to focus on selective feature maps that are more important to the final task, and it can also effectively model the relationship between feature maps.

RCAN uses 1 loss function network training. It has been observed that the recursive residual style architecture can lead to very deep networks with better convergence. In addition, it has better performance than modern methods such as VDSR and RDN. This illustrates the influence of the guided attention mechanism on low vision tasks. However, one disadvantage of the proposed framework is its high computational complexity compared with other frameworks.

3.3.11 SAN

The second-order attention network (SAN) [24] aims to provide more powerful expression of function-related functions and learning. In this structure, a novel trainable second-order channel attention (SOCA) module has been developed to better distinguish representations to adapt to channel direction features by using second-order feature statistics. In addition, the non-local enhanced residual group (NLRG) structure not only contains non-local operations for capturing remote spatial context information, but also contains repeated local source residual attention groups (LSRAG) for learning increasingly abstract features.

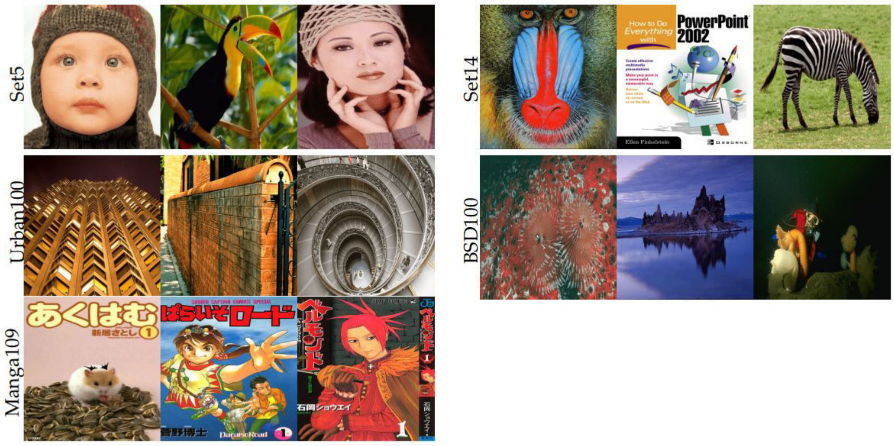
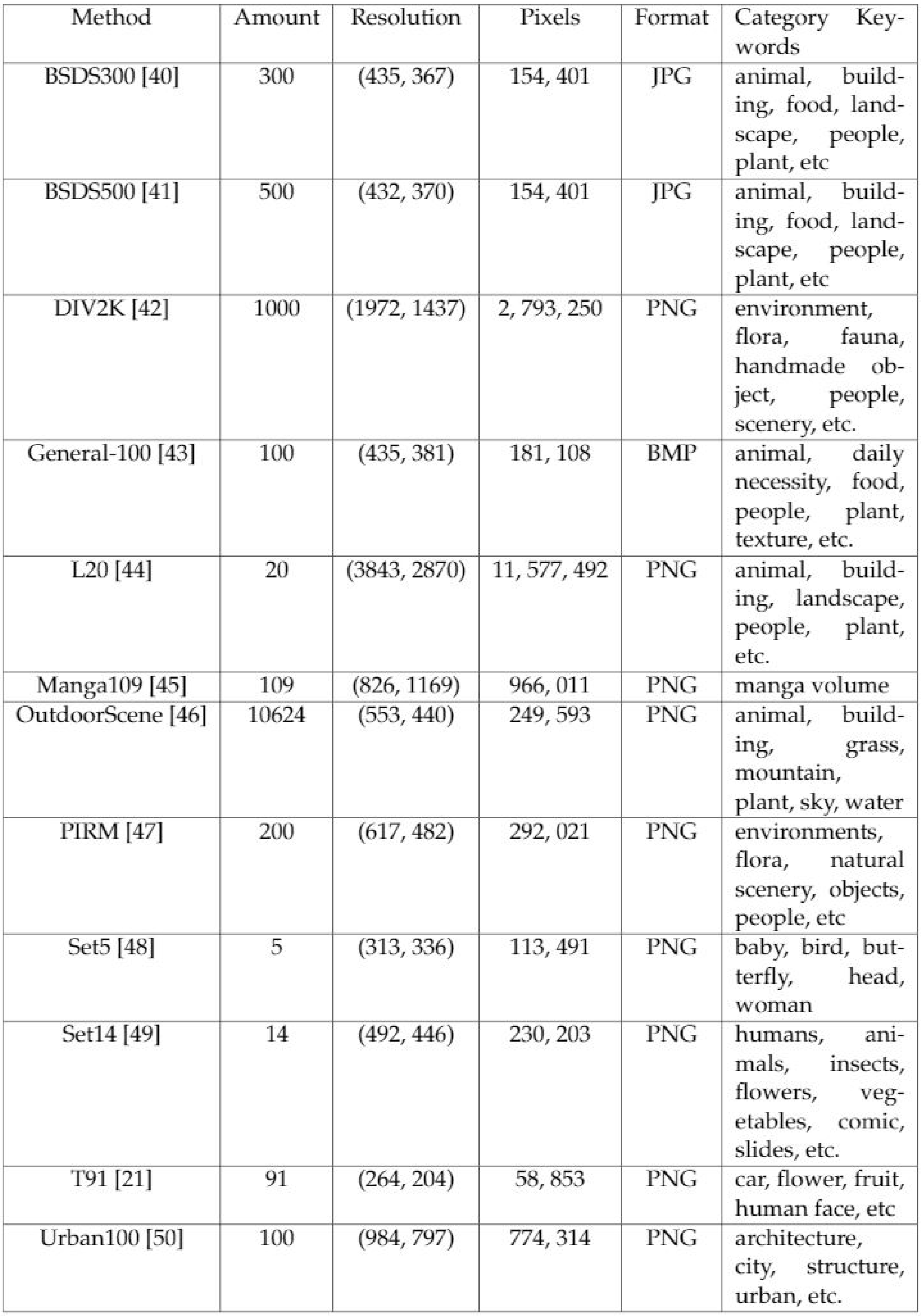


Figure 3.11: Super-resolution data set utilized in this study

3.4 Datasets

A number of data sets are now available that can be used for super-resolution images and have major differences in image quantity, accuracy, quality, and variety. Some have pairs of LR-HR images while others only send HR images. Typically the default imresize feature (i.e., bicubic anti-aliasing interpolation) is used to obtain LR images in this case). Numerous image data sets widely used in the SR community are listed in the table below, specifically their HR image number, average size, average number of pixels, image format and category keywords. In addition to these data sets, some data sets for other vision tasks are commonly used in SR such as ImageNet, MS-COCO. In addition, multiple training data sets, such as combining T91 and BSDS300 or combining DIV2K and Flickr2K, are also common. Finally, as shown below, we selected the five most common data sets of all of those data sets:



Method	Amount	Resolution	Pixels	Format	Category Keywords
BSDS300 [40]	300	(435, 367)	154, 401	JPG	animal, building, food, landscape, people, plant, etc
BSDS500 [41]	500	(432, 370)	154, 401	JPG	animal, building, food, landscape, people, plant, etc
DIV2K [42]	1000	(1972, 1437)	2, 793, 250	PNG	environment, flora, fauna, handmade object, people, scenery, etc.
General-100 [43]	100	(435, 381)	181, 108	BMP	animal, daily necessity, food, people, plant, texture, etc.
L20 [44]	20	(3843, 2870)	11, 577, 492	PNG	animal, building, landscape, people, plant, etc.
Manga109 [45]	109	(826, 1169)	966, 011	PNG	manga volume
OutdoorScene [46]	10624	(553, 440)	249, 593	PNG	animal, building, grass, mountain, plant, sky, water
PIRM [47]	200	(617, 482)	292, 021	PNG	environments, flora, natural scenery, objects, people, etc
Set5 [48]	5	(313, 336)	113, 491	PNG	baby, bird, butterfly, head, woman
Set14 [49]	14	(492, 446)	230, 203	PNG	humans, animals, insects, flowers, vegetables, comic, slides, etc.
T91 [21]	91	(264, 204)	58, 853	PNG	car, flower, fruit, human face, etc
Urban100 [50]	100	(984, 797)	774, 314	PNG	architecture, city, structure, urban, etc.

Table 3.1: Most widely used public image datasets for SISR

3.5 Image Quality Assessment

Image quality refers to the visual qualities of the image, and focuses on assessing the expectations of viewers. Methods for measuring image quality (IQA) typically include subjective methods based on human interpretation (i.e., how the picture looks real) and methods for objective measurement.

The former is more in line with our needs but is usually time-consuming and costly, which is why the latter is currently the norm. These methods do not necessarily coincide, however, since the target method is Normally human visual experience can't be very well described Precisely this may cause major variations in IQA outcomes.

In addition, objective IQA methods are further broken down into three types: complete reference assessment methods using reference images, simplified reference Methods to compare the extracted products, and No reference system, with no reference image. These are some of the most commonly used IQA methods which entail both subjective and objective methods.

3.5.1 PSNR

The PSNR is calculated as follows:

$$PSNR(x, y) = \frac{10 \log_{10}[\max(\max(x), \max(y))]^2}{x - y^2} \tag{3.8}$$

PSNR is one of the most common quality restoration metrics for loss transformations (such as compression of images and restore of images). For super-resolution image, the maximum pixel value (referred to as L) and the mean square error (MSE) between the image are specified as PSNR.

Since PSNR is connected only to pixel-level MSE, it focuses only on the difference between the corresponding pixels rather than visual perception. Thus, when representing the quality of reconstruction in a real scene, this typically leads to deterioration of results, and we typically pay more attention to human perception. However, due to the need for comparison with literary works, and the lack of an appropriate comparison PSNR is now the most commonly used method of assessment of SR models in terms of perception measures.

3.5.2 Structural Similarity

The formula of SSIM is defined as below:

$$(x, y) = \frac{(2\mu_x\mu_y + C_1) + (2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{3.9}$$

Given the human visual system (HVS) is very useful for capturing image structure, a structural similarity index (SSIM) is proposed to measure the structural resemblance between images based on the independent colour, contrast and form analysis.

3.6 Experiments

To check our network 's efficacy, I pick 5 datasets of benchmarks: Set5, Set14, BSD100, Urban100 and Manga109. I'm implementing resize feature with bicubic activity for the degradation model. I use PSNR and SSIM for assessing SR performance for the metrics.

The low-resolution images are increased for the NSAN training by horizontally tossing and arbitrarily spinning $90°, 180°, 270°$. I set 16 low-resolution image patches for each min-batch, with size 48×48as input. I use the ADAM algorithm to simplify our model with $\beta 1 = 0.9, \beta 2 = 0.99$,and $\epsilon = 10^{-8}$ and initialize learning speeds as 10^{-4} and then minimize them to half after 200 epochs. To train our proposed NSAN on an Nvidia 2080Ti GPU, I use the Pytorch system.

Due to time limit, I set the initial learning rate at 0.001 for ARSN training, and it will decrease after 60 epochs. The experiment also made use of the Adam optimizer to minimize the role of failure and equipped 120 epochs. The original residual scale factor is set to 0.25 and will be automatically changed as the model runs. Momentum is 0.9 and weight attenuation is 0.001. Place 256 pictures as mini batch for model entry.

Chapter 4

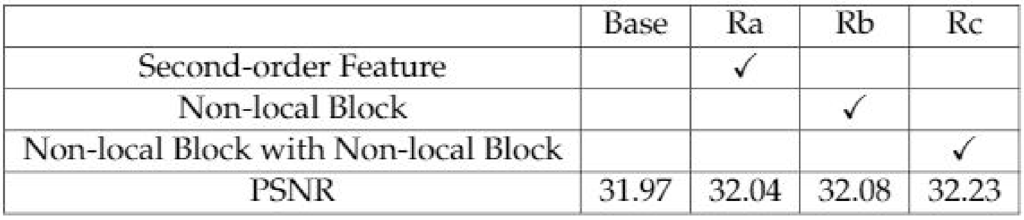
Results

4.1 Ablation Study

Our NSAN consists of two main components, including the high-order enhancement group (HEG) and non-local second-order attention module (NSA), as shown in Figure 3.1. Of comparative purposes I practiced and evaluated NSAN and its derivatives on the Set5 dataset to check the efficacy of specific modules. The specific performance is reflected in Table 4.1.

I set R_{BASE} as the standard baseline, comprising only the convolution layer containing 20 NHAGs and the remaining 10 blocks in each NHAG. I have introduced long jump and short jump connections to the basic model, after [38]. R_a and R_b show that the second-order extractor and the non-local component are contained in the basic structure, respectively. R_c represents a combination of the second-order extractor and non-local block function. It can be seen that the R_c output is higher than the R_a to R_b process.

Table 4.1: Effect of different module for Set5 dataset with 200 epoch



	Base	Ra	Rb	Rc
Second-order Feature		✓		
Non-local Block			✓	
Non-local Block with Non-local Block				✓
PSNR	31.97	32.04	32.08	32.23

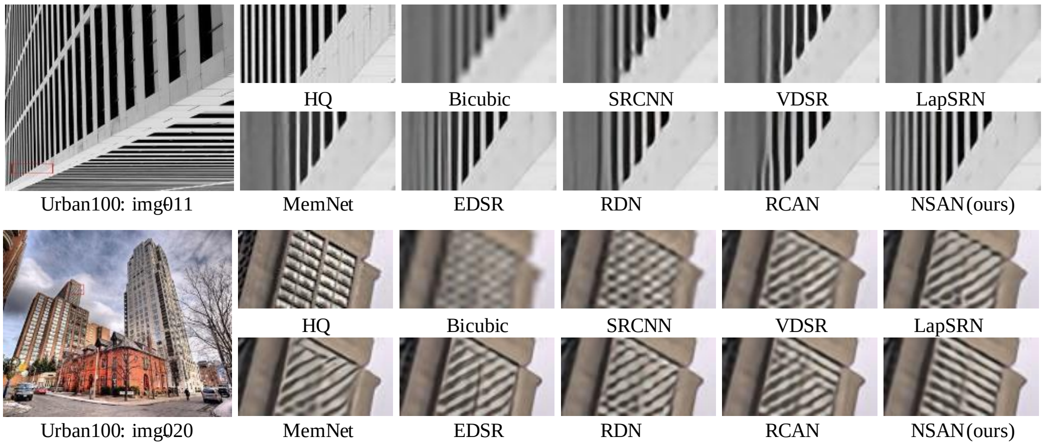
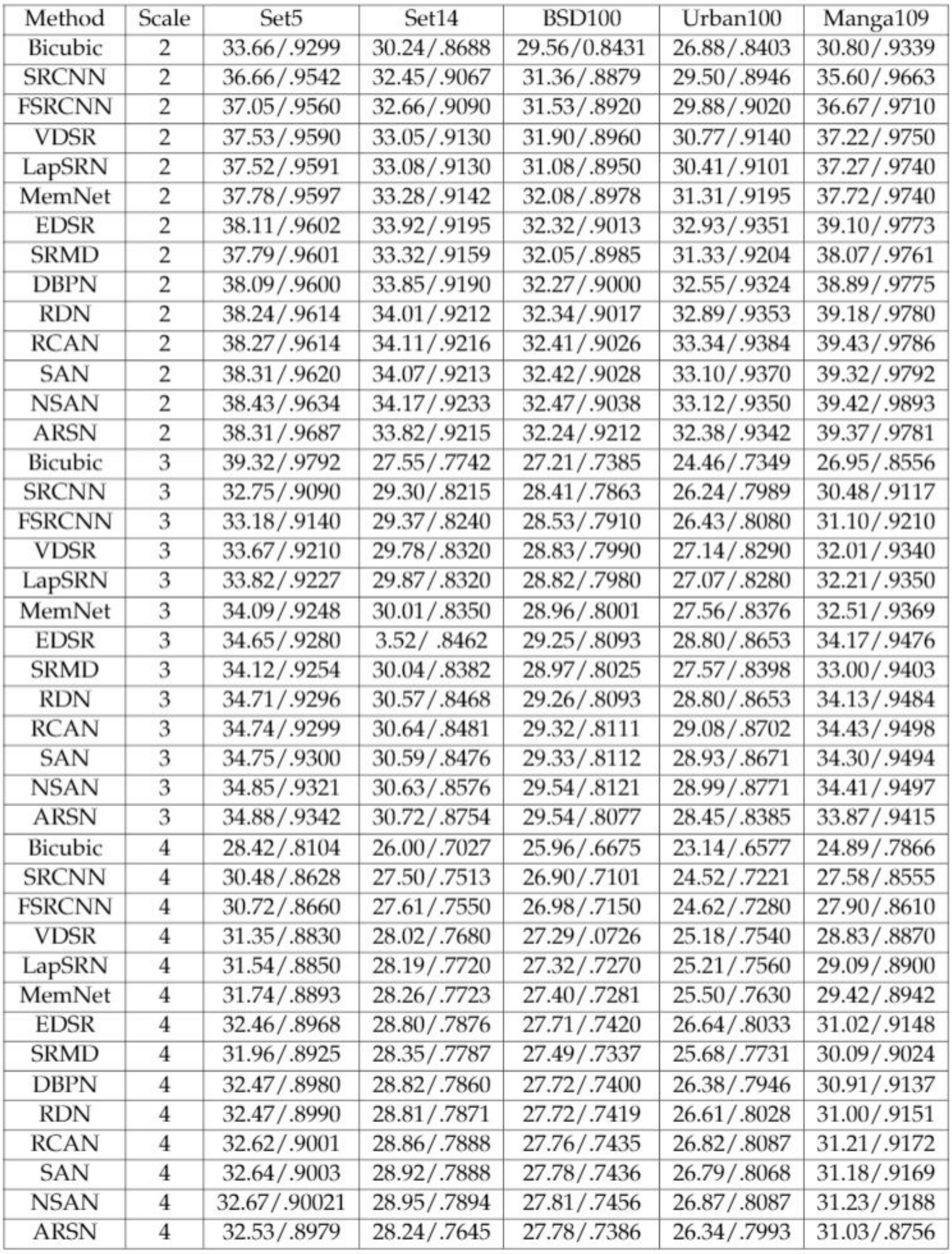


Figure 4.1: Visual Result for scale factor 2

Table 4.2: Quantitative results(PSNR/SSIM)



Method	Scale	Set5	Set14	BSD100	Urban100	Manga109
Bicubic	2	33.66/.9299	30.24/.8688	29.56/0.8431	26.88/.8403	30.80/.9339
SRCNN	2	36.66/.9542	32.45/.9067	31.36/.8879	29.50/.8946	35.60/.9663
FSRCNN	2	37.05/.9560	32.66/.9090	31.53/.8920	29.88/.9020	36.67/.9710
VDSR	2	37.53/.9590	33.05/.9130	31.90/.8960	30.77/.9140	37.22/.9750
LapSRN	2	37.52/.9591	33.08/.9130	31.08/.8950	30.41/.9101	37.27/.9740
MemNet	2	37.78/.9597	33.28/.9142	32.08/.8978	31.31/.9195	37.72/.9740
EDSR	2	38.11/.9602	33.92/.9195	32.32/.9013	32.93/.9351	39.10/.9773
SRMD	2	37.79/.9601	33.32/.9159	32.05/.8985	31.33/.9204	38.07/.9761
DBPN	2	38.09/.9600	33.85/.9190	32.27/.9000	32.55/.9324	38.89/.9775
RDN	2	38.24/.9614	34.01/.9212	32.34/.9017	32.89/.9353	39.18/.9780
RCAN	2	38.27/.9614	34.11/.9216	32.41/.9026	33.34/.9384	39.43/.9786
SAN	2	38.31/.9620	34.07/.9213	32.42/.9028	33.10/.9370	39.32/.9792
NSAN	2	38.43/.9634	34.17/.9233	32.47/.9038	33.12/.9350	39.42/.9893
ARSN	2	38.31/.9687	33.82/.9215	32.24/.9212	32.38/.9342	39.37/.9781
Bicubic	3	39.32/.9792	27.55/.7742	27.21/.7385	24.46/.7349	26.95/.8556
SRCNN	3	32.75/.9090	29.30/.8215	28.41/.7863	26.24/.7989	30.48/.9117
FSRCNN	3	33.18/.9140	29.37/.8240	28.53/.7910	26.43/.8080	31.10/.9210
VDSR	3	33.67/.9210	29.78/.8320	28.83/.7990	27.14/.8290	32.01/.9340
LapSRN	3	33.82/.9227	29.87/.8320	28.82/.7980	27.07/.8280	32.21/.9350
MemNet	3	34.09/.9248	30.01/.8350	28.96/.8001	27.56/.8376	32.51/.9369
EDSR	3	34.65/.9280	3.52/ .8462	29.25/.8093	28.80/.8653	34.17/.9476
SRMD	3	34.12/.9254	30.04/.8382	28.97/.8025	27.57/.8398	33.00/.9403
RDN	3	34.71/.9296	30.57/.8468	29.26/.8093	28.80/.8653	34.13/.9484
RCAN	3	34.74/.9299	30.64/.8481	29.32/.8111	29.08/.8702	34.43/.9498
SAN	3	34.75/.9300	30.59/.8476	29.33/.8112	28.93/.8671	34.30/.9494
NSAN	3	34.85/.9321	30.63/.8576	29.54/.8121	28.99/.8771	34.41/.9497
ARSN	3	34.88/.9342	30.72/.8754	29.54/.8077	28.45/.8385	33.87/.9415
Bicubic	4	28.42/.8104	26.00/.7027	25.96/.6675	23.14/.6577	24.89/.7866
SRCNN	4	30.48/.8628	27.50/.7513	26.90/.7101	24.52/.7221	27.58/.8555
FSRCNN	4	30.72/.8660	27.61/.7550	26.98/.7150	24.62/.7280	27.90/.8610
VDSR	4	31.35/.8830	28.02/.7680	27.29/.0726	25.18/.7540	28.83/.8870
LapSRN	4	31.54/.8850	28.19/.7720	27.32/.7270	25.21/.7560	29.09/.8900
MemNet	4	31.74/.8893	28.26/.7723	27.40/.7281	25.50/.7630	29.42/.8942
EDSR	4	32.46/.8968	28.80/.7876	27.71/.7420	26.64/.8033	31.02/.9148
SRMD	4	31.96/.8925	28.35/.7787	27.49/.7337	25.68/.7731	30.09/.9024
DBPN	4	32.47/.8980	28.82/.7860	27.72/.7400	26.38/.7946	30.91/.9137
RDN	4	32.47/.8990	28.81/.7871	27.72/.7419	26.61/.8028	31.00/.9151
RCAN	4	32.62/.9001	28.86/.7888	27.76/.7435	26.82/.8087	31.21/.9172
SAN	4	32.64/.9003	28.92/.7888	27.78/.7436	26.79/.8068	31.18/.9169
NSAN	4	32.67/.90021	28.95/.7894	27.81/.7456	26.87/.8087	31.23/.9188
ARSN	4	32.53/.8979	28.24/.7645	27.78/.7386	26.34/.7993	31.03/.8756

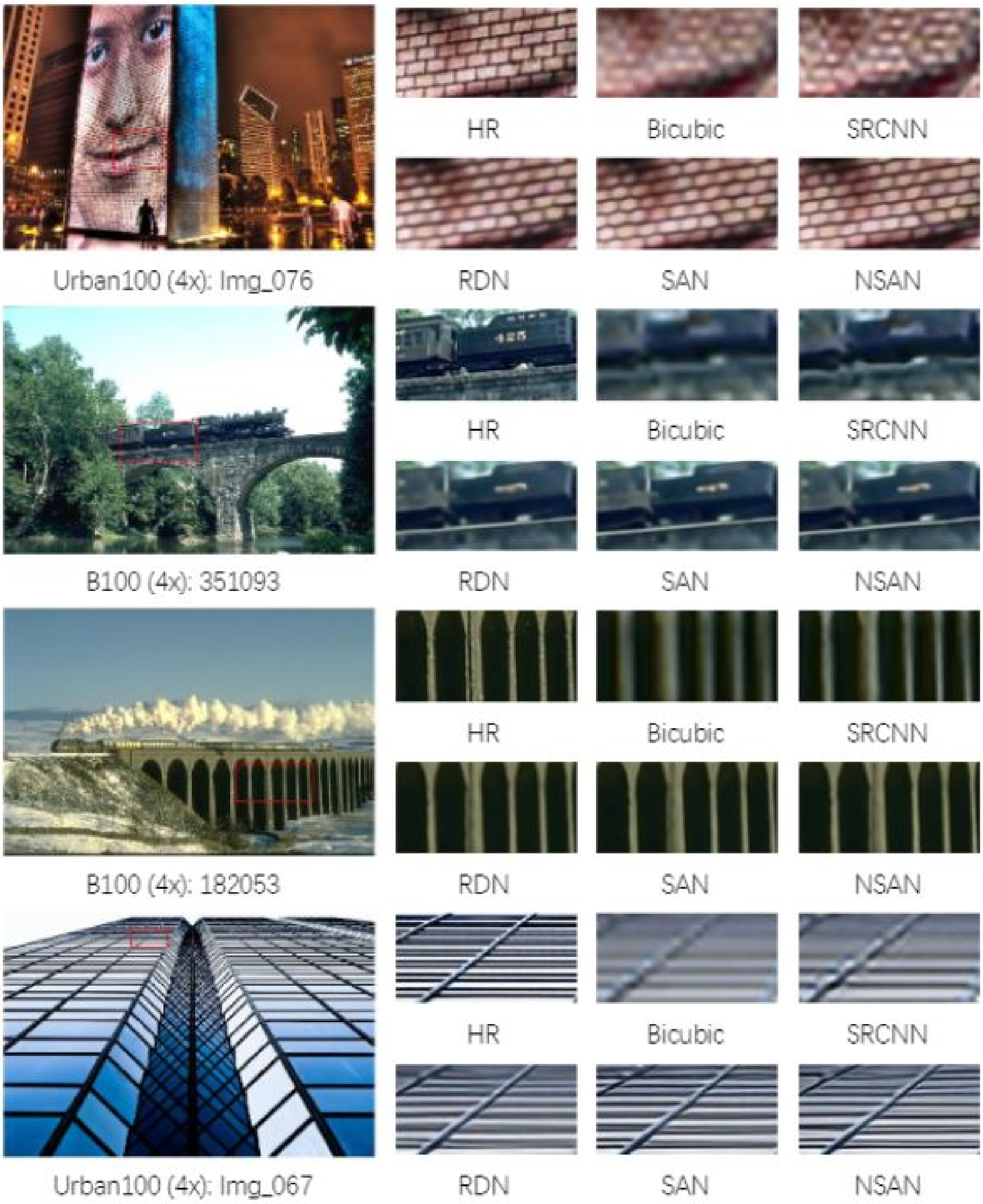
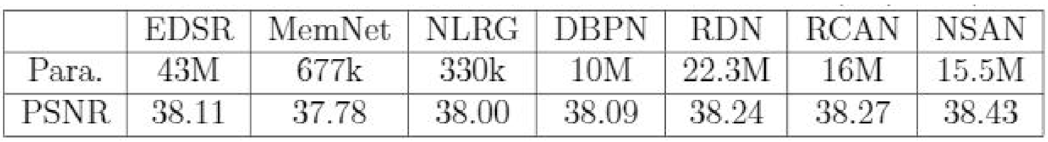


Figure 4.2: Visual Result for scale factor 2



	EDSR	MemNet	NLRG	DBPN	RDN	RCAN	NSAN
Para.	43M	677k	330k	10M	22.3M	16M	15.5M
PSNR	38.11	37.78	38.00	38.09	38.24	38.27	38.43

Table 4.3: Computational and parameter comparison (2X) Set5).

4.1.1 Results

I set up a comparative test with model 12 state-of-the-art CNN-based SR methods: SRCNN [4], FSRCNN [5], VDSR [11], LapSRN [6], MemNet [13], EDSR [?], RDN [38], and RCAN [9] to verify the effectiveness of NSAN. See Table 4.2 for the detailed findings of each scale element. Our NSAN worked better on all datasets as opposed to other models, with different scaling factors. NSAN and SAN can produce very close outcomes without self-integration, and are superior to other methods. This is mostly because they also use high-order features to understand the interdependence between users, which lets the network pay more attention to users of the details.

Our NSAN performed satisfactorily on data sets with rich texture information, such as Set5, Set14, and BSD100 relative to RCAN, and marginally worse results for data sets, such as Manga109 and BSD100 with rich reprocessing edge details. As we all know, layer is a higher-order pattern with more complicated statistical properties, while edge is a first-order pattern only a first-order operator can remove. Based on second-order attribute statistics and non-local operator, our NSA therefore performs best on images with higher-order details like texture.

I also display the visual effects of the various approaches as seen in Figure 4.1. I note that most SR models are unable to reproduce the lattices correctly, and have extreme fuzzy artifacts. Our NSAN, on the opposite, shows better performance and reconstructs more high-frequency information including high contrast and rough edges. Most of the comparative methods produce highly fuzzy objects in the case of ”img011” and ”img076” Bicubic, SRCNN, FSRCNN and LapSRN ’s early inventions have lost their principal architectures.

Compared to the ground-truth, NSAN obtains more accurate results and restores more information in the picture. Although recreating high-frequency information is difficult during LR 's restricted input information, our NSAN can still take full advantage of the restricted LR information by non-local second-order observation, thus taking advantage of the spatial function of both high-order characteristics correlated with more efficient pattern representation, resulting in more detailed outcomes.

This article also picked three photographs from the BSD100 dataset as a comparison example for checking the visual effects, as seen in Figure 4.3. This article contrasts conventional methods of restoration and alternative methods of restoration and bicubic and profound methods of learning.

4.1.2 Model Size Analyses

The Table 4.3 displays the scale and output of existing CNN SR models. MemNet and NLRG provide far fewer criteria for the output loss costs of these approaches. Not only did NSAN have less parameters than RDN, RCAN and SAN, but also achieved improved performance, meaning that NSAN may have a perfect performance trade-off between complexity and efficiency of the model.

4.1.3 Speed comparison

In order to make our method practically used, we conduct a experiment to demonstrate the speed of our proposed method. As shown in Table below, we test several methods on Kirin 970, which is a artificial intelligence mobile phone chip. It should be noted that our proposed ASRN achieve the best PSNR than other methods. Although the speed of ASRN are little large than VDSR, ASRN obtain a good balance between performance and speed.

Table 4.4: Speed test results for scale factor 4 on Set5

Method	VDSR	LapSRN	MemNet	NSAN	ARSN
Speed(ms)	2.45	3.12	5.78	3.83	2.97
PSNR(dB)	37.53/0.9587	37.52/0.9591	37.78/0.9597	38.43/.9634	38.31/.9687

4.1.4 Visual Comparison between ARSN and NSAN

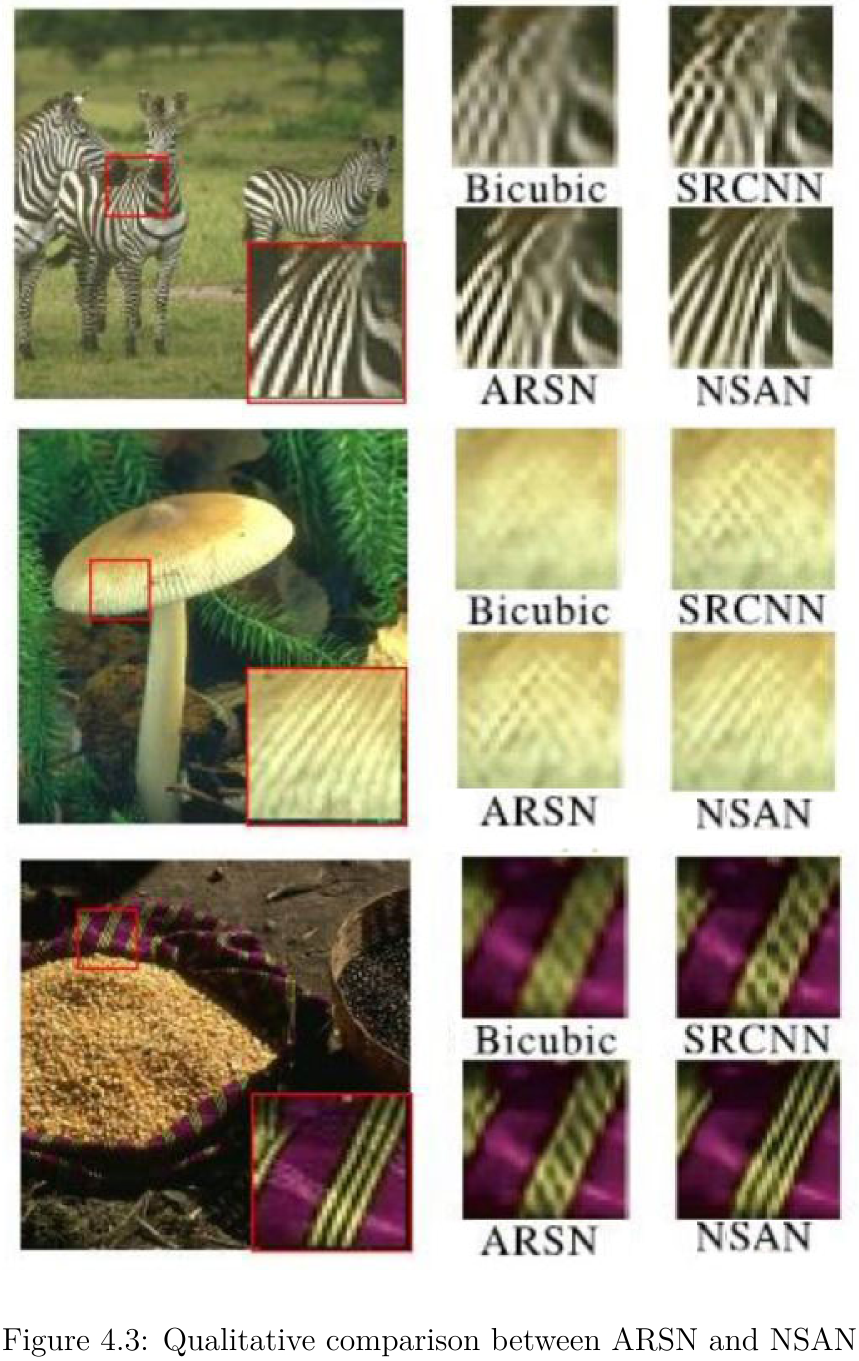


Figure 4.3: Qualitative comparison between ARSN and NSAN

Chapter 5

Conclusion

5.1 Main Contribution

In this dissertation, I proposed a network with attention mechanism (NSAN) for SISR. By utilizing this network structure, the high-order enhancement group supports NSAN to grab structural information and long-term dependences by integrating non-local operations. At the same time, NHAG allows a large amount of low-frequency data in the LR image to be utilized in the local skip connection. NSAN does not just use spatial correlation but also learns the interdependence of higher-order features through the global covariance pool to obtain a more discriminative representation through the NSA module. Experiments have shown NSAN 's efficacy in mathematical, visual, and evaluation analysis.

At the same time, I also introduced a lightweight network with automatic residual scaling algorithm, which can be used for super-resolution reconstruction. Through this algorithm, the number of parameters can be reduced, and the image can be reconstructed in real-time. The model proposed in this article has been tested in many ways. The performance of the standard test data set has reached a very high level. Compared with deepened super-resolution, the previously proposed convolutional neural network model has lower performance but higher real-time processing speed capacity; In addition, this method dramatically reduces model parameters and calculations. Therefore, the model has high practicality.

5.2 Future Work

5.2.1 Super Resolution Optical Microscope

This sort of microscopic image enhancement and transfer of microscopic LR to HR image is a fascinating area with little research. My experiments have shown that deep learning can be utilised to boost the efficiency of mobile microscopes in their imaging significantly.

The development of cost-effective portable microscopic imaging equipment based on cell phones has dramatically improved over the last few years, with possible impacts on environmental health and safety inspections. However, this handheld microscope system also has multiple bugs and causes of aberrations, and at the same time, the target data collection is small relative to the conventional SISR. With the advent of robust unsupervised learning, however, I believe this process, which will have a remarkable effect on the microscopic LR-to-HR super-resolution, will offer new functions and realise applications that cannot be done with today's optical microscope technologies.

5.2.2 Domain-specific Application

Not only can super-resolution be used specifically on particular fields for data and scenes, some vision functions will also, be significantly improved. It is also also a promising path to extend SR to more specialised areas, such as video monitoring, product analysis, object detection, medical imaging, and scene rendering.

Bibliography

[1] Y. Zhang, "Problems in the fusion of commercial high-resolution satelitte as well as landsat 7 images and initial solutions," *International Archives of Photogrammetry Remote Sensing and Spatial Information Sciences*, vol. 34, no. 4, pp. 587–592, 2002. 1.1

[2] L. Bo, C. Hong, S. Shan, and X. Chen, "Low-resolution face recognition via coupled locality preserving mappings," *IEEE Signal Processing Letters*, vol. 17, no. 1, pp. 20–23, 2009. 1.1

[3] J. A. Kennedy, O. Israel, A. Frenkel, R. Bar-Shalom, and H. Azhari, "Super-resolution in pet imaging," *IEEE transactions on medical imaging*, vol. 25, no. 2, pp. 137–147, 2006. 1.1

[4] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 2, pp. 295–307, 2015. 1.1, 2.1, 2.3.1, 3.3.1, 3.3.2, 3.3.3, 4.1.1

[5] C. Dong, C. C. Loy, and X. Tang, "Accelerating the super-resolution convolutional neural network," in *European conference on computer vision*, pp. 391–407, Springer, 2016. 1.1, 2.1, 3.1.1, 3.3.2, 4.1.1

[6] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Deep laplacian pyramid

networks for fast and accurate super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 624–632, 2017. 1.1, 2.1, 2.3.4, 3.3.4, 4.1.1

[7] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee, "Enhanced deep residual networks for single image super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 136–144, 2017. 1.1, 2.1, 2.3.2, 3.1.1, 3.1.1, 3.1.2, 3.3.6

[8] Y. Tai, J. Yang, and X. Liu, "Image super-resolution via deep recursive residual network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3147–3155, 2017. 1.1, 2.1

[9] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 286–301, 2018. 1.1, 2.1, 2.2, 3.3.10, 4.1.1

[10] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017. 1.1

[11] J. Kim, J. Kwon Lee, and K. Mu Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1646–1654, 2016. 1.1, 2.1, 2.3.1, 3.3.3, 4.1.1

[12] J. Kim, J. Kwon Lee, and K. Mu Lee, "Deeply-recursive convolutional network for image super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1637–1645, 2016. 1.1

[13] Y. Tai, J. Yang, X. Liu, and C. Xu, "Memnet: A persistent memory network

for image restoration," in *Proceedings of the IEEE international conference on computer vision*, pp. 4539–4547, 2017. 1.1, 3.3.5, 4.1.1

[14] A. Chakrabarti, A. Rajagopalan, and R. Chellappa, "Super-resolution of face images using kernel pca-based prior," *IEEE Transactions on Multimedia*, vol. 9, no. 4, pp. 888–892, 2007. 2.1

[15] D. Capel and A. Zisserman, "Super-resolution from multiple views using learnt image models," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, vol. 2, pp. II–II, IEEE, 2001. 2.1

[16] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE transactions on image processing*, vol. 19, no. 11, pp. 2861–2873, 2010. 2.1

[17] H. Chang, D.-Y. Yeung, and Y. Xiong, "Super-resolution through neighbor embedding," in *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, vol. 1, pp. I–I, IEEE, 2004. 2.1

[18] D. Glasner, S. Bagon, and M. Irani, "Super-resolution from a single image," in *2009 IEEE 12th international conference on computer vision*, pp. 349–356, IEEE, 2009. 2.1

[19] G. Freedman and R. Fattal, "Image and video upscaling from local self-examples," *ACM Transactions on Graphics (TOG)*, vol. 30, no. 2, p. 12, 2011. 2.1

[20] J.-B. Huang, A. Singh, and N. Ahuja, "Single image super-resolution from transformed self-exemplars," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5197–5206, 2015. 2.1

[21] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, *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. 2.1

[22] N. Ahn, B. Kang, and K.-A. Sohn, "Fast, accurate, and lightweight super-resolution with cascading residual network," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 252–268, 2018. 2.1

[23] J. Li, F. Fang, K. Mei, and G. Zhang, "Multi-scale residual network for image super-resolution," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 517–532, 2018. 2.1, 3.1.2

[24] T. Dai, J. Cai, Y. Zhang, S.-T. Xia, and L. Zhang, "Second-order attention network for single image super-resolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 11065–11074, 2019. 2.1, 2.2, 3.1.1, 3.1.1, 3.1.2, 3.3.11

[25] X. He, Z. Mo, P. Wang, Y. Liu, M. Yang, and J. Cheng, "Ode-inspired network design for single image super-resolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1732–1741, 2019. 2.1

[26] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 11, pp. 1254–1259, 1998. 2.2

[27] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132–7141, 2018. 2.2

[28] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in

Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7794–7803, 2018. 2.2

[29] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "Ccnet: Criss-cross attention for semantic segmentation," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 603–612, 2019. 2.2

[30] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3146–3154, 2019. 2.2

[31] Y. Zhang, K. Li, K. Li, B. Zhong, and Y. Fu, "Residual non-local attention networks for image restoration," *arXiv preprint arXiv:1903.10082*, 2019. 2.2

[32] S. Anwar, S. Khan, and N. Barnes, "A deep journey into super-resolution: A survey," *ACM Comput. Surv.*, vol. 53, May 2020. 2.3

[33] J. Lyn and S. Yan, "Non-local second-order attention network for single image super resolution," in *Machine Learning and Knowledge Extraction* (A. Holzinger, P. Kieseberg, A. M. Tjoa, and E. Weippl, eds.), (Cham), pp. 267–279, Springer International Publishing, 2020. 3.1.1

[34] 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," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1874–1883, 2016. 3.1.1

[35] K. Zhang, W. Zuo, and L. Zhang, "Learning a single convolutional super-resolution network for multiple degradations," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3262–3271, 2018. 3.3.7

[36] M. Haris, G. Shakhnarovich, and N. Ukita, “Deep back-projection networks for super-resolution,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1664–1673, 2018. 3.3.8

[37] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, “Residual dense network for image super-resolution,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2472–2481, 2018. 3.3.9

[38] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, “Residual dense network for image super-resolution,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2472–2481, 2018. 4.1, 4.1.1