SALIENT OBJECT SUPER-RESOLUTION

by

ATEENDRA GAUR
202111020

A Thesis Submitted in Partial Fulfilment of the Requirements for the Degree of

MASTER OF TECHNOLOGY
in
INFORMATION AND COMMUNICATION TECHNOLOGY
to

DHIRUBHAI AMBANI INSTITUTE OF INFORMATION AND COMMUNICATION TECHNOLOGY

May, 2023

Declaration

I hereby declare that

i) the thesis comprises of my original work towards the degree of Master of Technology in Information and Communication Technology at Dhirubhai Ambani Institute of Information and Communication Technology and has not been submitted elsewhere for a degree,

ii) due acknowledgment has been made in the text to all the reference material used.

ATEENDRA GAUR

Certificate

This is to certify that the thesis work entitled **SALIENT OBJECT SUPER-RESOLUTION** has been carried out by **ATEENDRA GAUR** for the degree of Master of Technology in Information and Communication Technology at *Dhirubhai Ambani Institute of Information and Communication Technology* under my supervision.

Prof. Srimanta Mandal
Thesis Supervisor

Acknowledgement

I would like to express my deepest appreciation and gratitude to all those who have supported me throughout the journey of completing this thesis.

First and foremost, I am immensely grateful to my supervisor, **Prof. Srimanta Mandal**, for their unwavering guidance, insightful suggestions, and continuous support. Their expertise, patience, and encouragement have been invaluable in shaping this research and pushing me to reach my fullest potential.

My gratitude extends to my family and friends, especially **Vraj**, for their unwavering support, understanding, and motivation throughout this process. Their belief in me and their willingness to listen and provide encouragement has been instrumental in my perseverance.

Additionally, I would like to acknowledge the **Dhirubhai Ambani Institute of Information and Communication Technology** for providing me with a conducive research environment, access to resources, and opportunities to grow as a researcher. The support from the faculty, staff, and fellow students has been invaluable in shaping my academic journey.

Lastly, I sincerely appreciate all the authors, researchers, and scholars whose work I have referenced in this thesis. Their contributions to the field have shaped my understanding and guided my research.

I offer my deepest gratitude to all those mentioned above and to anyone else who has contributed in any way, no matter how big or small. Your support, encouragement, and guidance have played a significant role in successfully completing this thesis.

ATEENDRA GAUR

Contents

Abstract

Salient object super-resolution refers to enhancing the resolution and details of salient objects or regions in an image. It is a sub-field of image super-resolution, which aims to generate high-resolution salient object images from low-resolution inputs. Several approaches have been used for either salient object detection or image super-resolution, but no method employs both in a single mechanism.

The aim behind salient object super-resolution is to provide a more focused, informative, and visually pleasing representation of images by prioritizing and enhancing the most relevant and eye-catching regions. This can lead to improved performance in various applications like surveillance, medical imaging etc. and a better viewing experience for users.

We propose a salient object super-resolution approach that addresses the challenges inherent in this task, like fine details preservation, inconsistent saliency map quality, computational complexity, ambiguity and uncertainty etc. This approach involves salient object detection, salient object segmentation, salient object super-resolution, restacking of salient objects, and guided image smoothening. Each step is designed to improve salient objects' resolution and visual quality while preserving the remaining image content.

For the super-resolution task, we employed three different models, namely SRGAN (Super-Resolution Generative Adversarial Network) [26], NLSN (Non-Local Sparse Attention Network) [39], and DRT (Deraining Recursive Transformer) [32]. We used the "Salient Object Detection with Robust Background Detection" method [58] for saliency detection.

Further, we explore the potential of a hybrid model that combines the DRT and Non-Local Sparse Attention techniques for the super-resolution task. The DRT model, initially designed for deraining tasks, is adapted for super-resolution to restore fine details and textures within the low-resolution image effectively. The Non-Local Sparse Attention mechanism is incorporated to selectively attend to relevant spatial and channel information, improving the preservation of essential features while suppressing noise and artefacts.

Overall, our work contributes to advancing salient object super-resolution tech-

niques and explores the potential of a hybrid model, TraNLSN, for further improvements. Analyzing the results from the ongoing training phase will provide insights into the effectiveness of the hybrid model and its potential applications in various domains.

List of Tables

List of Figures

CHAPTER 1
Introduction

In recent years, image super-resolution has garnered significant attention in the field of computer vision. The task of increasing the resolution and enhancing the visual details of low-resolution images has various applications, including surveillance, medical imaging, and satellite imagery analysis. Among the various sub-fields of image super-resolution, the focus on salient object super-resolution has emerged as a promising direction. This sub-field specifically aims to generate high-resolution salient object images from low-resolution inputs, preserving the essential features and improving visual quality.

Salient objects play a crucial role in human perception and cognition, as they are the visually significant regions that attract immediate attention. However, existing super-resolution methods often treat the entire image uniformly, neglecting the importance of preserving salient object details. Addressing this limitation, we propose a comprehensive salient object super-resolution approach that incorporates salient object detection, salient object segmentation, salient object super-resolution, restacking of salient objects, and guided image smoothening. Each step is meticulously designed to improve the resolution and visual quality of salient objects while maintaining the fidelity of the rest of the image content.

In salient object super-resolution, the primary focus is preserving and enhancing salient objects' resolution while potentially omitting or downscaling non-salient regions. This approach is motivated by the understanding that not all parts of an image contribute equally to the overall visual quality or the informative content. In some cases, only specific regions of interest within an image, which are relatively small compared to the entire image, require high-resolution reconstruction. Fig. 1.1 (b), (d), and (f) exemplify such scenarios, where the regions of interest are confined to a very small portion of the image, rendering it unnecessary to perform super-resolution on the entire image.

Salient object super-resolution has numerous applications in computer vision and image processing. It can enhance image resolution in various domains, in-

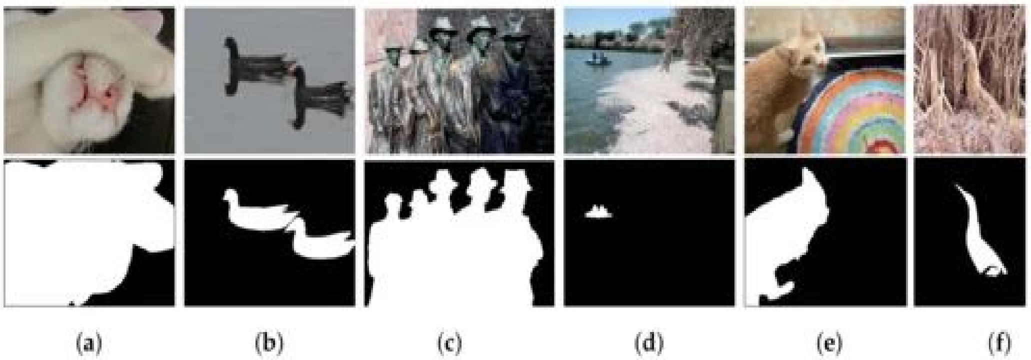


Figure 1.1: Images and their Respective masks displaying ROIs [10]

cluding surveillance, remote sensing, medical imaging, and digital photography. Improving salient objects' visual quality and details enables better analysis, interpretation, and understanding of images in these domains.

The goal of salient object super-resolution is to generate high-resolution details and textures while preserving the overall structure and context of the salient objects. This task is challenging due to the inherent limitations of low-resolution images, such as loss of fine details, blurriness, and aliasing artefacts. Super-resolution algorithms employ various computational methods to estimate and reconstruct the missing details, producing a high-resolution version of the salient object.

Salient object detection often leverages low-level image priors, such as edge information and local image statistics, and high-level semantic knowledge about the salient objects. In contrast, image super-resolution uses patch-based, deep learning, and hybrid approaches combining multiple strategies.

Patch-based methods divide the low-resolution image into patches and search for similar patches in a high-resolution training set. The high-resolution patches then reconstruct the salient object at a higher resolution. Deep learning-based methods utilize convolutional neural networks (CNNs) to learn the mapping between low-resolution and high-resolution images. These neural networks are trained on large datasets to effectively capture the complex relationships between the input and output images. Hybrid approaches combine the strengths of both patch-based and deep learning-based methods, often achieving superior results.

Salient object super-resolution is challenging due to several factors and limitations associated with the process. Some of the critical challenges in salient object super-resolution include:

1. **Limited information in low-resolution images:** Low-resolution images inherently contain less information and fewer details than their high-resolution

counterparts. The loss of fine details and textures in salient objects poses a significant challenge in accurately reconstructing them at a higher resolution.

2. **Ambiguity and uncertainty:** Salient objects can exhibit complex shapes, textures, and appearances. It can be challenging to disentangle the precise details of the salient objects from noise, artefacts, or background information in low-resolution images. This ambiguity and uncertainty make the super-resolution task challenging.

3. **Non-uniform degradation:** Low-resolution images can suffer from various types of degradation, such as blurring, aliasing, and compression artefacts. However, the degradation is not necessarily uniform across the entire image. Different regions within the salient object may exhibit different types and levels of degradation, making the super-resolution process more complex.

4. **Limited training data:** Training deep learning models for salient object super-resolution often requires many paired high-resolution and low-resolution images. However, obtaining such paired data can be challenging and time-consuming, especially when capturing high-resolution ground truth images. Limited training data can impact the generalization ability and performance of super-resolution algorithms.

5. **Computational complexity:** Super-resolution algorithms can be computationally intensive, especially those based on deep learning. The complexity increases further when dealing with salient object super-resolution, as the algorithms must focus on specific regions of interest within the image. Efficiently processing and enhancing the resolution of salient objects while maintaining real-time or near-real-time performance can be a significant challenge.

6. **The trade-off between details and artefacts:** In enhancing the resolution of salient objects, there is a delicate balance between generating high-quality information and avoiding introducing artefacts. Super-resolution methods must avoid over-sharpening artefacts, noise amplification, or unrealistic textures that can degrade the overall visual quality of the salient objects.

7. **Subjectivity and perception:** Salient object super-resolution is often evaluated based on subjective visual quality assessments. Human perception and

preferences play a crucial role in judging the success of super-resolution algorithms. Accommodating the subjective nature of evaluation and meeting diverse perceptual expectations pose additional challenges.

Addressing these challenges requires ongoing research and development of innovative algorithms, data augmentation techniques, and evaluation methodologies. Advances in machine learning, deep learning, and computer vision are continuously being explored to overcome these difficulties and improve the performance of salient object super-resolution techniques.

In summary, salient object super-resolution is a specialized field within image super-resolution that focuses on enhancing the resolution and details of visually important regions in an image. It utilizes various computational techniques, including patch-based methods, deep learning-based approaches, and hybrid strategies, to generate high-resolution versions of salient objects while preserving their overall structure and context.

Let's delve a bit deeper into the basics of salient object detection and super-resolution one by one.

1.1 Salient Object Detection

According to a survey paper [44], visual saliency prediction can be approached from two main directions: visual-attention prediction and saliency detection in computer vision. Each direction focuses on different aspects of saliency and employs distinct approaches to achieve accurate saliency prediction.

Visual-attention prediction models are primarily concerned with understanding human visual attention and predicting eye-gaze patterns. These models typically utilize simple feature channels, such as colour, orientation, and intensity, to capture salient regions in images. By analyzing these features, they aim to estimate where humans will likely focus their attention.

On the other hand, saliency detection models can be categorized into heuristic-based and learning-based approaches. Heuristic-based methods use bottom-up features like contrast, location, and texture to identify salient regions. Within the heuristic-based approach, several subcategories exist based on different feature priors:

1. Local contrast-based methods [8, 16, 18, 21, 22, 29, 30, 36, 37, 57] calculate the saliency value map by considering local features between different regions,

such as colour, illumination, orientation, and motion information. By analyzing the contrast between neighbouring regions, these methods identify salient regions.

2. Global contrast-based methods [1, 7, 9, 11, 15, 42, 49, 54] calculate the saliency value map by considering global features. They assign high saliency scores to regions with similar values indicating their salient nature. These methods leverage overall contrast information to identify salient regions.

3. Center-prior methods [21, 22, 41, 48, 49] aim to highlight the centre region of an image or combine centre-based cues with other features to identify salient regions or objects. The central region is often considered a spatial feature that attracts attention.

4. Background-prior methods [7, 19, 20, 27, 30, 47, 50, 59] treat the narrow border region as the background of the image. They calculate the saliency score by considering the contrast against the background, using the background seeds as a reference. This approach helps distinguish salient objects from the background.

5. Objectness-prior methods [3, 6, 17, 23, 28, 45] employ object proposals to assist in salient object detection. They measure the probability that a whole object exists in an image by evaluating the objectness score for each random window. Object proposals provide valuable information for identifying salient objects.

6. Bayesian framework methods [35, 48] use Bayesian principles to estimate the probability of pixels or regions being salient based on visual features and prior knowledge about saliency. These methods typically involve probabilistic models to compute the saliency maps.

Exploring these different approaches allows researchers to develop effective saliency prediction and detection models. The survey paper [44] serves as a valuable resource, providing a comprehensive overview of these approaches and shedding light on the advancements and trends in saliency prediction within the field of computer vision.

1.2 Super-Resolution

The super-resolution survey paper [46] highlights the importance of formulating an observation model that connects the original high-resolution (HR) image to the observed low-resolution (LR) images in the super-resolution image reconstruction challenge. Typically, the observed LR images are obtained by sub-sampling and blurring operations applied to the HR image, as depicted in Figure 1.2.

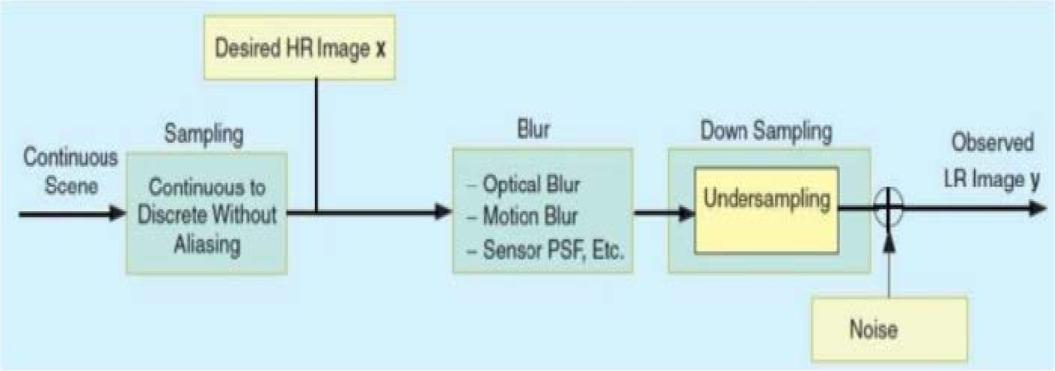


Figure 1.2: Observation model relating LR images to HR images [46]

The observation model can be represented by Equation 1.1 as follows:

$$f = KUg + m \tag{1.1}$$

In this equation, $f \in \mathbb{R}^y$ represents the LR image, $g \in \mathbb{R}^x$ represents the original HR image, $K \in \mathbb{R}^{y \times x}$ is the downsampling operator, $U \in \mathbb{R}^{x \times x}$ is the blurring operator and $m \in \mathbb{R}^y$ is the noise component. Here, $x > y$. The block diagram in Figure 1.2 illustrates the observation model. Since the matrix KU has more columns than rows, there are infinitely many solutions, necessitating regularization in super-resolution.

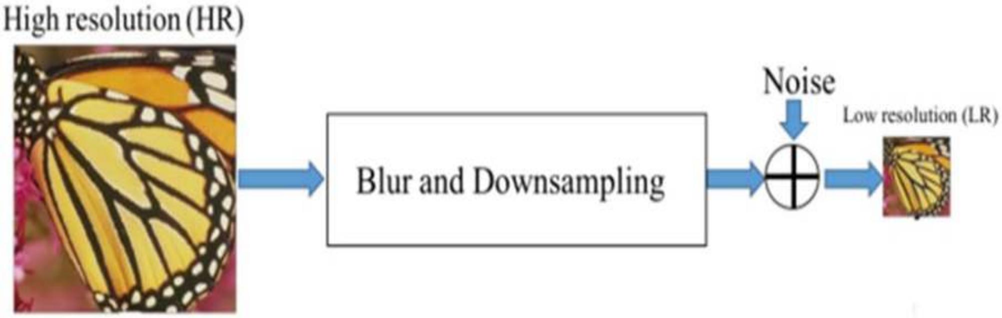


Figure 1.3: Overall Framework Sketch of LR Image formation [51]

Super-resolution can be categorized into two types based on the number of LR images: multiple-image super-resolution and single-image super-resolution. Multiple-image super-resolution utilizes multiple LR images of the same scene to generate a high-resolution image, exploiting the redundancy present in the input images. These multiple images provide complementary information, such as fine details and texture, which can be aligned and fused to generate high-resolution images. However, in many cases, multiple LR images are unavailable, leading to a focus on single-image super-resolution algorithms.

Single-image super-resolution relies on the self-similarity property, which states that small patches within an image can exhibit similar patterns and structures that recur in different parts of the same image. This property arises from common visual elements such as textures and edges present in natural scenes. Single-image super-resolution algorithms leverage this property to infer missing high-frequency information in LR patches by exploiting the redundancy within the image.

Figure 1.3 provides a basic sketch of how an LR image can be represented using a simple equation 1.2:

$$f = Kgb + m \tag{1.2}$$

In this equation, f represents the LR image, g represents the unknown HR image, b represents the blurry kernel, and m represents noise. An LR image is formed by applying a blurry kernel to the HR image, downsampling, and adding noise.

1.3 Objective

Salient object super-resolution aims to enhance the visual quality of salient objects within an image and improve object recognition and understanding. Increasing the resolution and reconstructing fine details make salient objects more visually appealing and informative, making them easier to identify and distinguish from the background. This enhancement in visual quality is essential for applications such as image editing, object detection, tracking, and scene understanding, where accurate recognition and understanding of salient objects are crucial.

Additionally, salient object super-resolution aims to preserve the semantics and context of the salient objects during the resolution enhancement process. It ensures that the reconstructed objects retain their original appearance and characteristics, maintaining the integrity of the visual content. This preservation of semantics and context is vital for maintaining the visual consistency and coherency of the image.

Moreover, high-resolution salient objects enable detailed analysis and inspection of specific regions of interest. Researchers, professionals, and analysts often need to examine particular areas within an image in detail, and by enhancing the resolution of salient objects, salient object super-resolution enables a more thorough and comprehensive examination. This provides valuable insights and supports decision-making processes in various fields.

Furthermore, salient object super-resolution contributes to advancements in graphics and visual communication applications. It enables the creation of visually appealing images with enhanced salient objects, which find applications in domains such as advertising, digital media, and user interfaces. The improved resolution and visual quality of salient objects enhance visual content's overall aesthetic appeal and impact.

Salient object super-resolution aims to enhance visual quality, improve object recognition and understanding, preserve semantics and context, enable detailed analysis and inspection, and advance graphics and visual communication applications. By achieving these objectives, salient object super-resolution contributes to the broader field of image processing and computer vision, benefiting various domains and applications.

1.4 Contribution

This thesis work makes several significant contributions to the field of salient object super-resolution. The main contributions of this research can be summarized as follows:

1. **Selective Salient Object Super-Resolution Approach:** The proposed approach introduces a novel selective strategy for salient object super-resolution. By leveraging salient object detection and segmentation techniques, the method identifies and isolates regions of interest within the low-resolution image that require high-resolution reconstruction. This selective approach significantly reduces computational complexity and processing time by avoiding unnecessary super-resolution processing on non-salient areas.

2. **TraNLSN: Hybrid Model for Effective Super-Resolution:** The development and analysis of the hybrid model TraNLSN represent a noteworthy contribution. Combining the strengths of the DRT (Deraining Recursive Transformer) model and the Non-Local Sparse Attention technique, TraNLSN is specifically designed and optimized for super-resolution tasks. Its adaptation for restoring fine details and textures within the identified regions of interest improves the overall quality of the salient object super-resolution process.

Overall, the contributions of this thesis work have implications for various applications, including image analysis, object recognition, and computer vision tasks where the emphasis on preserving and enhancing salient objects is crucial. The proposed approach's resource efficiency and improved visual quality open up possibilities for real-world applications in domains such as surveillance, medical imaging, and remote sensing, where accurate and visually appealing representations are vital.

1.5 Organization

The remaining work is structured as follows to provide a comprehensive understanding of the research conducted.

Chapter 2 serves as a literature review, delving into the state-of-the-art models for visual saliency prediction and super-resolution. This chapter examines the different loss functions utilized in these models, their relationships with various evaluation metrics, and their architectural complexity and training requirements. By reviewing existing research, Chapter 2 sets the foundation for understanding the advancements and challenges in the field.

In Chapter 3, our proposed method is presented, along with the introduction of a new model called TraNLSN. This chapter outlines the different stages involved in the salient object super-resolution process and provides a detailed description of the architecture of the TraNLSN model. Chapter 3 highlights our research's novel aspects and contributions by explaining the proposed method and model.

Chapter 4 focuses on the experimental setup used to evaluate the presented method. This chapter describes the datasets utilized, the evaluation metrics employed to assess the performance of the proposed method, and the results obtained through comparisons with state-of-the-art methods. The experimental setup provides insights into the effectiveness and superiority of our approach in enhancing salient object super-resolution.

Finally, Chapter 5 and Chapter 6 conclude the thesis by summarizing the main findings and conclusions derived from the research. These chapters also discuss the potential future directions and scopes for further improvement and application of the presented method and the proposed model TraNLSN. By closing the thesis with a reflection on the research outcomes and future possibilities, Chapter 5 and Chapter 6 offer a comprehensive understanding of the contributions made and the potential impact of the research in the field of salient object super-resolution.

Overall, the organization of the remaining work ensures a logical flow of information, starting with the literature review, followed by the presentation of the proposed method, experimental evaluation, and finally, the conclusion and future prospects.

CHAPTER 2
Related Works

While super-resolution and salient object detection have been extensively studied as separate domains, the research on salient object super-resolution, explicitly focusing on enhancing the resolution and visual quality of salient objects within low-resolution images, is relatively limited. This specific combination of tasks poses unique challenges and requires tailored approaches to address them effectively.

Most existing super-resolution methods primarily focus on enhancing the resolution of the entire image without considering the saliency of objects within the image. On the other hand, salient object detection methods aim to identify and localize salient objects but do not explicitly tackle the problem of super-resolving those objects. As a result, the integration of these two tasks to achieve salient object super-resolution remains relatively unexplored.

The limited research in salient object super-resolution can be attributed to the complexity of the task and the lack of available datasets and evaluation metrics specifically designed for this purpose. Developing effective algorithms for salient object super-resolution requires addressing challenges such as accurately identifying salient objects, preserving their structure and details during the super-resolution process, and seamlessly integrating the enhanced objects back into the original image.

Despite the scarcity of work in salient object super-resolution, recent advancements in deep learning, attention mechanisms, and image processing techniques provide a promising foundation for exploring and developing novel approaches in this domain. By combining the strengths of salient object detection, super-resolution, and related techniques, researchers can contribute to filling the gap in the literature and advancing the field of salient object super-resolution.

Therefore, the proposed method in this thesis aims to bridge this research gap by providing a comprehensive framework for salient object super-resolution. By breaking down the task into stages and leveraging advanced techniques such as

SRGAN, Non-Local Sparse Attention, and the Deraining Recursive Transformer, we aim to enhance the resolution and visual quality of salient objects within low-resolution images, thus contributing to the limited body of work in salient object super-resolution.

2.1 Super-Resolution

SRCNN (Super-Resolution Convolutional Neural Network), VDSR (Very Deep Super-Resolution), SRGAN (Super-Resolution Generative Adversarial Network), EDSR (Enhanced Deep Super-Resolution), RCAN (Residual Channel Attention Networks), and NLSN (Non-Local Sparse Attention) are all notable methods in the field of image super-resolution.

2.1.1 SRCNN

SRCNN was one of the pioneering deep learning-based super-resolution methods introduced by Dong et al. It utilizes a three-layer convolutional neural network to learn the mapping between low-resolution and high-resolution image patches. By effectively exploiting the hierarchical representations, SRCNN demonstrated significant improvement in super-resolution performance [46].

The network of SRCNN is not deep. It consists of three parts as shown in figure 2.1:

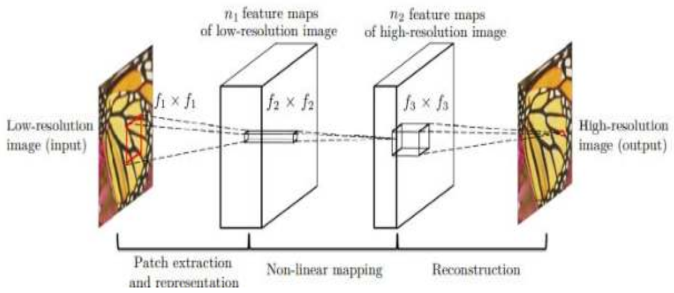

n_1 feature maps of low-resolution image
n_2 feature maps of high-resolution image
$f_1 \times f_1$
$f_2 \times f_2$
$f_3 \times f_3$
Low-resolution image (input)
High-resolution image (output)
Patch extraction and representation
Non-linear mapping
Reconstruction

Figure 2.1: Architecture of Super-Resolution Convolutional Neural Network [46]

a) **Patch Extraction and Representation:** The low-resolution input is first upscaled to the required size using bicubic interpolation before being fed into the SRCNN network. The first layer performs a convolution with Relu to get $E_1(Y)$.

$$E_1(Y) = max(0, W_1 * Y + B_1) \tag{2.1}$$

Here, X is a High-resolution ground truth image, Y is a the low-resolution image that has been bicubic up-sampled and size of W_1 is $c \times f_1 \times f_1 \times n_1$ where c is number of channels, f_1 is filter size and n_1 is number of filters.B_1 is a n_1 dimensional bias vector.

b) **Non-linear mapping:** After calculating $E_1(Y)$, nonlinear mapping is performed by:

$$E_2(Y) = max(0, W_1 * E_1(Y) + B_2) \tag{2.2}$$

Here, size of W_2 is $n_1 \times 1 \times 1 \times n_2$ and B_2 is a n_2 dimensional bias vector. So here mapping of n_1 dimensional vector to n_2 dimensional vector is done.

c) **Reconstruction:** Reconstruction is required after the nonlinear mapping. Thus convolution is again done here:

$$E(Y) = W_3 * E_2(Y) + B_3 \tag{2.3}$$

Here size of W_3 is $n_2 \times f_3 \times f_3 \times c$ and B_3 is bias of size c.

The standard loss function average of mean squared error is used in this network for training. Loss function $P(/theta)$ can be given as:

$$P(\theta) = \frac{1}{n} \sum_{i=1}^{n} ||E(Y_i; \theta) - X_i||^2 \tag{2.4}$$

where $E(Y_i; \theta)$ denotes reconstructed super-resolution image, X_i denotes ground truth high resolution image and n denotes total no. of samples.

2.1.2 VDSR

VDSR, proposed by Kim et al., introduced a deep residual network architecture for super-resolution. It utilizes a very deep network with residual learning to recover high-resolution details. By employing a skip connection, VDSR effectively addresses the vanishing gradient problem and enables the training of deeper networks. It achieved state-of-the-art results by surpassing the performance of previous methods [25].

2.1.3 SRGAN

SRGAN, an adversarial learning-based approach proposed by Ledig et al., combines the power of generative adversarial networks (GANs) with super-resolution.

It employs a generator network to produce high-resolution images and a discriminator network to distinguish between generated and real high-resolution images. SRGAN demonstrated the ability to generate photo-realistic and visually appealing super-resolved images with enhanced perceptual quality [26].

It consists of two major components:

- **Generator:** The generator uses residual networks which use skip connections. While training, First, an HR image is downsampled to an LR image, and then the generator tries to upsample the LR image to SR. After that discriminator tries to distinguish between an SR image and an HR image and generates an adversarial loss, which is backpropagated into the generator.

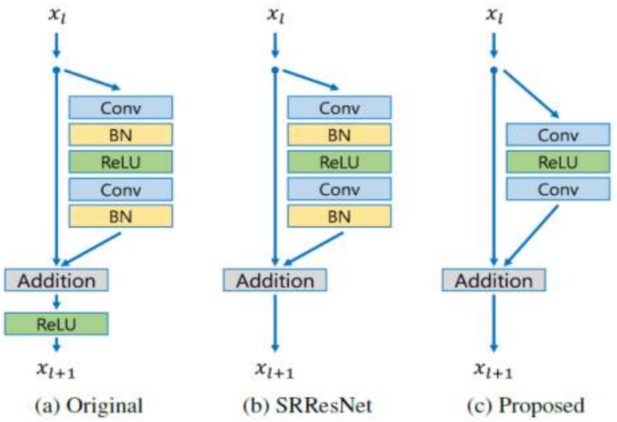
- **Discriminator:** The task here is to discriminate between generated SR images and real HR images. To solve the the adversarial min-max problem of a discriminator network D_{θ_D} is optimized in alternate manner with Generator G_{θ_G} using following equation 2.5 where θ_D and θ_G denotes the weights and biases of the deep network:

$$\min_{\theta_G} \max_{\theta_D} \mathbb{E}_{I^{HR}\ P_{min}(I^{HR})}[\log D_{\theta_D}(I^{HR})] + \mathbb{E}_{I^{LR}\ P_{G}(I^{LR})}[\log(1 - D_{\theta_D}(G_{\theta_G}(I^{LR})))] \tag{2.5}$$

2.1.4 EDSR

EDSR (Enhanced Deep Super-Resolution) is a method that enhances the performance of single-image super-resolution by making specific modifications to the architecture used in SRGAN (Super-Resolution Generative Adversarial Network). EDSR introduces a deeper network with more convolutional layers and utilizes residual connections to capture intricate details and improve reconstruction. Notably, EDSR removes batch normalization, reducing computational complexity and avoiding potential biases during inference. The model is trained efficiently using pixel-wise mean squared error loss and gradient clipping. .

It modified the SR-Resnet block by removing the batch normalization layer, due to which GPU memory usage was reduced by 40% while training compared to SR-Resnet as shown in Figure: 2.2.



(a) Original (b) SRResNet (c) Proposed

Figure 2.2: SR-Resnet Modifications [33]

These modifications result in superior performance compared to SRGAN and other state-of-the-art methods, as measured by the peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM). EDSR offers improved image quality and fidelity, making it a promising approach for super-resolution tasks [33]

2.1.5 RCAN

RCAN [26], proposed by Zhang et al., introduced a residual channel attention mechanism to improve super-resolution performance. By explicitly modelling the interdependencies between channels, RCAN effectively enhances the representation capacity of the network. It introduces channel attention mechanisms to emphasize important image features during super-resolution selectively. By adaptively recalibrating channel-wise features, RCAN achieves significant improvements in perceptual quality and quantitative measures. It outperforms several other CNN-based methods, including SRCNN and EDSR.

As shown in architecture Figure 2.3 , RCAN consists of four parts:

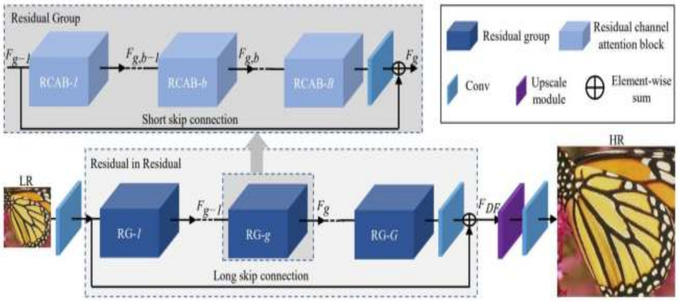


Figure 2.3: Architecture of Residual Channel Attention Network [26]

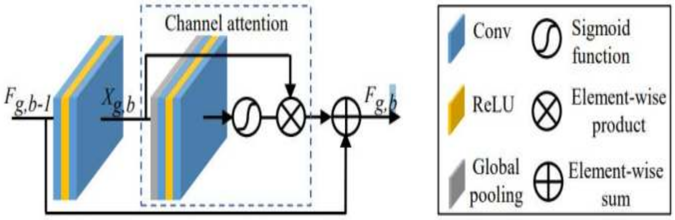


Figure 2.4: Architecture of Residual Channel Attention Block [26]

1. **Shallow-Feature Extraction:** In this part, only one convolutional layer is used to extract the shallow features F_0 from LR input.

   $$F_0 = H_{SF}(I_{LR}) \tag{2.6}$$

   Here H_{SF} is the convolution function.

2. **Residual in Residual (RIR):** Also known as Deep Feature Extraction in this step previous feature is used in the RIR module for deep feature extraction. Here H_{RIR} denotes the proposed RIR structure with G residual groups(RG).

   $$H_{DF} = H_{RIR}(F_0) \tag{2.7}$$

   This proposed RIR for deep feature extraction can achieve the largest depth and provide a large receptive field size. Figure 5 describes the architecture of the Residual Channel Attention Block(RCAB), which consist of a channel attention mechanism.

3. **Upscale module and reconstruction part:** The output from RIR is then upscaled,

   $$F_{UP} = H_{UP}(F_{DF}) \tag{2.8}$$

   and the upscaled feature is then reconstructed using one convolution layer.

   $$I_{SR} = H_{REC}(F_{UP}) = H_{RCAN}(I_{LR}) \tag{2.9}$$

   Here H_{REC} and H_{RCAN} denote the reconstruction layer and the function of the RCAN, respectively. The standard loss function used in this network is:

   $$L(\theta) = \frac{1}{N}\sum_{i=1}^{N} ||H_{RCAN}(I^i_{LR}) - I^i_{HR}||_1 \tag{2.10}$$

2.1.6 NLSN

NLSN (Non-Local Sparse Attention) [39] is a method that utilizes non-local operations and sparsity regularization to capture long-range dependencies and enhance fine details within images. It selectively attends to relevant spatial and channel information, preserving important features while suppressing noise and artefacts. It offers a unique approach to super-resolution and can be combined with other methods to enhance performance.

2.2 Visual Saliency

2.2.1 Global Contrast Based Salient Region Detection

Global Contrast Based Salient Region Detection [7] is a computer vision method that automatically identifies visually salient regions in images. It operates by analyzing the contrast differences between a region and its surroundings. The assumption is that salient regions exhibit higher contrast compared to their neighbours. The method involves preprocessing the image, constructing a contrast map, computing a saliency map, refining the saliency map, and performing post-processing. The resulting saliency map highlights the most visually interesting areas. This method has been widely used, has inspired variations and extensions, and has been evaluated against other state-of-the-art saliency detection methods. It is an effective technique for automatic salient region detection in images, making it valuable for various applications in computer vision.

They proposed a histogram-based method to detect the contrast of images. According to the paper, the Saliency of a pixel is defined using its colour contrast values to all other pixels in the image in $L^*a^*b^*$ (CIELAB) colour space. L^* for perceptual lightness; a* and b* for 4 unique colours red, green, blue, and yellow.

The saliency of an image pixel I_k is defined as:

$$S(I_k) = \sum_{i=1}^{N} D(I_k, I_i) \tag{2.11}$$

Here $D()$ is the colour distance metric, i.e. the Euclidean colour distance between each pixel. Rearranging eq. As the same colour value, pixels will have the same saliency.

$$S(I_k) = S(c_l) = \sum_{j=1}^{n} f_j D(c_l, c_j) \tag{2.12}$$

Here n represents distinct pixels, and f_j represents the probability of colour pixel c_j.

Since there are 2563 combinations possible in the worst case so, time complexity would be very high to compute. They quantised the RGB space while measuring distance in the L*a*b* colour space. So they quantized each colour channel RGB into 12 different values to reduce complexity. They chose 95% most frequently occurring colours, and the rest, 5%, were the colours they replaced by the closest histogram colour.

Quantization may introduce artefacts as similar colours quantised to different values, so they used a smoothing procedure to refine the saliency value for each colour. they took the weighted average of the saliency value of similar colours measured in $L^*a^*b^*$ colour space given as:

$$\frac{1}{m-1}T\sum_{i=1}^{m}(T - D(c, c_i))S(c_i) \tag{2.13}$$

here $m = \frac{n}{4}$ and $T = \sum_{i=1}^{m} D(c, c_i)$ is the sum of distances between colour c and its m nearest neighbours c_i. They used a linearly-varying smoothing weight $(T - D(c, c_i))$ to assign larger weights to colours closer to c in the colour feature space.

2.2.2 The Saliency Filtering

While the Global Contrast Based Salient Region Detection method calculates contrast values and generates a saliency map, **The Saliency Filtering method** [40] takes a step further by applying contrast-based filtering operations, these operations aim to amplify the contrast differences between the salient regions and their surroundings. By doing so, the Saliency Filtering method emphasizes the salient regions more prominently, making them stand out even further.

The key distinction lies in the filtering step of the Saliency Filtering method, where contrast-based techniques are employed to enhance the regions with high contrast values. This additional step allows for a more refined and focused highlighting of the salient regions, resulting in a more robust detection of visually significant areas in the image.

They break down their algorithm into 4 stages, each having its own importance.

(a) **Abstraction:** They decomposed the image into basic elements that preserve the relevant structure but abstract undesirable detail. Specifically, each element should locally abstract the image by clustering pixels with similar properties (like colour) into perceptually homogeneous regions.

For this, they used K-means Clustering in geodesic image distance in Cielab colour space.

(b) **Element Uniqueness:** In the first contrast measure, image regions, which stand out from other regions in certain aspects, catch our attention and should be labelled more salient.

Element uniqueness of the current pixel i, given its position p_i. And color c_i is given as:

$$U_i = \sum_{j=1}^{N} ||(c_i - c_j)||^2.w_{ij}^{(p)} \tag{2.14}$$

$$U_i = \sum_{j=1}^{N} [D(I_i, I_j)]2.w_{ij}^{(p)} \tag{2.15}$$

$$w_{ij}^{(p)} = \frac{1}{Z_i}(\exp -\frac{1}{2}(\sigma_p^2)||(p_i - p_j)||^2)) \tag{2.16}$$

$$\sigma_p^2 = 0.25 \tag{2.17}$$

Where Z_i is normalization factor ensuring:

$$\sum_{j=1}^{n} w_{ij}^{p} = 1 \tag{2.18}$$

This way they handled both local and global contrast.

(c) **Element Distribution:** Ideally colours belonging to the background will be distributed over the entire image exhibiting a high spatial variance, whereas foreground objects are generally more compact. they define the element distribution measure for a segment i using the spatial variance D_i of its color c_i as,

$$D_i = \sum_{j=1}^{N} ||p_i - \mu_i||^2.w_{ij}^{(c)} \tag{2.19}$$

Where,

$$w_{ij}^{(p)} = \frac{1}{Z_i}(\exp{(-\frac{1}{2}(\sigma_p^2))}||(c_i - c_j)||^2)) \tag{2.20}$$

$$\mu_i = \sum w_{ij}^c p_j \tag{2.21}$$

$$\sigma_c = 20 \tag{2.22}$$

(d) **Saliency Assignment:** They Normalized U_i and D_i in the range $[0, 1]$ and further assigned saliency. They found the distribution measure D_i of higher significance and discriminative power. So they used exponential to emphasize D_i as,

$$S_i = U_i \exp{(-k.D_i)} k = 6 \tag{2.23}$$

Naive up-sampling assignment of S_i, carries segmentation errors of abstraction. So they used,

$$S_i^{'} = \sum_{j=1}^{N} w_{ij}.S_j$$

Where,

$$w_{ij} = \frac{1}{Z_i} \exp -\frac{1}{2}(\alpha.||c_i - c_j||^2 + \beta.||p_i - p_j||^2)$$

$$\alpha = \beta = \frac{1}{30}$$

In summary, the Saliency Filtering method builds upon the Global Contrast Based Salient Region Detection method by further incorporating contrast-based filtering techniques to enhance salient regions' detection and emphasis. This refinement step enables a more pronounced contrast between the salient regions and their surroundings, ultimately improving the accuracy and quality of salient region detection in images.

2.2.3 Frequency Tuned Salient Region Detection

The Frequency Tuned Salient Region Detection method [1] is a computer vision technique that identifies salient image regions by analyzing their frequency components. It focuses on the understanding that salient regions exhibit distinct frequency characteristics compared to less significant areas. This method involves preprocessing the image, analyzing its frequency components, applying frequency tuning to enhance relevant frequencies, generating a saliency map based on the tuned frequencies, and performing optional post-processing.

This method utilizes the Frequency-Tuned (FT) channel to highlight salient regions based on their frequency content. They first convert image to CIELAB color space. Then, The center-surround contrast is calculated for each pixel by taking the absolute difference between its intensity value (L value) and the average intensity of its surrounding pixels. This operation enhances the differences between the center region and its surrounding regions.

$$C(x,y) = |L(x,y) - \overline{L}(x,y)| \tag{2.24}$$

where $C(x,y)$ is the center-surround contrast at pixel (x,y), $L(x,y)$is the luminance value at pixel (x,y), and $\overline{L}(x,y)$ is the average luminance of the surrounding pixels.

Then, the center-surround contrast image is convolved with a Gaussian kernel to introduce smoothing and reduce noise.

$$BlurC(x,y) = GaussianBlur(C(x,y)) \tag{2.25}$$

The next step is to take the Fourier Transform of the Gaussian-blurred center-surround contrast image.

$$F(u,v) = FourierTransform(BlurC(x,y)) \tag{2.26}$$

where $F(u,v)$ represents the Fourier Transform at frequency (u,v).

Similar to the previous methods, the amplitude spectrum represents the magnitudes of the frequency components in the Fourier domain. This is typically obtained by calculating the absolute values of the Fourier Transform.

$$A(u,v) = |F(u,v)| \tag{2.27}$$

After that, the amplitude spectrum is modulated by a weighting function that is designed to emphasize the frequency components associated with salient regions while suppressing others. The Frequency-Tuned (FT) channel is obtained by applying the following frequency tuning function:

$$FT(u,v) = A(u,v) * \exp(-(D(u,v)/(2 * \sigma_D)^2)) \tag{2.28}$$

where $D(u,v)$ is the Euclidean distance from each frequency component (u,v)

to a center frequency (u_0, v_0), and σ_D controls the width of the Gaussian in frequency domain.

The modulated amplitude spectrum (FT channel) is transformed back into the spatial domain using the inverse Fourier Transform.

$$FTChannel(x, y) = InverseFourierTransform(FT(u, v)) \tag{2.29}$$

The final saliency map is obtained by normalizing the FT channel values to the range $[0, 1]$.

$$Saliency(x, y) = \frac{(FTChannel(x, y) - \min(FTChannel))}{(\max(FTChannel) - \min(FTChannel))} \tag{2.30}$$

When compared to the previous methods, the Frequency Tuned Salient Region Detection method offers a different perspective by leveraging frequency information. While the Global Contrast Based Salient Region Detection and Saliency Filtering methods primarily rely on contrast analysis, the Frequency Tuned method emphasizes the analysis of frequency components. It considers that salient regions often possess unique frequency signatures, which can be enhanced to highlight their significance.

2.2.4 Saliency Optimization from Robust Background Detection

Saliency Optimization from Robust Background Detection [58] is a computer vision method that extracts salient regions from images by accurately detecting the background and optimizing the saliency map. The method involves two main steps: robust background detection and saliency optimization.

The robust background detection step focuses on identifying the background regions using robust background modelling techniques. This helps differentiate less visually significant parts of the image from foreground objects or salient regions. The saliency optimization step then refines the saliency map by incorporating the detected background information. Optimization techniques are used to assign low saliency scores to the background or suppress its influence on the saliency map, ensuring that the salient regions stand out.

Basically, they calculated the ratio of a region's perimeter on the boundary to the region's overall perimeter or square root of its area as shown in Equation 2.31.

$$BC(R) = \frac{|\{z | z \in R, z \in B\}|}{\sqrt{|\{z | z \in R\}|}} \tag{2.31}$$

where B is the set of image boundary patches, and z is an image patch.

The image is first abstracted as a set of nearly regular superpixels using the SLIC method. They then construct an undirected weighted graph by connecting all adjacent superpixels (x, y) and assign their weight $d_{euc}(x, y)$ as the Euclidean distance between their average colours in the CIE-Lab colour space.

The geodesic distance between any two superpixels $d_{geo}(x, y)$ is defined as the accumulated edge weights along their shortest path on the graph, as shown in Equation 2.32.

$$d_{geo}(x, y) = \min_{x_1 = x, x_2, \ldots, x_n = y} \sum_{i=1}^{n-1} d_{euc}(x_i, x_{i+1}) \tag{2.32}$$

For convenience, they define $d_{geo}(x, x) = 0$. Then they define the "spanning area" of each superpixel x as shown in Equation 2.33.

$$A(x) = \sum_{i=1}^{N} \exp\left(-\frac{d_{geo}^2(x, x_i)}{2\sigma_c^2}\right) = \sum_{i=1}^{N} Z(x, x_i) \tag{2.33}$$

where N is the number of superpixels, equation 2.33 computes a soft area of the region that x belongs to. The operand $Z(x, x_i)$ in the summation is in the range $(0, 1]$ and characterizes how much superpixel x_i contributes to $x's$ area.

Similarly, they define the length along the boundary as shown in Equation 2.34.

$$Lb(p) = \sum_{i=1}^{N} Z(x, x_i).\delta(x_i \in B) \tag{2.34}$$

where $\delta(*)$ is 1 for superpixels on the image boundary and 0 otherwise.

Finally, they compute the boundary connectivity similarly to Equation 2.31.

$$BC(x) = \frac{L_b(x)}{\sqrt{A(x)}} \tag{2.35}$$

They further add edges between any two boundary superpixels to enlarge the boundary connectivity values of background regions, which has little effect on the object regions. This is useful when a physically connected background region is separated due to the occlusion of foreground objects. To compute Equation 2.35, the shortest paths between all superpixel pairs are efficiently calculated using Johnson's algorithm [24] since their graph is very sparse.

Thus, they calculate the saliency map of the salient objects.

The Saliency Optimization from the Robust Background Detection method extracts salient regions by accurately detecting the background and optimizing the saliency map. It combines robust background detection techniques with saliency optimization to ensure that the saliency map primarily highlights the visually salient foreground objects. This method offers a distinct approach to saliency detection when compared to frequency-based methods, emphasizing the accurate separation of salient regions from the background.

CHAPTER 3
Proposed Method

Our proposed method aims to generate a high-resolution output image that explicitly enhances the salient object while preserving the remaining regions of the image in their original low-resolution form.

For the initials, instead of taking an overall deep-learning-based approach to generate salient object super-resolution, we took the hybrid approach.

We divided the problem statement into multiple sub-tasks, each task having its importance. By breaking down the problem statement into multiple sub-tasks, each with its specific importance, we effectively tackled the challenge of dealing with low training data in salient object super-resolution.

By approaching the problem in a modular manner, we were able to focus on different aspects of the super-resolution task separately, addressing the specific challenges associated with each sub-task. The division of the problem into smaller components allowed us to leverage the limited training data more efficiently and effectively.

These sub-tasks are:

1. Salient Object Detection
2. Salient Object Segmentation
3. Salient Object Super-Resolution
4. Re-stacking Salient Object
5. Image Smoothing using Guided Image Filter

By combining these stages, as illustrated in Fig.3.1, we achieve a high-resolution output image that focuses on enhancing the salient object while maintaining the remaining image as it is. This approach allows us to preserve important details and features of the salient object while improving its visual quality and resolution.

Let's expand every segment into a bit of detail.

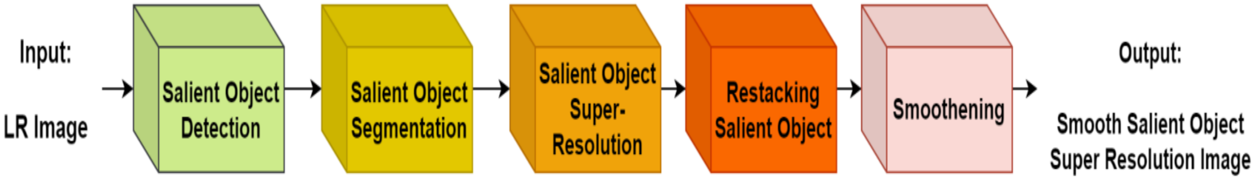

Input:
LR Image
Salient Object Detection
Salient Object Segmentation
Salient Object Super-Resolution
Restacking Salient Object
Smoothening
Output:
Smooth Salient Object Super Resolution Image

Figure 3.1: Overall Flow chart of the proposed method

3.1 Salient Object Detection

Here's the modified text with the equations formatted:

As explained in Chapter 2, to begin the salient object super-resolution process, we generate a saliency map for the low-resolution input image using a specific method mentioned in reference [58] for the respective datasets.

The saliency map serves as a visual representation highlighting the regions in the low-resolution image considered salient or visually distinctive. It helps to identify the areas of interest or importance within the image most likely to contain the salient object.

We can effectively extract the saliency information from the low-resolution image and generate a corresponding saliency map by employing the mentioned method. This map will provide valuable guidance for subsequent stages of the salient object super-resolution process, aiding in the accurate reconstruction of the salient object in high resolution.

We used saliency optimization from robust background detection to detect salient object regions.

3.2 Salient Object Segmentation

After obtaining the saliency map from the previous step, which is a grayscale image highlighting the salient regions, it requires further processing to segment the salient object effectively. As a grayscale image, the saliency map cannot be directly used for segmentation purposes. Therefore, it must be converted into a binary map to facilitate the segmentation process.

The conversion from a grayscale saliency map to a binary map involves thresholding, where a particular threshold value is applied to the saliency map to distinguish between salient and non-salient regions. Pixels in the saliency map with values above the threshold are classified as belonging to the salient object, while pixels below the threshold are considered non-salient.

To get the appropriate threshold value, we employ Otsu's thresholding method [52], which calculates an optimal threshold value to separate foreground and background pixels in an image. The threshold is determined by maximizing the between-class variance or minimizing the within-class variance.

After computing the optimal threshold value, it can be used to separate the foreground and background pixels by assigning them different labels or values.

By converting the saliency map into a binary map, we obtain a clear separation between the salient object and the background, enabling more accurate segmentation. The resulting binary map represents a different mask where the salient thing is represented by foreground (white) pixels, and the non-salient background is represented by background (black) pixels. This binary representation allows for easier identification and extraction of the salient object during the subsequent segmentation steps.

Once the thresholding of the binary map is done, we need the proper connectivity between the salient objects in order to extract the whole region as a single region for further steps. In order to get the connectivity of salient objects, we employed the morphological closing operation [5]. Morphological closing is a fundamental operation in mathematical morphology, which is a branch of image processing and analysis. It removes small dark regions or holes in an image's bright regions while preserving the objects' overall structure and shape. We employed morphological closing operation on the binary map obtained in the previous step, enhancing the connectivity of the salient objects in the given image.

For segmentation, we calculate the minimum upright rectangle encompassing all non-zero (foreground) pixels in the binary map obtained from the previous step. By considering the position and size of this bounding rectangle, we can accurately extract the salient object from the low-resolution image. The bounding rectangle provides essential information regarding the spatial location and dimensions of the salient object, allowing for its precise isolation.

Upon obtaining the bounding rectangle, we crop the region of interest from the low-resolution image, discarding the non-salient background regions. This selective extraction ensures that subsequent super-resolution processing is focused solely on enhancing the details and quality of the salient object itself.

Utilizing robust salient object detection and segmentation methods helps identify and isolate visually significant regions. By focusing on these regions during super-resolution, the limited available information in low-resolution is better preserved and enhanced.

3.3 Salient Object Super-Resolution

After successfully segmenting the salient object from the low-resolution image, the next step in our process is the salient object super-resolution. The objective here is to enhance the resolution and quality of the extracted salient object while keeping the remaining parts of the image unchanged.

By identifying and isolating the regions of interest (salient objects) within the low-resolution image, the super-resolution process can be limited to only these specific areas. This targeted approach significantly reduces the computational burden, as the model does not need to process or enhance the entire image. Instead, it can concentrate its efforts on refining the resolution and visual quality of the visually significant regions.

To achieve salient object super-resolution, we employ specific techniques and algorithms that focus on the salient regions of the image. These techniques leverage the high-frequency information present in the salient object and aim to restore its fine details and sharpness.

The super-resolution process involves generating a high-resolution version of the salient object by exploiting the available low-resolution information and incorporating additional details from external sources or through sophisticated reconstruction methods. This allows us to enhance the visual quality and level of detail specifically within the salient object, providing a more realistic and visually appealing representation.

For salient object super-resolution, we have explored three approaches: SRGAN (Super-Resolution Generative Adversarial Network), Non-Local Sparse Attention, and a trained version of Deraining Recursive Transformer. Each approach has its own distinct methodology and focuses on enhancing the resolution and quality of the salient object in a unique way.

For the super-resolution step, we employed three different models.

3.3.1 SRGAN

SRGAN [26] is a deep learning-based approach that utilizes a generative adversarial network to generate high-resolution images. It is trained to learn the mapping from low-resolution to high-resolution images and produces visually realistic and sharp results by leveraging adversarial training.

The SRGAN (Super-Resolution Generative Adversarial Network) architecture is a deep learning model designed specifically for image super-resolution. It consists of a generator network and a discriminator network, working together in an adversarial manner to enhance the resolution of low-resolution images.

The generator network takes a low-resolution image as input and aims to generate a corresponding high-resolution output. It uses convolutional layers with activation functions to capture image features and progressively upsamples the input to increase its resolution while preserving finer details, as shown in figure 3.2. Skip connections are often incorporated to facilitate the flow of information across different levels of features, aiding in the reconstruction of fine details.

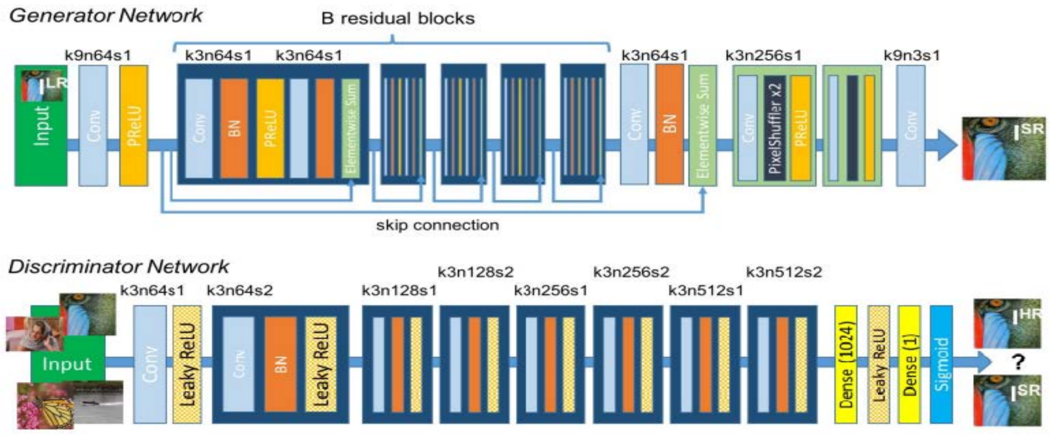


Figure 3.2: The architecture of Super-Resolution Generative Adversarial Network [26]

On the other hand, the discriminator network is responsible for distinguishing between generated high-resolution images and real high-resolution images. It examines either the generated or real images and outputs a probability score indicating the authenticity of the input image. The discriminator typically consists of convolutional layers with activation functions and is trained to correctly classify real and generated images.

During training, the generator and discriminator are trained in an adversarial manner. The generator aims to generate high-resolution images that the discriminator cannot distinguish from real images, while the discriminator aims to classify the images correctly. This adversarial training encourages the generator to produce visually convincing and realistic high-resolution outputs.

To guide the training process, SRGAN utilizes multiple loss functions. The primary loss function is the adversarial loss, which pushes the generator to generate images that fool the discriminator. Additionally, a content loss is employed to ensure that the generated images retain the content of the low-resolution input. Other loss components, such as pixel-wise mean squared error (MSE), may also be used to encourage similarity between the generated and target high-resolution images.

Overall, the SRGAN architecture harnesses the power of adversarial training to generate high-quality, visually appealing high-resolution images from low-resolution inputs, making it a valuable tool for super-resolution tasks.

3.3.2 NLSN

Non-Local Sparse Attention [39] is a technique that employs non-local self-attention mechanisms to capture long-range dependencies and improve the representation of salient objects. Integrating non-local sparse attention modules into the super-resolution pipeline enhances the reconstruction and preserves fine details by effectively modelling the relationships between salient object regions and their surrounding context.

For the Non-Local Sparse Attention method, we utilized a pre-trained model. This approach involves leveraging a model that has been trained on a large dataset or specific task beforehand. In our case, we employed a pre-existing model that was trained on a relevant dataset or task, which captures the essential features and knowledge required for salient object super-resolution.

Using a pre-trained model offers several advantages. Firstly, it saves computational resources and time as we don’t need to train the model from scratch. Secondly, pre-trained models have already learned intricate patterns and representations from a vast amount of data, enabling them to generalize well to new unseen data. This can lead to improved performance and faster convergence during the testing phase.

By utilizing a pre-trained model for the Non-Local Sparse Attention method, we could take advantage of its learned knowledge and leverage it in the super-resolution task, enhancing the quality and accuracy of the results.

The Non-Local Sparse Attention (NLSA) aims to improve the performance of SISR models by leveraging the benefits of both non-local and sparse representations.

The key component of the NLSA approach is the Non-Local Sparse Attention module, which incorporates a dynamic sparse attention pattern. This module lets the model focus on important image features and exploit their correlations. By dynamically adjusting the sparse attention pattern, the NLSA model can adaptively capture long-range dependencies and selectively attend to relevant information while also reducing computational overhead.

In the context of SISR, the NLSA module is inserted after every eight residual blocks, as shown in figure 3.3. This placement allows the NLSA module to enhance the representations learned by the preceding layers and contribute to the overall super-resolution process.

By combining non-local operations and sparse attention in the NLSA approach, the model can effectively capture long-range dependencies and exploit the correlations between image features, leading to improved super-resolution results.

Incorporating attention mechanisms can selectively focus on areas with non-uniform degradation, one of the major challenges, directing the super-resolution process to the regions that require more attention and mitigate the issue.

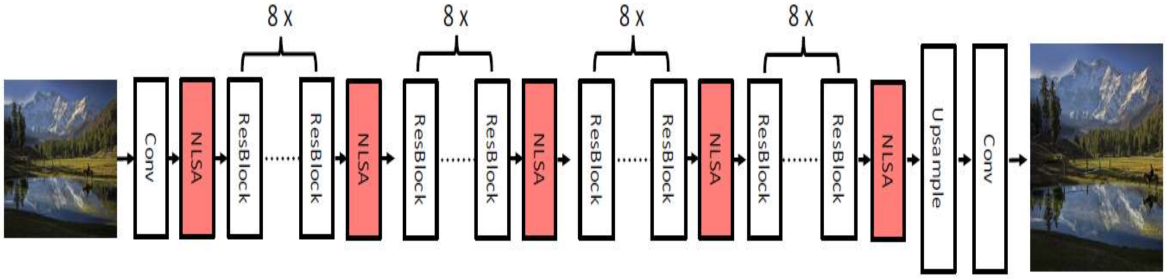


Figure 3.3: NLSN [26]

3.3.3 DRT

Although originally designed for deraining tasks, the Deraining Recursive Transformer (DRT) model [32] has been repurposed and trained for salient object super-resolution in our approach. This trained version of the DRT model is utilized further to enhance the quality and resolution of the salient object. By leveraging the capabilities of the DRT model, we aim to address any rain-induced blur or artefacts present in the salient object and restore its fine details.

We incorporate a specific network structure, as Figure 3.4 illustrates, for our experiments. The model consists of several stages: patch embedding, image reconstruction, deep feature extraction, and hierarchical feature representations.

In the patch embedding and image reconstruction stages, only one convolution layer is used without any activation function. This layer helps capture important features from the input images.

The deep feature extraction stage comprises six Recursive Transformer Blocks (RTBs), denoted as N = 6. Each RTB consists of three recursive calls (L = 3) on two Spatial Transformer Blocks (STBs). At the end of each RTB, a single convolution layer is employed without any activation function. It's worth noting that all the convolution operations maintain the input size, ensuring that there is no down-

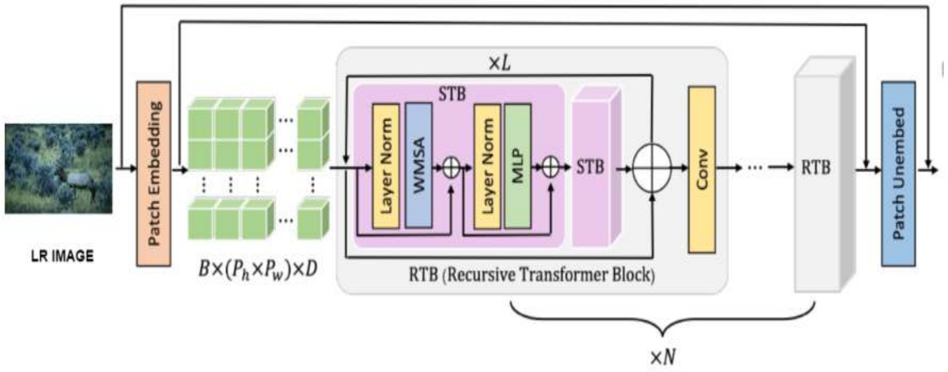


Figure 3.4: DRT [32]

scaling or upsampling of the image during the super-resolution process. This approach prevents the loss of pixel-level information.

Each STB in our model has a fixed local window dimension of 7×7 and a patch size of size 4. Additionally, each STB employs two attention heads to capture different aspects of the input.

The depth of the hierarchical feature representations, denoted as D, is set to 96. This depth value determines the level of abstraction and complexity in the hierarchical features learned by the model.

The recursive design of the DRT allows it to iteratively refine the super-resolution predictions. Each iteration updates the generated high-resolution output, progressively reducing ambiguity and uncertainty by incorporating more information from the low-resolution input and learned features.

3.4 Re-stacking Salient Object

After performing salient object super-resolution, we proceed to the restacking stage, where we recombine the enhanced salient object with the remaining parts of the image to create the final output. Re-stacking aims to integrate the high-resolution salient object back into its original context, ensuring that the overall image maintains a coherent and visually pleasing appearance.

In our approach, we utilize the saliency map's bounding rectangle to facilitate the restacking process of the salient object. The saliency map provides valuable information about the regions in the image that are deemed salient or attention-worthy. By extracting the bounding rectangle of the salient object from the saliency map, we obtain a tight and accurate representation of its spatial extent.

The bounding rectangle is a reference for aligning and repositioning the high-resolution salient object within the low-resolution image. By preserving the relative position and scale of the salient object, we ensure that it is reintegrated seamlessly into its original context.

We align the high-resolution salient object with the corresponding region in the low-resolution image to achieve restacking using the saliency map's bounding rectangle. This alignment can involve resizing the salient object to match the size of the bounding rectangle or translating it to the correct position within the image.

By utilizing the bounding rectangle, we can effectively merge the high-resolution salient object with the remaining parts of the image while maintaining its spatial coherence. This approach ensures that the salient object appears visually consistent and integrated within the overall image, enhancing the overall quality and realism of the final output.

3.5 Image Smoothing using Guided Image Filter

To achieve smoother blending and reduce artefacts around the boundaries of the salient object after the restacking stage, we employ guided image smoothening. In this process, we use the low-resolution image as guidance to guide the smoothing operation.

The low-resolution image means preserving important details and structures within the salient object while smoothing the sharp edges. By incorporating the guidance image, we can ensure that the smoothing process considers the desired level of detail and texture, resulting in a more natural and visually pleasing integration of the salient object.

The guided image smoothening [12] process involves applying a smoothing algorithm or filter to the salient object region while considering the guidance image. The algorithm considers the pixel values and features of both the salient object and the guidance image to determine the degree of smoothing. By adjusting the parameters of the smoothing algorithm, we can control the balance between preserving details and achieving a smoother transition.

The guided image filtering technique can be mathematically described as a weighted averaging process at each target image pixel; a local window is defined around it. The weights for the averaging process are determined based on the similarity between the pixel intensities in the guidance image and the pixels in the local window of the target image.

The benefits of guided image filtering include edge preservation and reduced artefacts compared to traditional filtering methods like Gaussian filtering. This makes it particularly useful for image enhancement, denoising, and tone-mapping tasks.

3.6 Proposed Model TraNLSN

In our research, we propose a hybrid model TraNLSN, that combines the Deraining Recursive Transformer (DRT) model and the Non-Local Sparse Attention mechanism. The DRT model, initially developed for deraining tasks, is adapted for super-resolution to restore fine details and textures within salient objects effectively.

In the context of salient object super-resolution, the challenges are distinct from deraining tasks. Super-resolution involves the reconstruction of high-resolution details within regions of interest, particularly salient objects, from low-resolution input images. This task demands the extraction of intricate details and the preservation of relevant information within salient regions.

The DRT model's strength in capturing long-range dependencies becomes highly relevant for salient object super-resolution, as it allows the model to understand the spatial relationships and context between different regions within an image. By incorporating the DRT model into our hybrid framework, we aim to exploit this capability to enhance the resolution of salient objects while considering their global context.

Furthermore, the DRT model's recursive nature enables it to iteratively refine feature representations iteratively, leading to more accurate and comprehensive image reconstructions. This iterative process can be highly beneficial in dealing with complex and diverse salient objects, as it allows the model to learn and refine the fine details within these regions iteratively.

By integrating the DRT model into our hybrid framework for salient object super-resolution, we expect to leverage its advanced feature extraction capabilities and recursive refinement process to enhance the resolution of salient objects in images effectively.

Additionally, we incorporate the Non-Local Sparse Attention mechanism into our hybrid model. This mechanism lets the model selectively attend to the input image's relevant spatial and channel information. By attending to specific areas and channels, the model can better preserve important features while suppressing noise and artefacts that may be present in the low-resolution input.

The combination of the DRT model and the Non-Local Sparse Attention mechanism results in a hybrid model that excels at the super-resolution of salient objects. The DRT model provides the foundation for capturing detailed information, while the Non-Local Sparse Attention mechanism enhances the model's ability to focus selectively on important image features.

Our hybrid model offers a comprehensive solution for the super-resolution of low-resolution images, effectively restoring fine details and textures and preserving important features while mitigating noise and artefacts.

In the super-resolution process of our proposed model TraNLSN, we have designed a specific architecture, as depicted in Figure 3.5. This architecture comprises several key blocks, including patch embedding, RTB (Deraining Recursive Transformer Block), patch unembedding, NLSA (Non-Local Sparse Attention) upsample, and more. Let's delve into the details of each block.

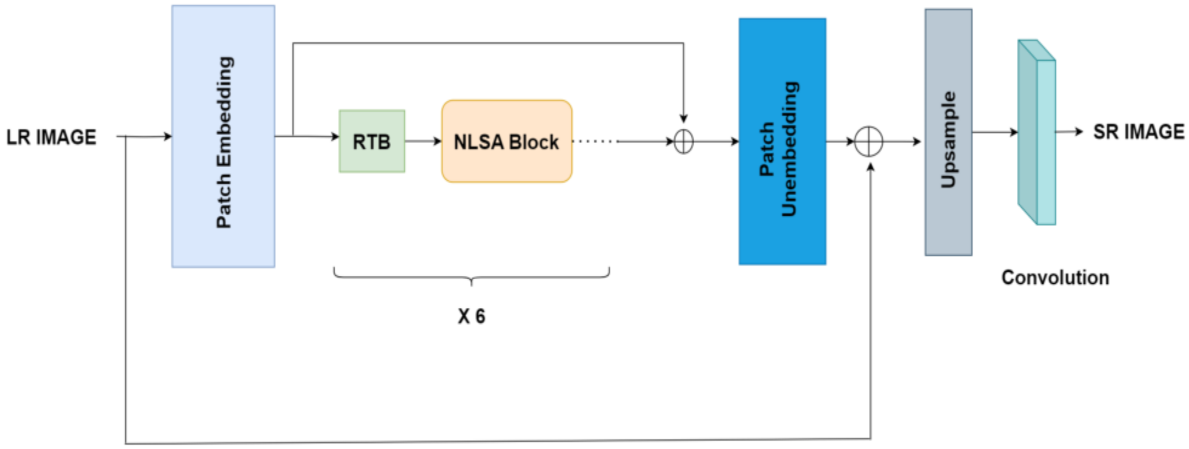


Figure 3.5: Architecture for TraNLSN

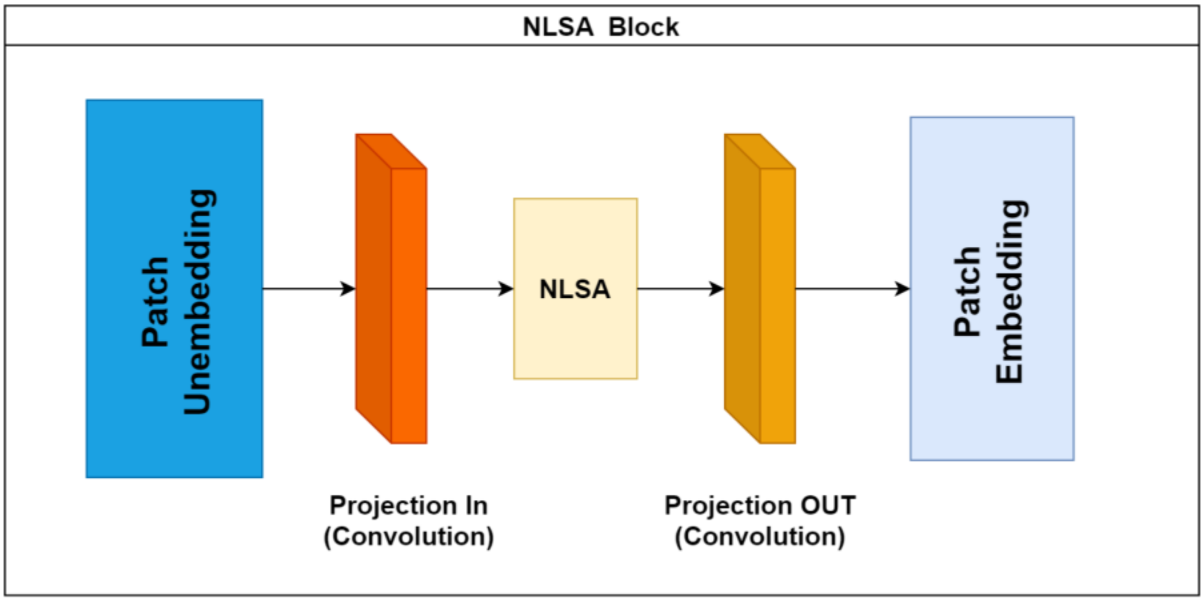


Figure 3.6: NLSA Block

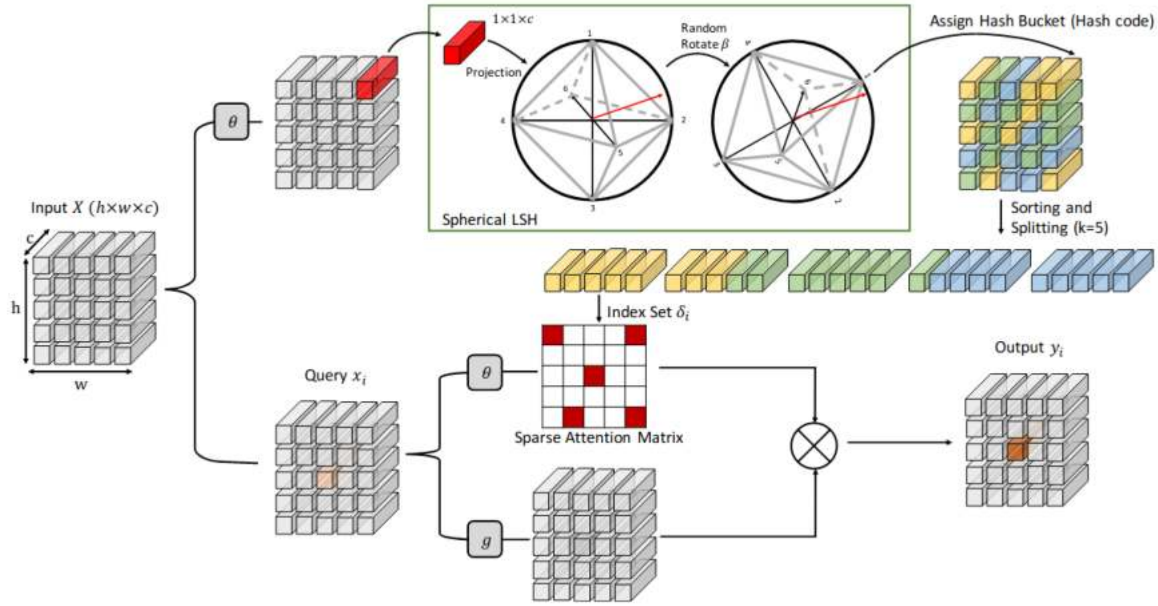


Figure 3.7: Non-Local Sparse Attention (NLSA) [39]

To integrate the Deraining Recursive Transformer (DRT) model and the Non-Local Sparse Attention (NLSN) model, we introduced the NLSA Block as shown in Figure 3.6. This block allows for the fusion of the NLSN attention modules with the transformer architecture by incorporating patch unembedding, projection convolution, the original NLSN attention block, projection output convolution, and patch embedding layers. The NLSA Block seamlessly integrates the NLSN attention mechanism, as shown in Figure 3.7, into the transformer architecture.

The overall architecture of our hybrid model, TraNLSN, is constructed by alternating the Recursive Transformer Blocks (RTBs) and NLSA Blocks. This arrangement is repeated N = 6 times to form the final architecture. Each RTB consists of three recursive calls on two Spatial Transformer Blocks (STBs), followed by a convolution layer without activation. The STBs have a fixed local window dimension of 7×7 and a patch size of 4. We also utilize two attention heads to capture different aspects of the input data.

The hierarchical feature representations in TraNLSN have a depth of 96, determining the level of abstraction and complexity in the learned features. This depth allows the model to capture intricate details and patterns in the image, contributing to more accurate super-resolution results.

The model incorporates two skip connections: an inner skip connection after the first patch embedding, which is combined with the NLSA block output, and an outer skip connection before the first patch embedding, added at the post-patch embedding stage. An upsample and convolution block are then applied to upsample the image to the desired scale and generate the final output.

By combining the strengths of the DRT model and the NLSN attention mechanism in this modified architecture, TraNLSN aims to leverage the benefits of both approaches for super-resolution tasks, enhancing the restoration of fine details, preservation of important features, and suppression of noise and artefacts.

CHAPTER 4

Experiment and Results

In this section, we describe the experimental setup and present the results obtained from conducting salient object super-resolution experiments using three different methods. Each method follows the same five basic steps, the only difference being the choice of super-resolution architecture.

4.1 Datasets

We utilized three benchmark datasets for our experiments: SET5 [4], SET14 [53], and BSDS100 [38]. SET5, SET14 and BSDS100 consist of five, fourteen and a hundred images, respectively, commonly used for image super-resolution evaluation, while All these datasets are benchmark datasets for testing super-resolution algorithms.

4.2 Methodology

4.2.1 Method-1 (SRGAN)

Experimental setup for SRGAN Model

During the training phase of the SRGAN model, the following steps are undertaken:

1. **Dataset Preparation:** The training dataset is prepared by selecting appropriate image datasets for super-resolution tasks, such as MIRFLICKR [14], Set5, or Set14. The images in the dataset are resized to a specific resolution and divided into patches of suitable sizes. Data augmentation techniques, such as random flipping or rotation, may be applied to increase the diversity of the training samples.

2. **Model Configuration:** The SRGAN model architecture, as described earlier, is implemented using a deep learning framework. The model's hyperparameters are set based on prior research and empirical analysis. The optimizer, such as Adam, is chosen for training the model.

3. **Loss Function:** During the training phase of the SRGAN model, the loss function plays a crucial role in guiding the model's learning process. We utilise a perceptual loss function instead of using the traditional mean squared error (MSE) loss commonly employed in super-resolution tasks.

   The perceptual loss function is based on the concept of perceptual similarity, which aims to measure the similarity between images based on high-level perceptual features rather than pixel-level differences. This approach is motivated by the understanding that humans perceive images based on their visual appearance and structure rather than the exact pixel values.

   We employ a pre-trained deep neural network to compute the perceptual loss, such as VGG-19 [43], which has been trained on large-scale image classification tasks. The network is utilized as a feature extractor to capture high-level features from both the generated high-resolution images and the corresponding ground truth images.

   The perceptual loss is calculated as the mean squared error between the feature representations of the generated and ground truth images at multiple layers of the VGG-19 network. The model is encouraged to generate images with similar pixel values and high-level visual characteristics by comparing the feature maps at different levels.

   This perceptual loss function enables the SRGAN model to focus on capturing and reproducing important perceptual features in the output images, such as textures, edges, and structures. The model learns to produce visually pleasing and perceptually accurate high-resolution images by optimising this loss during training.

   The perceptual loss function is defined as

   $$L_{per} = L_c + 10^{-3} L_{adv} \tag{4.1}$$

   where L_c is content loss and L_{adv} is adversarial loss.

The following is how this loss is defined: Content loss 4.2:

$$p^{sr}_{mse} = \frac{1}{r^2wh} \sum_{x=1}^{rw} \sum_{y=1}^{rh} (I^{hr}_{x,y} - G_{\theta_g}(I^{lr})_{x,y}) \tag{4.2}$$

VGG content loss 4.3:

$$p^{sr}_{vgg/i,j} = \frac{1}{w_{i,j}h_{i,j}} \sum_{x=1}^{w_{i,j}} \sum_{y=1}^{h_{i,j}} (\phi_{i,j}(I^{hr}_{x,y} - \phi_{i,j}(G_{\theta_g}(I^{lr})_{x,y}))) \tag{4.3}$$

Here r is the downsampling factor,$w_{i,j}$ and $h_{i,j}$ describe the dimensions of the respective feature maps in the VGG network.And $\phi_{i,j}$ indicate the feature map obtained by the j^{th} convolution(after the activation) and before the i^{th} max-pooling layer within the VGG19 network.

The adversarial loss can be calculated as

$$L_{adv} = \sum_{n=1}^{N} -logH_{\theta_H}(J_{\theta_J}(I^{LR})) \tag{4.4}$$

where $H_{\theta_H}(J_{\theta_J}(I^{LR}))$ is the probability that the super-resolution image $J_{\theta_J}(I^{LR})$ is a natural HR image.

In summary, using a perceptual loss function enhances the training process of the SRGAN model by promoting the generation of images that closely resemble ground truth images in terms of high-level perceptual features. This leads to improved perceptual quality and visual fidelity in the super-resolution output.

4. **Training Iterations:** The SRGAN model is trained using the prepared dataset and loss function. Training is performed in iterations, where each iteration involves feeding a batch of training samples to the model. The Adam optimizer updates the model's parameters based on the computed gradients, gradually improving the model's performance.

By following these steps, the SRGAN model undergoes the training phase, gradually learning to extract relevant features and enhance the resolution of input images. The model's performance is continuously evaluated and refined during this phase to achieve the desired super-resolution capabilities.

Training Setup

In the SRGAN scratch training process, we trained the model using patches of size 96×96 extracted from the Image-Celeb dataset. The training was performed with a batch size of 16 and an Adam optimizer. The learning rate was set to 0.00008, and we applied Adam's decay with a first-order momentum decay (b1) of 0.5 and a second-order momentum decay of 0.999. The training process was performed for a total of 100 epochs for a scale factor of ×4.

Since the SRGAN model typically requires a large training dataset (around 350,000 images), but we had limited training samples available, we applied data augmentation techniques to augment our dataset. Specifically, we employed random horizontal flipping with a probability of 0.5 and random rotation of 90° to increase the diversity and variability of our training data.

By incorporating data augmentation, we aimed to enhance the generalization capability of the SRGAN model and improve its performance in handling variations and complexities present in salient object super-resolution tasks.

In our experiment, where the low-resolution images are upsampled, we modified the SRGAN architecture to adapt it to our objective. The SRGAN architecture is primarily designed for single-image super-resolution, which involves enhancing the resolution of low-resolution images.

To align the SRGAN architecture with our objective of salient object super-resolution, where the low-resolution image is already upsampled, we removed the upsampling block from the SRGAN architecture. The upsampling block is responsible for increasing the resolution of the input image.

Removing the upsampling block ensured that the modified SRGAN architecture was suitable for our specific scenario, where the low-resolution images were already upsampled. This modification allowed us to focus solely on the remaining components of the SRGAN architecture, such as the generator and discriminator networks, which play crucial roles in enhancing the quality of the salient object super-resolution results.

Table 4.1 illustrates the sequential images depicting the various stages involved in Method-1.

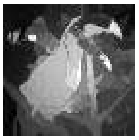
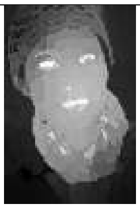
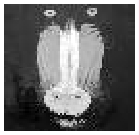
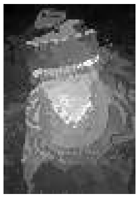
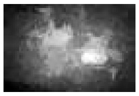
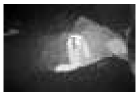
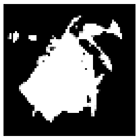
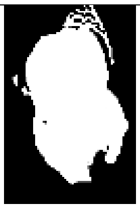
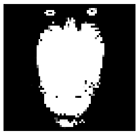
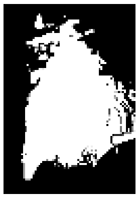
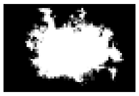
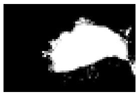
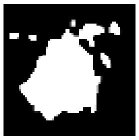
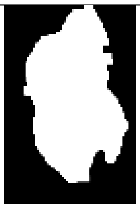
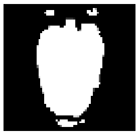
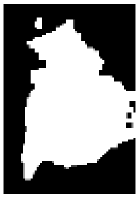
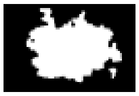
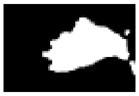
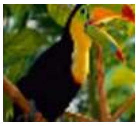
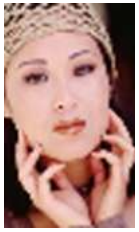
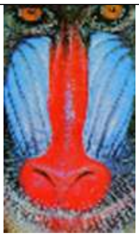
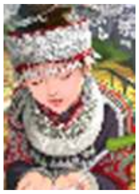
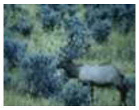
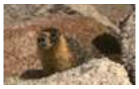
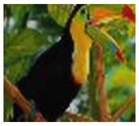
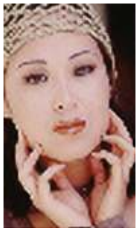
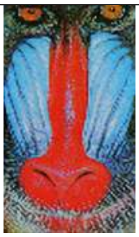
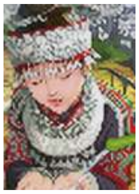
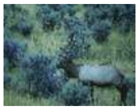
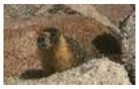
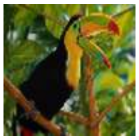
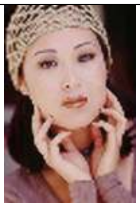
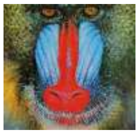
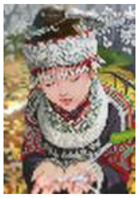
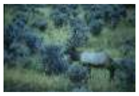
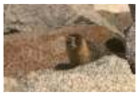
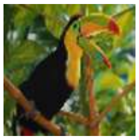
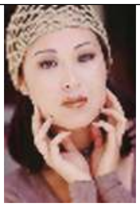
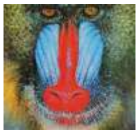
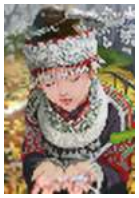
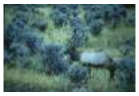
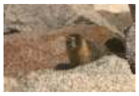
Method	Set5		Set14		BSDS100	
LR Image						
Map						
Thresh						
Fill						
Segment						
SOS						
Restack						
Smooth						

Table 4.1: Stagewise Method-1 depiction

4.2.2 Method-2 (NLSN)

Experimental setup for NLSN Model

In the experimental setup for the Non-Local Sparse Attention (NLSA) model, we utilized a pre-trained model to evaluate its performance in the super-resolution task. Here are the details of the setup:

1. **Pre-Trained Model:** We obtained a pre-trained NLSA model that was already trained on a large-scale dataset, such as DIV2K [2], which contains high-resolution images. The model was trained using appropriate optimization techniques and loss functions to learn the super-resolution task effectively.

2. **Datasets:** To evaluate the performance of the pre-trained NLSA model, we selected benchmark datasets commonly used in super-resolution, such as Set5, Set14, and Urban100 [13]. These datasets consist of low-resolution images along with their corresponding high-resolution ground truth images. The low-resolution images were used as input to the pre-trained NLSA model for super-resolution.

3. **Preprocessing:** We performed the necessary preprocessing steps before feeding the low-resolution images into the pre-trained NLSA model. This involved resizing the low-resolution images to match the desired scale and applying any required transformations or augmentations to enhance the quality of the inputs.

4. **Evaluation Metrics:** To assess the quality of the super-resolved images generated by the pre-trained NLSA model, we used various evaluation metrics. Commonly used metrics such as Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) were calculated by comparing the super-resolved images with their corresponding ground truth high-resolution images. These metrics provided quantitative measures of the similarity between the super-resolved and ground truth images.

5. **Comparison to Baseline Models:** In addition to evaluating the pre-trained NLSA model, we compared its performance to baseline models and state-of-the-art super-resolution techniques. The baseline models could include traditional interpolation methods like bicubic interpolation, while state-of-the-art techniques could involve other deep learning-based models or advanced algorithms. The comparison was made based on the evaluation metrics to

assess the superiority of the pre-trained NLSA model in generating high-quality super-resolved images.

By utilizing the pre-trained NLSA model and following this experimental setup, we were able to assess its performance in the super-resolution task without the need for training from scratch. This approach saves computational time and resources while still allowing us to evaluate the capabilities and effectiveness of the pre-trained model.

Testing Setup

We comprehensively evaluated the Non-Local Sparse Attention (NLSA) model during our research's testing phase. This evaluation aimed to assess the model's performance in super-resolution tasks. The testing setup involved the utilization of a pre-trained NLSA model, which had been trained on a large dataset.

The NLSA model was set to inference mode during testing to ensure consistent evaluation. This mode disabled any training-specific operations and ensured that the model behaved consistently throughout the evaluation process. The model was loaded with pre-trained weights, enabling it to produce reliable and accurate results.

The attention mechanism of the NLSA model played a crucial role in its performance. For testing purposes, we set the attention bucket size (also known as the chunk size) to 144. Additionally, the number of hashing rounds was set to 4, which determined how the attention mechanism operated and contributed to the super-resolution process.

The NLSA model was built upon an Enhanced Deep Super-Resolution (EDSR) backbone with 32 residual blocks. This architecture, combined with the specific hyperparameters used during training, formed the basis of the model's structure during testing. The convolutional layers within the model had a kernel size of 3x3, and the intermediate features had 256 channels, consistent with the training configuration.

During testing, the NLSA model was designed to upscale the resolution of input images by a scale factor of 4. This meant that the model aimed to increase the level of detail and clarity in the images, producing high-resolution outputs that were four times larger than the input resolution.

The output configuration of the NLSA model was carefully considered. The final convolutional layer transformed the deep features into RGB images with three channels, ensuring compatibility with the evaluation metrics and comparison with the ground truth high-resolution images.

Throughout the testing phase, we inputted preprocessed low-resolution images into the NLSA model and evaluated the model's ability to generate super-resolved images with improved resolution. The performance of the NLSA model was assessed by comparing the generated high-resolution images against the corresponding ground truth high-resolution images, using appropriate evaluation metrics.

For NLSN we used L_1 loss which can be given in equation:

$$L_{NLSN}(G) = \frac{1}{N}\sum_{i=1}^{N} ||G(x_i) - y_i|| \tag{4.5}$$

where G is the NLSN model, x_i are the input data samples, y_i are the corresponding ground truth samples and N is the total no. of samples.

Table 4.2 illustrates the sequential images depicting the various stages involved in Method-2. Each image represents a specific step in the overall process.

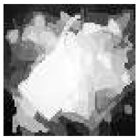
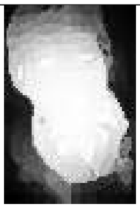
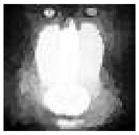
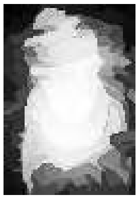
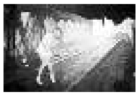
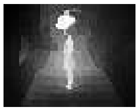
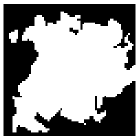
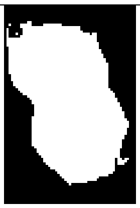
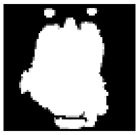
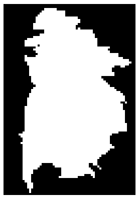
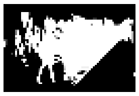
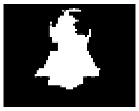
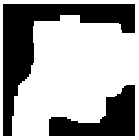
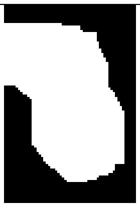
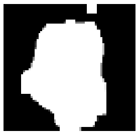
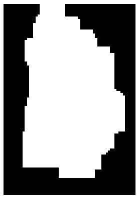
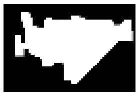
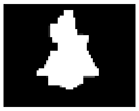
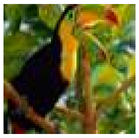
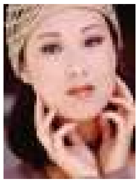
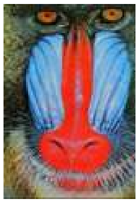
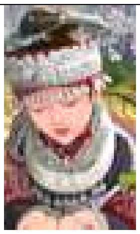
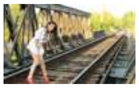
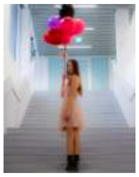
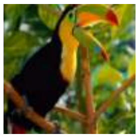
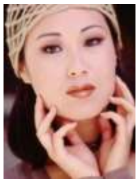
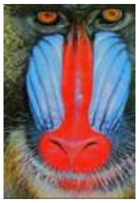
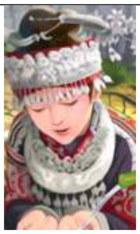
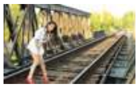
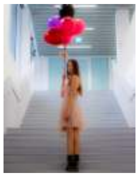
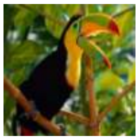
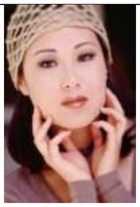
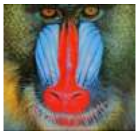
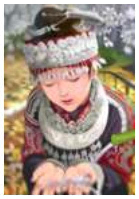
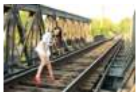
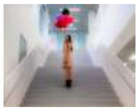
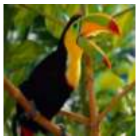
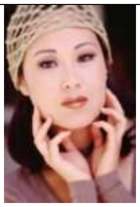
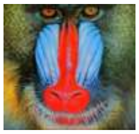
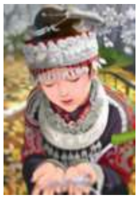
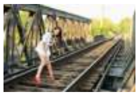
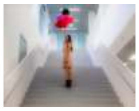
Method	Set5		Set14		Urban100	
LR Image						
Map						
Thresh						
Fill						
Segment						
SOS						
Restack						
Smooth						

Table 4.2: Stagewise Method-2 depiction

4.2.3 Method-3 (DRT)

Experimental Setup for DRT

During the training phase of the DRT (Deraining Recursive Transformer) model, the following steps are undertaken:

1. **Dataset Preparation:** The training dataset is prepared by selecting appropriate image datasets for super-resolution tasks, such as DIV2K, Set5, or Set14. The images in the dataset are resized to a specific resolution and divided into patches of suitable sizes. Data augmentation techniques, such as random flipping or rotation, may be applied to increase the diversity of the training samples.

2. **Model Configuration:** The DRT model architecture, as described earlier, is implemented using a deep learning framework. The model's hyperparameters, including the number of RTBs, STBs, attention heads, and patch size, are set based on prior research and empirical analysis. The optimizer, such as Adam, is chosen for training the model.

3. **Loss Function:** The mean squared error (MSE) loss function is commonly used for super-resolution tasks. It calculates the pixel-wise difference between the output image generated by the model and the corresponding high-resolution ground truth image. The goal is to minimize this loss during training. The loss function L_{DRT} can be given as:

   $$L_{DRT}(T) = \frac{1}{N} \sum_{i=1}^{N} ||T(x_i) - y_i||^2 \tag{4.6}$$

   where T is the DRT model, x_i is the input data samples, y_i is the ground truth samples and N is total no. of data samples.

4. **Training Iterations:** The DRT model is trained using the prepared dataset and loss function. Training is performed in iterations, where each iteration involves feeding a batch of training samples to the model. The Adam optimizer updates the model's parameters based on the computed gradients, gradually improving the model's performance.

5. **Validation:** During the training process, a separate validation dataset is used to monitor the model's performance and prevent overfitting. The validation dataset consists of high-resolution images, and the model's output

is compared with the ground truth using evaluation metrics such as PSNR and SSIM. This helps in selecting the best-performing model and adjusting the hyperparameters if necessary.

6. **Testing and Evaluation:** Once the training is completed, the trained DRT model is evaluated on a separate testing dataset. The low-resolution test images are fed into the model, which generates high-resolution output images. The performance of the model is assessed using quantitative metrics like PSNR and SSIM, which measure the similarity between the generated images and the ground truth. Visual inspection and qualitative assessment of the output images are also conducted to assess the perceptual quality and the model's ability to enhance image resolution.

7. **Comparison with Baselines:** The performance of the DRT model is compared with other state-of-the-art super-resolution methods, including EDSR, SRGAN, and others. The comparison is made in terms of quantitative metrics (PSNR, SSIM) and qualitative evaluation to demonstrate the superiority and effectiveness of the proposed model.

By following these steps, the DRT model undergoes the training phase, gradually learning to extract relevant features and enhance the resolution of input images. During this phase, the model's performance is continuously evaluated and refined to achieve the desired super-resolution capabilities.

Training Setup

During the training process of the Deraining Recursive Transformer (DRT) model, we followed a specific setup to optimize its performance for super-resolution tasks. We extracted non-overlapping high-resolution patches of size 56×56 from the training dataset as the target output and downsampled it to the size 14×14, then again upsampled it to the same size 56×56, which served as the input for the model. We utilized the Adam optimizer, a widely used optimization algorithm in deep learning to optimise the model's parameters.

The training was performed using a batch size of eight, meaning that eight patches were processed simultaneously during each training iteration. This batch size strikes a balance between computational efficiency and the model's ability to learn from diverse examples.

To measure the discrepancy between the model's output and the target high-resolution image, we employed the mean squared error (MSE) loss function. The

MSE loss calculates the average squared difference between the predicted and target values, providing a measure of how well the model approximates the ground truth image.

The initial learning rate was set to 1e-4, a common choice for training deep learning models. The learning rate determines the step size at each optimization iteration and influences the convergence of the model.

We utilized an NVIDIA RTX GEFORCE 2080 Ti GPU to accelerate the training process, which offers high-performance computing capabilities suitable for training deep neural networks. The GPU's parallel processing capabilities expedite the model's forward and backward computations, resulting in faster training times.

By following this training setup, we aimed to optimize the DRT model's parameters, enabling it to learn and extract relevant features from the input patches and enhance its performance in super-resolution tasks.

Table 4.3 illustrates the sequential images depicting the various stages involved in Method-3. Each image represents a specific step in the overall process.

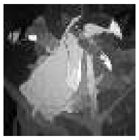
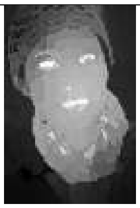
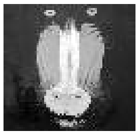
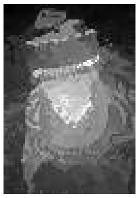
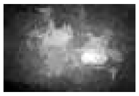
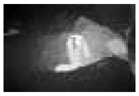
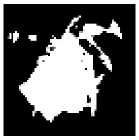
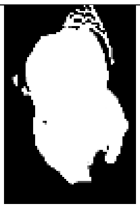
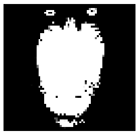
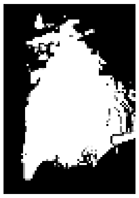
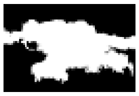
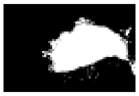
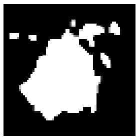
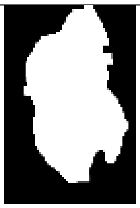
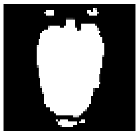
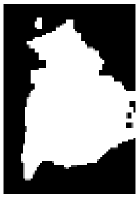
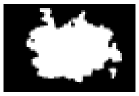
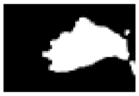
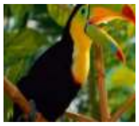
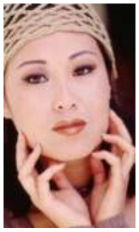
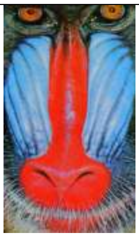
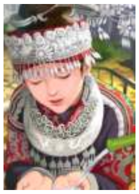
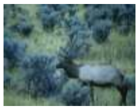
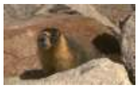
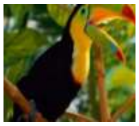
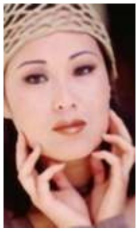
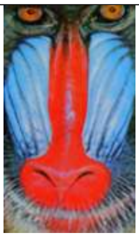
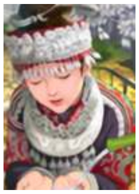
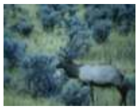
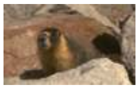
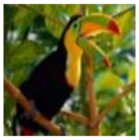
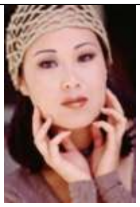
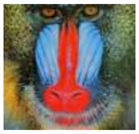
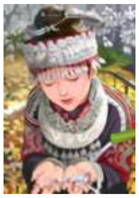
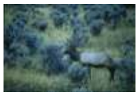
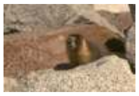
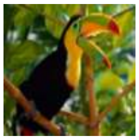
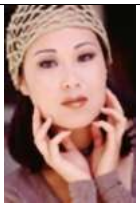
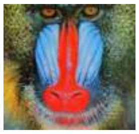
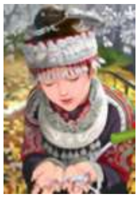
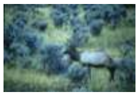
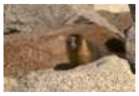
Method	Set5		Set14		BSDS100	
LR Image						
Map						
Thresh						
Fill						
Segment						
SOS						
Restack						
Smooth						

Table 4.3: Stagewise Method-3 depiction

4.2.4 Proposed Model (TraNLSN)

Experimental Setup

During the training phase of the TraNLSN (Transformer with Non-Local Sparse Attention) model, the following steps are undertaken:

1. **Dataset Preparation:** The training dataset is prepared by selecting appropriate image datasets for super-resolution tasks, such as DIV2K, Set5, or Set14. The images in the dataset are resized to a specific resolution and divided into patches of suitable sizes. Data augmentation techniques, such as random flipping or rotation, may be applied to increase the diversity of the training samples.

2. **Model Configuration:** The TraNLSN model architecture, as described earlier, is implemented using a deep learning framework. The model's hyperparameters, including the number of RTBs, STBs, attention heads, and patch size, are set based on prior research and empirical analysis. The optimizer, such as Adam, is chosen for training the model.

3. **Loss Function:** The mean squared error (MSE) loss function is commonly used for super-resolution tasks. It calculates the pixel-wise difference between the output image generated by the model and the corresponding high-resolution ground truth image. The goal is to minimize this loss during training as mentioned in equation 4.6.

4. **Training Iterations:** The TraNLSN model is trained using the prepared dataset and loss function. Training is performed in iterations, where each iteration involves feeding a batch of training samples to the model. The Adam optimizer updates the model's parameters based on the computed gradients, gradually improving the model's performance.

By following these steps, the TraNLSN model undergoes the training phase, gradually learning to extract relevant features and enhance the resolution of input images. During this phase, the model's performance is continuously evaluated and refined to achieve the desired super-resolution capabilities.

Training Setup

During the training process of the TraNLSN model, we followed a specific setup to optimize its performance for super-resolution tasks. We extracted non-overlapping high-resolution patches of size 64 × 64 from the training dataset as the target output and downsampled it to the size 16 × 16, which served as the input for the model. We utilized the Adam optimizer, a widely used optimization algorithm in deep learning to optimise the model's parameters.

The training was performed using a batch size of eight, meaning that eight patches were processed simultaneously during each training iteration. This batch size strikes a balance between computational efficiency and the model's ability to learn from diverse examples.

To measure the discrepancy between the model's output and the target high-resolution image, we employed the mean squared error (MSE) loss function. The MSE loss calculates the average squared difference between the predicted and target values, measuring how well the model approximates the ground truth image.

The initial learning rate was set to 1e-4, a common choice for training deep learning models. The learning rate determines the step size at each optimization iteration and influences the convergence of the model.

We utilized an NVIDIA RTX GEFORCE 2080 Ti GPU to accelerate the training process, which offers high-performance computing capabilities suitable for training deep neural networks. The GPU's parallel processing capabilities expedite the model's forward and backward computations, resulting in faster training times.

By following this training setup, we aimed to optimize the TraNLSN model's parameters, enabling it to learn and extract relevant features from the input patches and enhance its performance in super-resolution tasks.

4.3 Evaluation

To assess the effectiveness of our proposed method, we employed standard evaluation metrics commonly used in the respective domains. Specifically, we utilized two widely used metrics for image super-resolution tasks: peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM).

PSNR is a metric that measures the quality of the reconstructed image by comparing it to the ground truth. It quantifies the ratio between the maximum possible power of a signal (in this case, the original high-resolution image) and the power of the noise or distortion in the reconstructed image. Higher PSNR values indicate better reconstruction quality, indicating a smaller difference between the reconstructed image and the ground truth.

SSIM, on the other hand, measures the structural similarity between two images by considering their luminance, contrast, and structural information. It evaluates the perceived similarity of the reconstructed image with respect to the ground truth, taking into account both global and local image features. SSIM values range from -1 to 1, with a value of 1 indicating a perfect match between the reconstructed image and the ground truth.

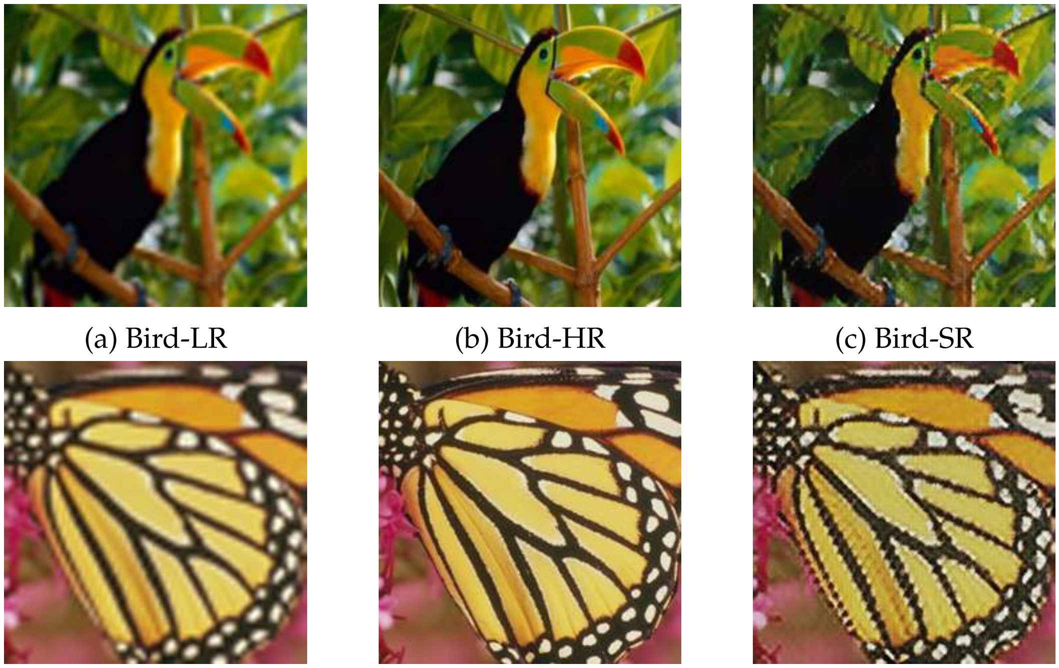
By using these metrics, we aim to quantitatively evaluate the performance of our proposed method and compare it with other state-of-the-art approaches. These metrics provide valuable insights into the quality and similarity of the generated high-resolution images, helping us assess our method's effectiveness in enhancing image resolution.

4.4 Results

4.4.1 Method-1 (SRGAN)

Super-Resolution Model Results (SRGAN)

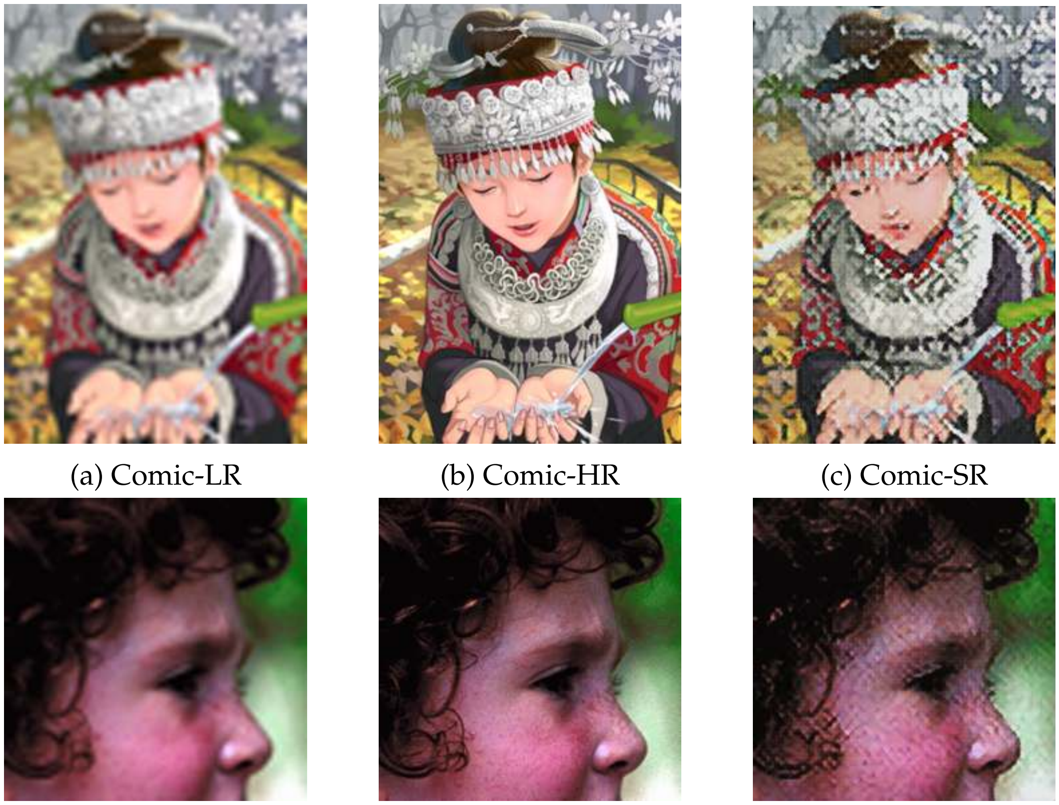
Following figures 4.1,4.2 and 4.3 contain low-resolution input images, high-resolution reference images, and super-resolution output images of the respective dataset.



(a) Bird-LR (b) Bird-HR (c) Bird-SR

(d) Butterfly-LR (e) Butterfly-HR (f) Butterfly-SR

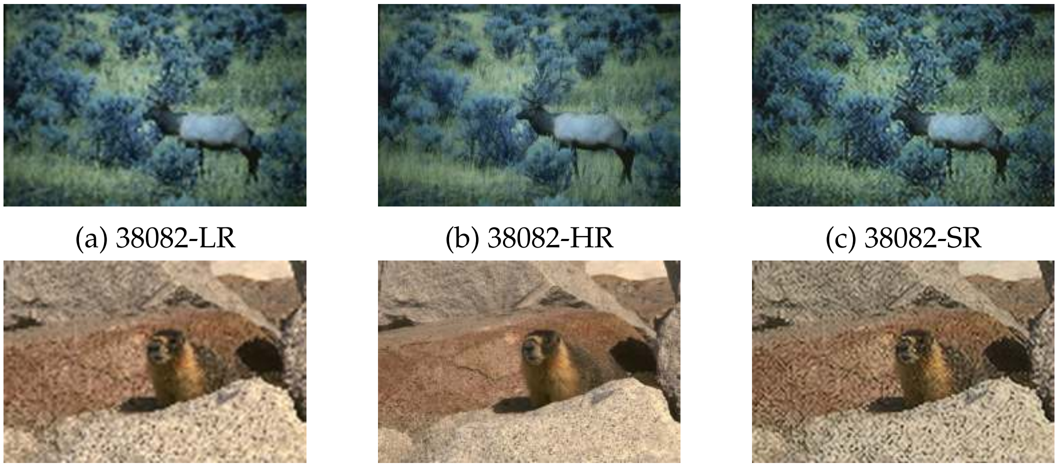
Figure 4.1: SET5 (SRGAN)



(a) Comic-LR (b) Comic-HR (c) Comic-SR

(d) Face-LR (e) Face-HR (f) Face-SR

Figure 4.2: SET14 (SRGAN)



(a) 38082-LR (b) 38082-HR (c) 38082-SR

(d) 41069-LR (e) 41069-HR (f) 41069-SR

Figure 4.3: BSDS100 (SRGAN)

Proposed Method Results

Table 4.4 presents a breakdown of the time taken in seconds at each stage in Method-1. The table provides an analysis of the time duration for each step, highlighting the relative time consumption for different stages of the method.

Method	Set5(s)	Set14(s)	BSDS100(s)
Map Generation	3.07	4.19	2.94
Thresholding	0.008	0.004	0.004
Hole Filling	0.013	0.003	0.003
Segmentation	0.008	0.011	0.010
SOS	**0.099**	**0.157**	**0.104**
Restacking	0.015	0.019	0.015
Smoothing	3.27	6.69	4.56
Total Time	6.483	11.074	7.636
Entire SR	**0.148**	**0.252**	**0.193**

Table 4.4: Stagewise Time Analysis of Different Stages for Method-1

Table 4.5 shows a comparative analysis of our Method-1 with state-of-the-art methods using two evaluation metrics: PSNR (Peak Signal-to-Noise Ratio) and SSIM (Structural Similarity Index). The table provides a comparison of the performance of different methods in terms of these metrics, indicating the quality of the output generated by each method.

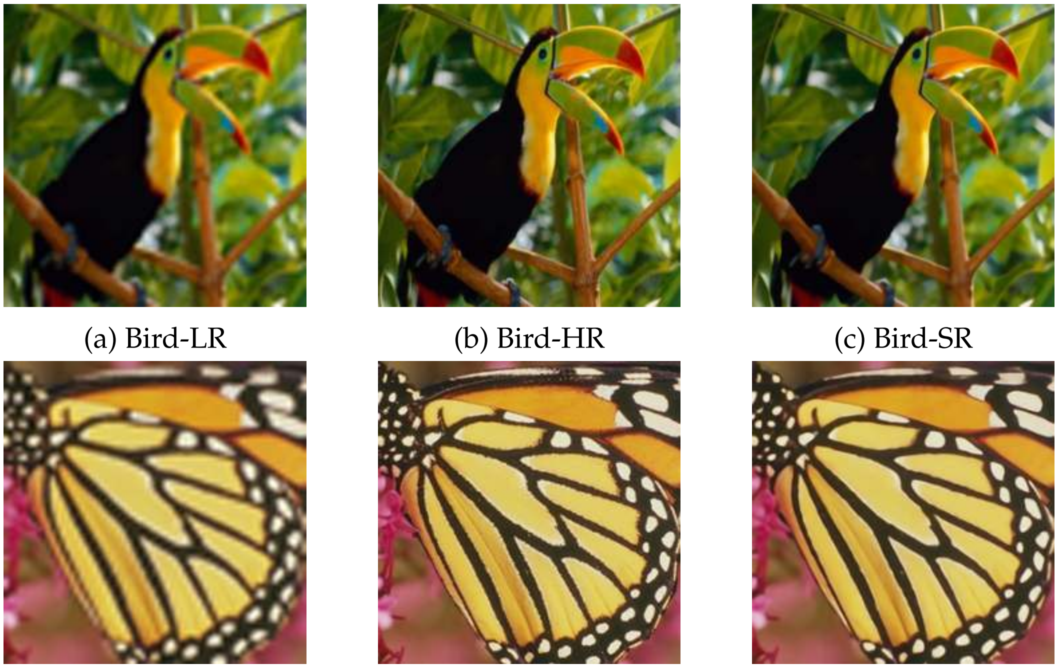
Method	Scale	Set5		Set14		BSDS100	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Bicubic	×4	28.42	0.8104	26.00	0.7027	25.96	0.6675
SRCNN [46]	×4	30.48	0.8628	27.50	0.7513	26.90	0.7101
VDSR [25]	×4	31.35	0.8830	28.02	0.7680	27.29	0.7026
EDSR [33]	×4	32.46	0.8968	28.80	0.7876	27.71	0.7420
NLRN [34]	×4	31.92	0.8916	28.36	0.7745	27.48	0.7306
RNAN [55]	×4	32.49	0.8982	28.83	0.7878	27.72	0.7409
SRFBN [31]	×4	32.47	0.8983	28.81	0.7868	27.72	0.7409
RDN [56]	×4	32.47	0.8990	28.81	0.7871	27.72	0.7419
RCAN [26]	×4	32.63	0.9002	28.87	0.7889	27.77	0.7436
NLSN [39]	×4	32.59	0.9000	28.87	0.7891	27.78	0.7444
Entire SR	×4	22.59	0.6921	20.77	0.5907	20.91	0.5587
Cropped SR	×4	22.36	0.7022	20.31	0.5820	20.07	0.5344
SOS	×4	24.08	0.7518	21.94	0.6550	21.72	0.6273
Entire SR (Y)	×4	23.03	0.7112	21.15	0.6087	20.96	0.5629
Cropped SR (Y)	×4	22.90	0.7240	20.73	0.6009	20.10	0.5391
SOS (Y)	×4	24.72	0.7802	22.33	0.6760	21.78	0.6345

Table 4.5: Quantitative analysis of different Architectures (Scale 4) with Method-1

4.4.2 Method-2 (NLSN)

Super-Resolution Model Results (NLSN)

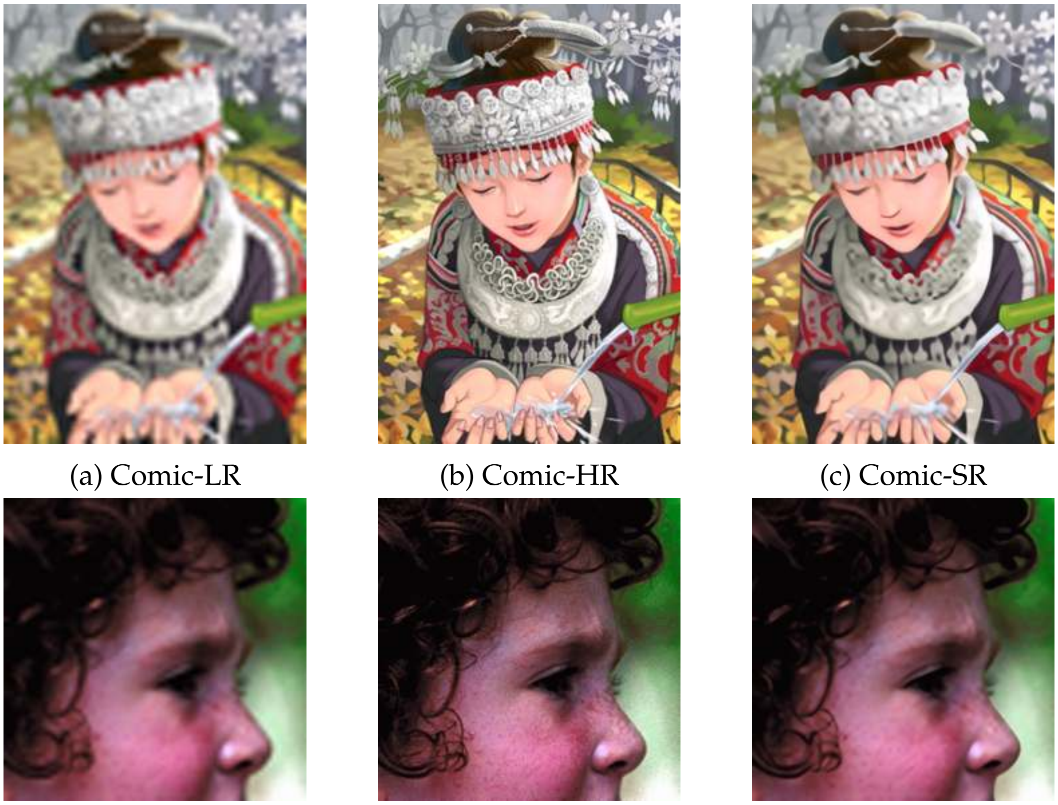
Following figures 4.4,4.5 and 4.6 contain low-resolution input images, high-resolution reference images, and super-resolution output images of the respective dataset.



(a) Bird-LR (b) Bird-HR (c) Bird-SR

(d) Butterfly-LR (e) Butterfly-HR (f) Butterfly-SR

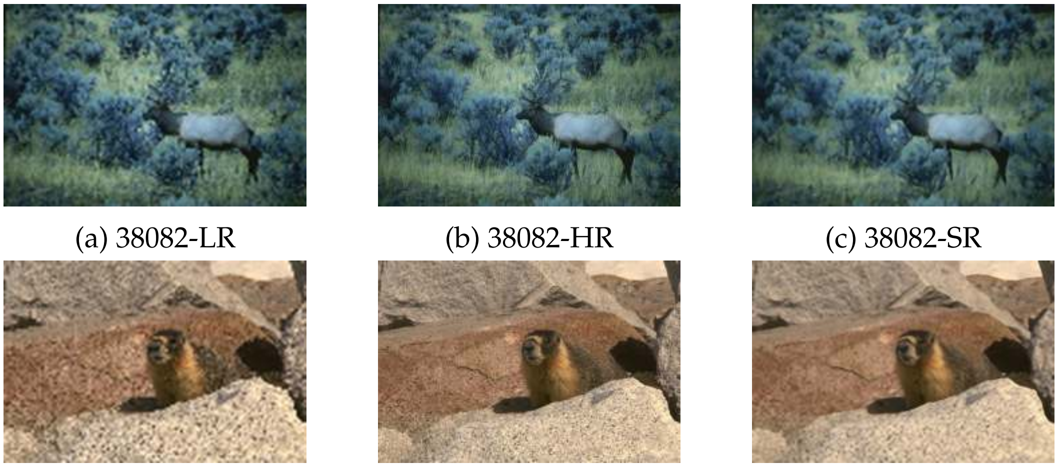
Figure 4.4: SET5 (NLSN)



(a) Comic-LR (b) Comic-HR (c) Comic-SR

(d) Face-LR (e) Face-HR (f) Face-SR

Figure 4.5: SET14 (NLSN)



(a) 38082-LR (b) 38082-HR (c) 38082-SR

(d) 41069-LR (e) 41069-HR (f) 41069-SR

Figure 4.6: BSDS100 (NLSN)

Proposed Method Results

Table 4.6 presents a breakdown of the time taken at each stage in Method-2. The table provides an analysis of the time duration for each step, highlighting the relative time consumption for different stages of the method.

Method	Set5(s)	Set14(s)	Urban100(s)
Map Generation	2.98	2.79	3.38
Thresholding	0.004	0.003	0.005
Hole Filling	0.011	0.005	0.005
Segmentation	0.007	0.007	0.009
SOS	**3.85**	**1.78**	**1.35**
Restacking	0.92	1.07	1.47
Smoothing	0.95	0.93	0.94
Total Time	8.722	6.585	7.159
Entire SR	**3.94**	**2.06**	**1.82**

Table 4.6: Stagewise Time Analysis of Different Stages for Method-2

Table 4.7 shows a comparative analysis of our Method-2 with state-of-the-art methods using two evaluation metrics: PSNR (Peak Signal-to-Noise Ratio) and SSIM (Structural Similarity Index). The table provides a comparison of the performance of different methods in terms of these metrics, indicating the quality of the output generated by each method.

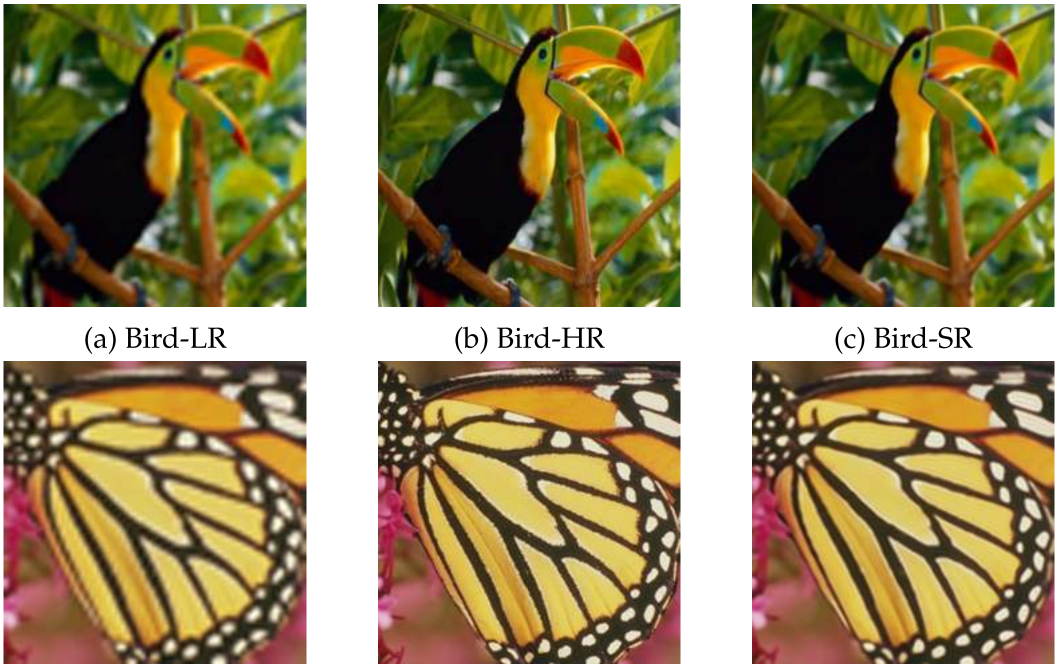
Method	Scale	Set5(s)		Set14(s)		Urban100(s)	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Bicubic	×4	28.42	0.8104	26.00	0.7027	23.14	0.6577
SRCNN [46]	×4	30.48	0.8628	27.50	0.7513	24.52	0.7221
VDSR [25]	×4	31.35	0.8830	28.02	0.7680	25.18	0.7540
EDSR [33]	×4	32.46	0.8968	28.80	0.7876	26.64	0.8033
NLRN [34]	×4	31.92	0.8916	28.36	0.7745	25.79	0.7729
RNAN [55]	×4	32.49	0.8982	28.83	0.7878	26.61	0.8023
SRFBN [31]	×4	32.47	0.8983	28.81	0.7868	26.60	0.8015
RDN [56]	×4	32.47	0.8990	28.81	0.7871	26.61	0.8028
RCAN [26]	×4	32.63	0.9002	28.87	0.7889	26.82	0.8087
NLSN [39]	×4	32.59	0.9000	28.87	0.7891	26.96	0.8109
Entire SR	×4	30.68	0.8842	25.60	0.7402	23.31	0.7517
Cropped SR	×4	30.92	0.8832	25.14	0.7374	22.95	0.7404
SOS	×4	28.31	0.8413	24.08	0.7026	22.24	0.6863
Entire SR (Y)	×4	31.26	0.8990	27.33	0.7906	25.76	0.8159
Cropped SR(Y)	×4	31.57	0.8764	25.63	0.7189	23.12	0.7150
SOS(Y)	×4	28.82	0.8339	24.40	0.6929	22.38	0.6752

Table 4.7: Quantitative analysis of different Architectures (Scale 4) with Method-2

4.4.3 Method-3 (DRT)

Super-Resolution Model Results(DRT)

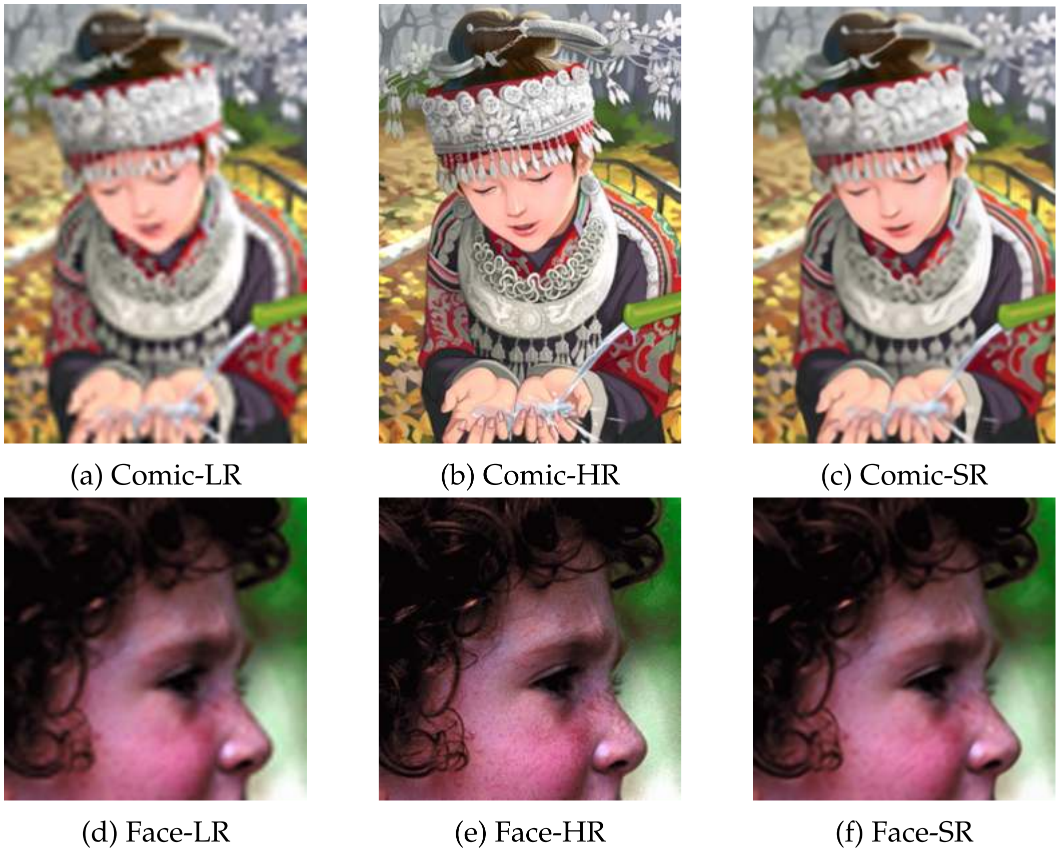
Following figures 4.7,4.8, and 4.9 contain low-resolution input images, high-resolution reference images and super-resolution output images of the respective dataset.



(a) Bird-LR (b) Bird-HR (c) Bird-SR

(d) Butterfly-LR (e) Butterfly-HR (f) Butterfly-SR

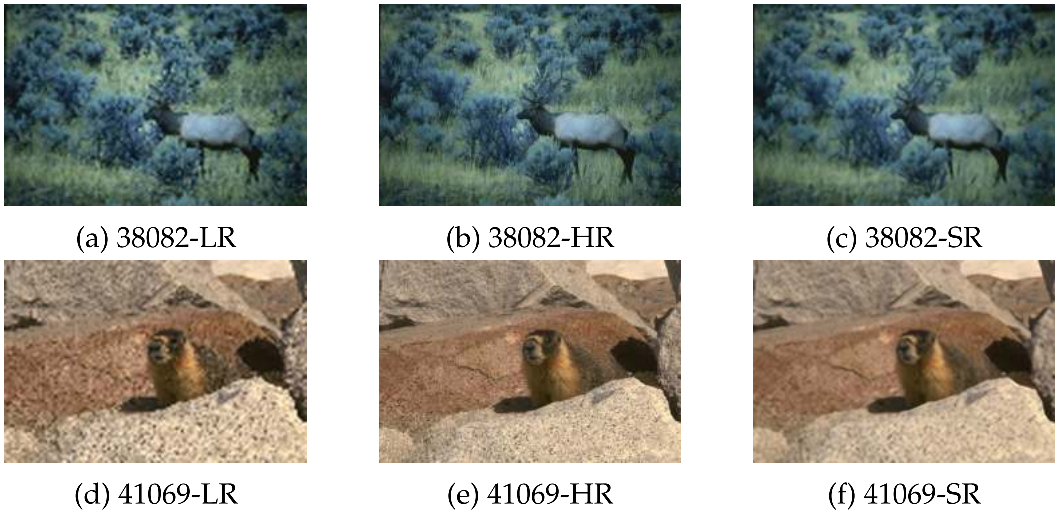
Figure 4.7: SET5 (DRT)



(a) Comic-LR (b) Comic-HR (c) Comic-SR

(d) Face-LR (e) Face-HR (f) Face-SR

Figure 4.8: SET14 (DRT)



(a) 38082-LR (b) 38082-HR (c) 38082-SR

(d) 41069-LR (e) 41069-HR (f) 41069-SR

Figure 4.9: BSDS100 (DRT)

Proposed Method Results

Table 4.8 presents a breakdown of the time taken at each stage in Method-3. The table provides an analysis of the time duration for each step, highlighting the relative time consumption for different stages of the method.

Method	Set5	Set14	BSDS100
Map Generation	2.73	3.03	2.76
Thresholding	0.032	0.002	0.001
Hole Filling	0.040	0.001	0.001
Segmentation	0.005	0.007	0.005
SOS	**64.84**	**150.74**	**81.06**
Restacking	0.11	0.018	0.012
Smoothing	2.461	4.840	3.258
Total Time	70.189	158.638	87.097
Entire SR	**102.86**	**238.39**	**156.34**

Table 4.8: Stagewise Time Analysis of Different Stages for Method-3

Table 4.9 shows a comparative analysis of our Method-3 with state-of-the-art methods using two evaluation metrics: PSNR (Peak Signal-to-Noise Ratio) and SSIM (Structural Similarity Index). The table provides a comparison of the performance of different methods in terms of these metrics, indicating the quality of the output generated by each method.

Method	Scale	Set5		Set14		BSDS100	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Bicubic	×4	28.42	0.8104	26.00	0.7027	25.96	0.6675
SRCNN [46]	×4	30.48	0.8628	27.50	0.7513	26.90	0.7101
VDSR [25]	×4	31.35	0.8830	28.02	0.7680	27.29	0.7026
EDSR [33]	×4	32.46	0.8968	28.80	0.7876	27.71	0.7420
NLRN [34]	×4	31.92	0.8916	28.36	0.7745	27.48	0.7306
RNAN [55]	×4	32.49	0.8982	28.83	0.7878	27.72	0.7409
SRFBN [31]	×4	32.47	0.8983	28.81	0.7868	27.72	0.7409
RDN [56]	×4	32.47	0.8990	28.81	0.7871	27.72	0.7419
RCAN [26]	×4	32.63	0.9002	28.87	0.7889	27.77	0.7436
NLSN [39]	×4	32.59	0.9000	28.87	0.7891	27.78	0.7444
Entire SR	×4	30.06	0.8720	26.68	0.7678	26.35	0.7390
Cropped SR	×4	30.43	0.8735	26.20	0.7695	25.55	0.7295
SOS	×4	27.69	0.8343	24.74	0.7224	23.61	0.6878
Entire SR (Y)	×4	30.58	0.8872	27.13	0.7847	26.36	0.7436
Cropped SR(Y)	×4	31.03	0.8918	26.70	0.7876	25.55	0.7347
SOS(Y)	×4	28.15	0.8507	25.10	0.7394	23.65	0.6934

Table 4.9: Quantitative analysis of different Architectures (Scale 4) with Method-3

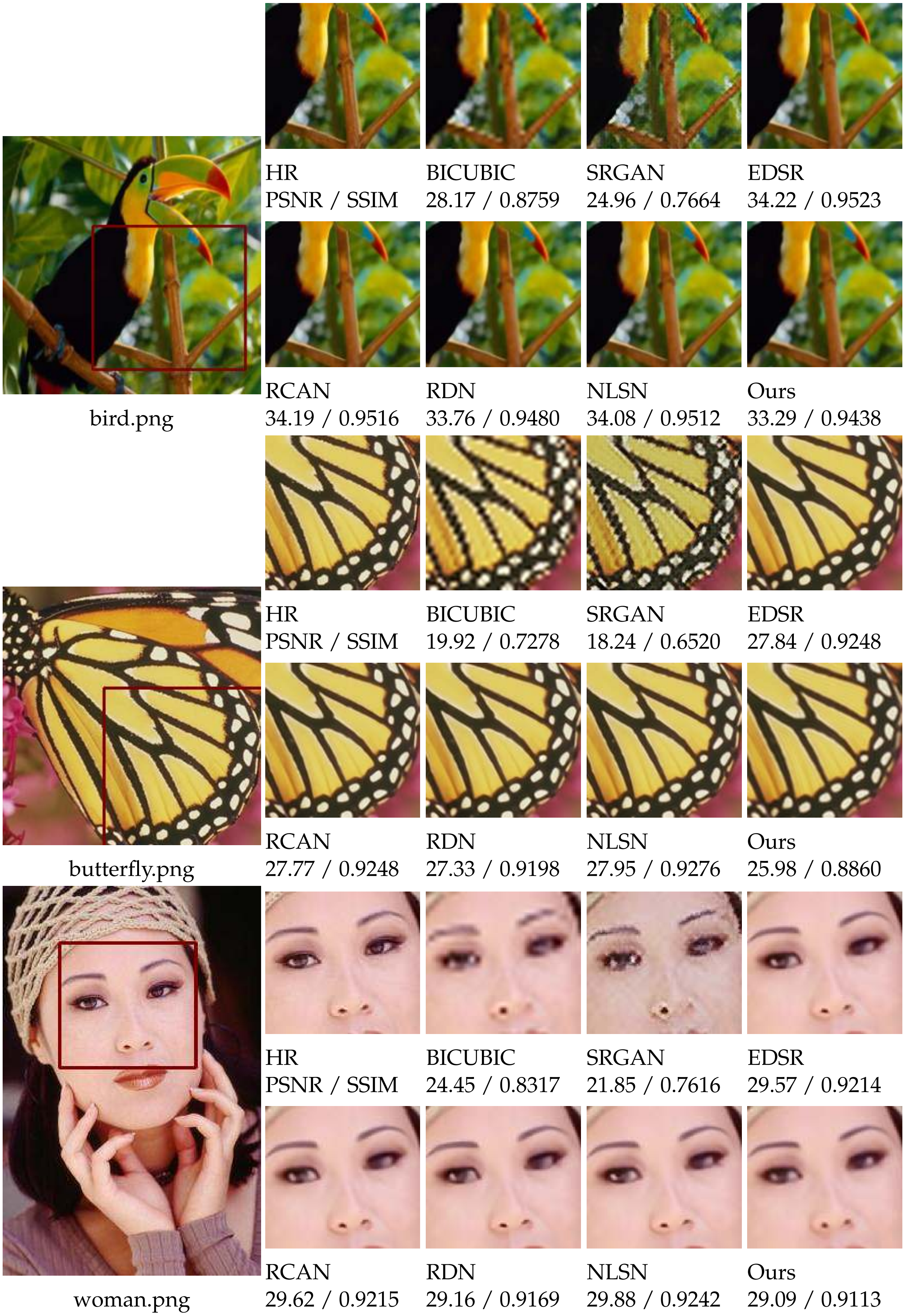


Figure 4.10: Qualitative Results on Scale-4 (SET5) (SOS)

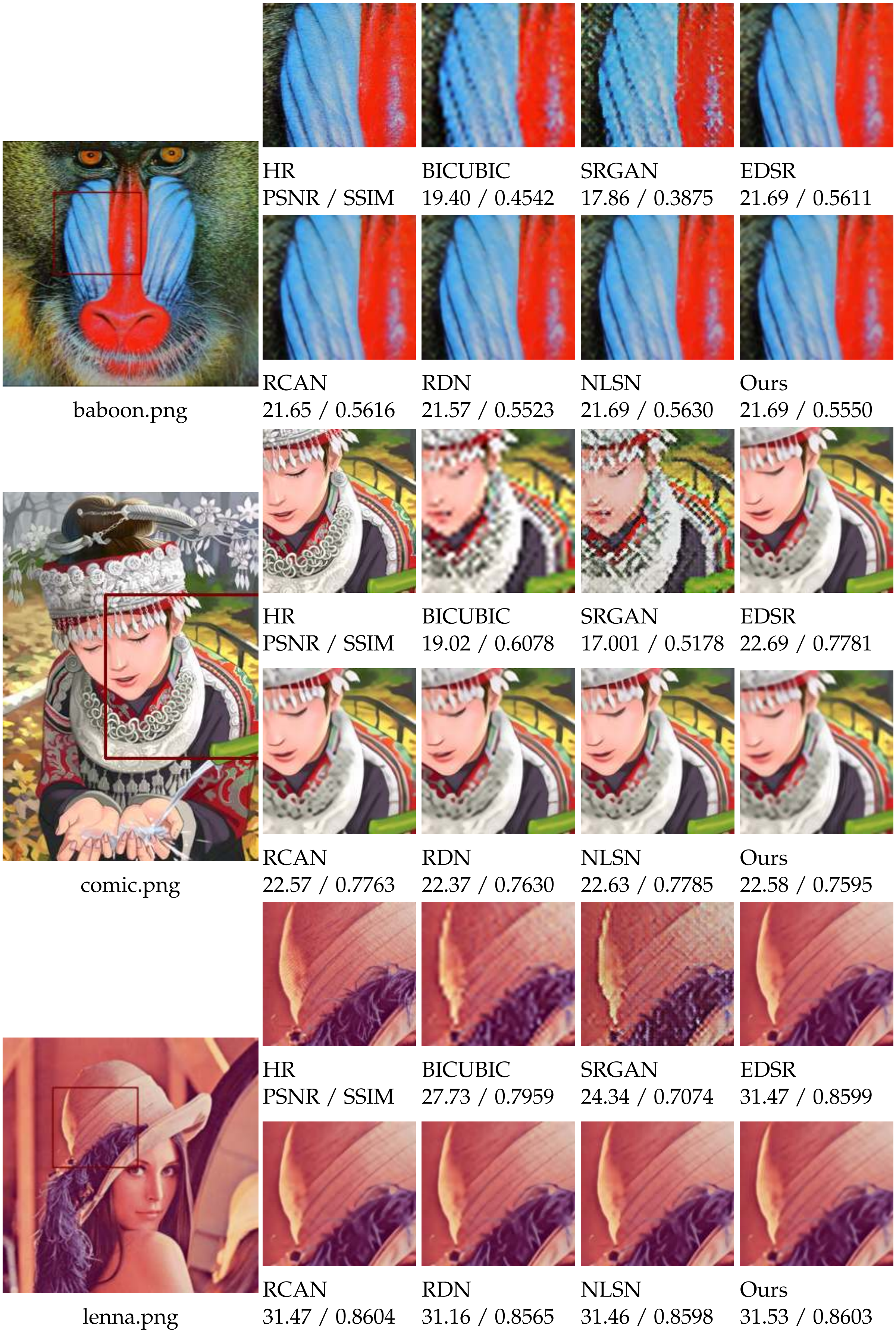


Figure 4.11: Qualitative Results on Scale-4 (SET14) (SOS)

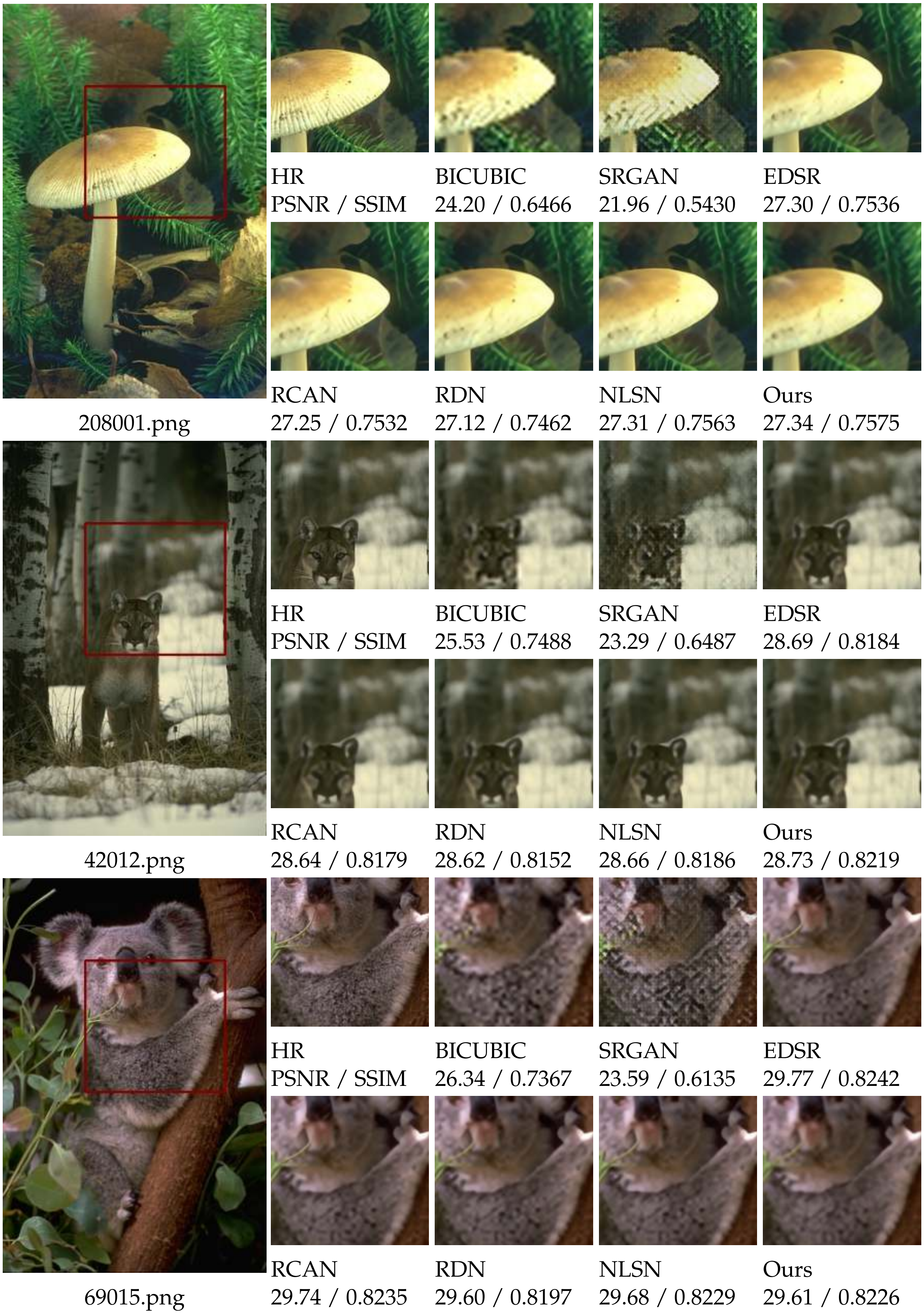


Figure 4.12: Qualitative Results on Scale-4 (BSDS100) (SOS)

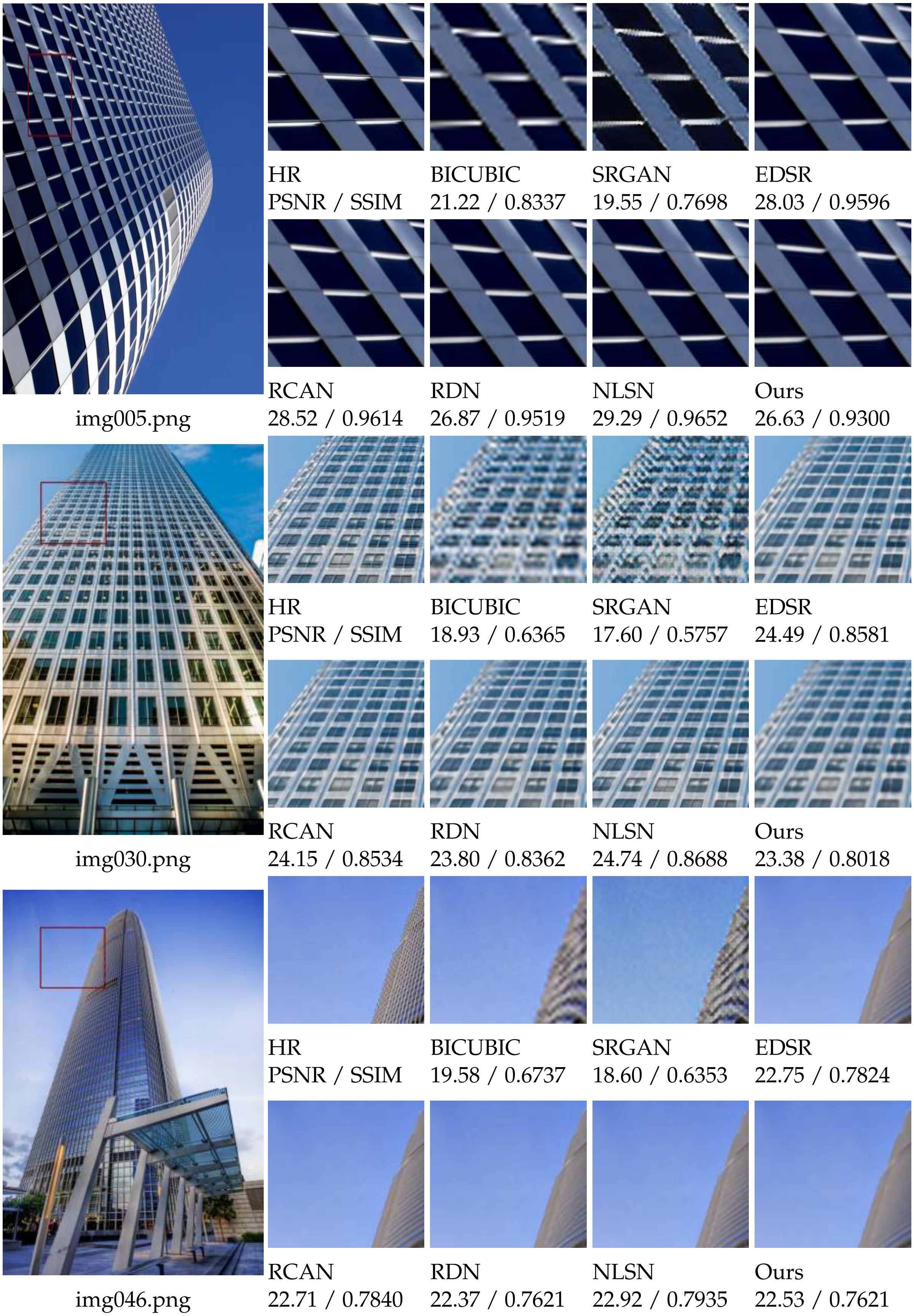


Figure 4.13: Qualitative Results on Scale-4 (Urban100) (SOS)

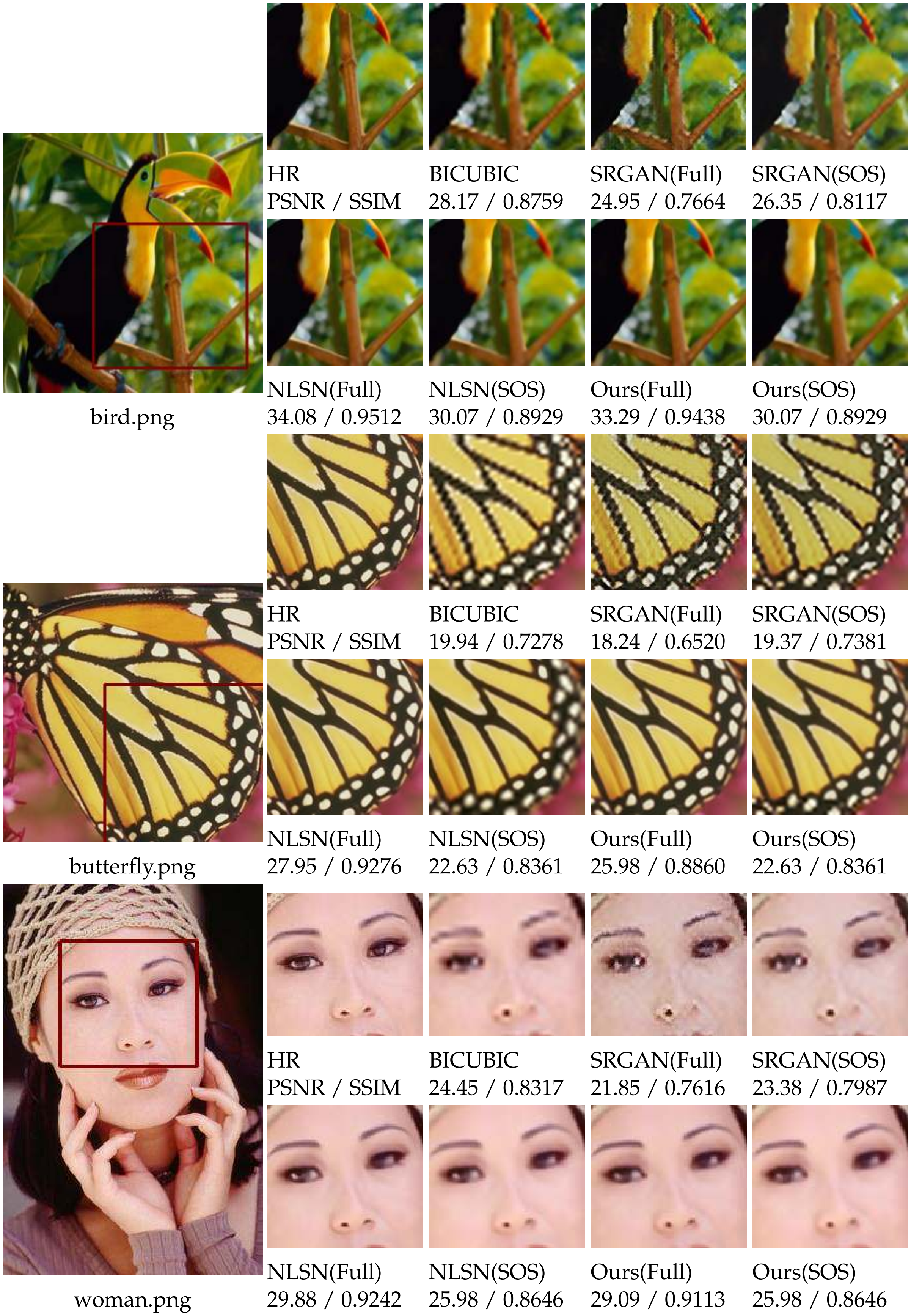


Figure 4.14: Comparative analysis of Various used methods on Scale-4 (Set5)

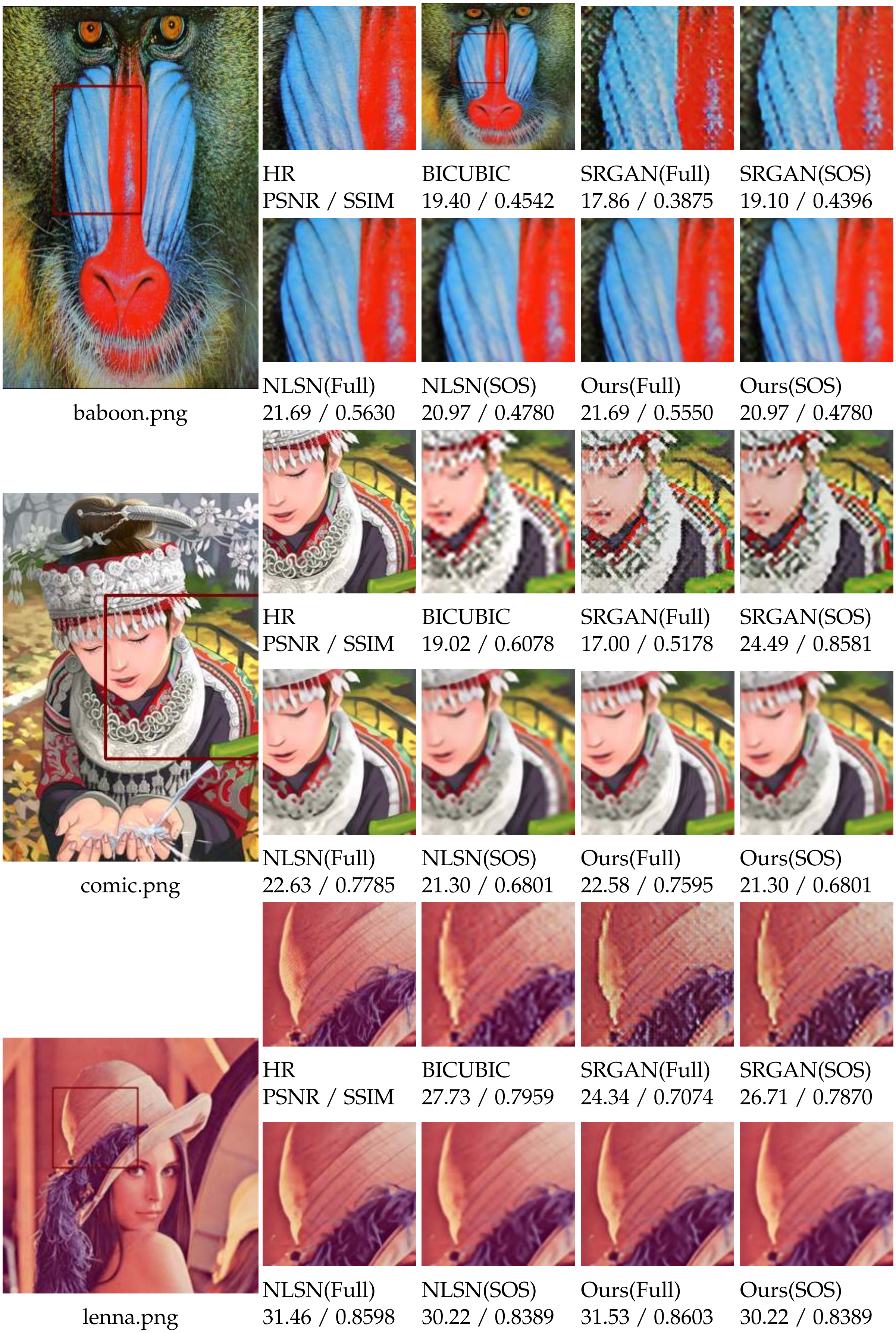


Figure 4.15: Comparative Analysis of various used methods on Scale-4 (Set14)

4.4.4 Proposed model TraNLSN Results

Table 4.10 shows a comparative analysis of our Proposed Model TraNLSN with state-of-the-art methods using two evaluation metrics: PSNR (Peak Signal-to-Noise Ratio) and SSIM (Structural Similarity Index). The table provides a comparison of the performance of different methods in terms of these metrics, indicating the quality of the output generated by each method. Figures 4.16, 4.17, 4.18, and 4.19 shows comparative qualitative analysis of our Proposed Model TraNLSN with state-of-the-art methods.

Method	Scale	Set5		Set14		BSDS100	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Bicubic	×4	28.42	0.8104	26.00	0.7027	25.96	0.6675
SRCNN [46]	×4	30.48	0.8628	27.50	0.7513	26.90	0.7101
VDSR [25]	×4	31.35	0.8830	28.02	0.7680	27.29	0.7026
EDSR [33]	×4	32.46	0.8968	28.80	0.7876	27.71	0.7420
NLRN [34]	×4	31.92	0.8916	28.36	0.7745	27.48	0.7306
RNAN [55]	×4	32.49	0.8982	28.83	0.7878	27.72	0.7409
SRFBN [31]	×4	32.47	0.8983	28.81	0.7868	27.72	0.7409
RDN [56]	×4	32.47	0.8990	28.81	0.7871	27.72	0.7419
RCAN [26]	×4	32.63	0.9002	28.87	0.7889	27.77	0.7436
NLSN [39]	×4	32.59	0.9000	28.87	0.7891	27.78	0.7444
Ours	×4	28.05	0.8305	25.25	0.7347	25.43	0.7117
Ours (Y)	×4	28.50	0.8462	25.60	0.7503	25.43	0.7161

Table 4.10: Quantitative analysis of different Architectures (Scale 4) with Model TraNLSN

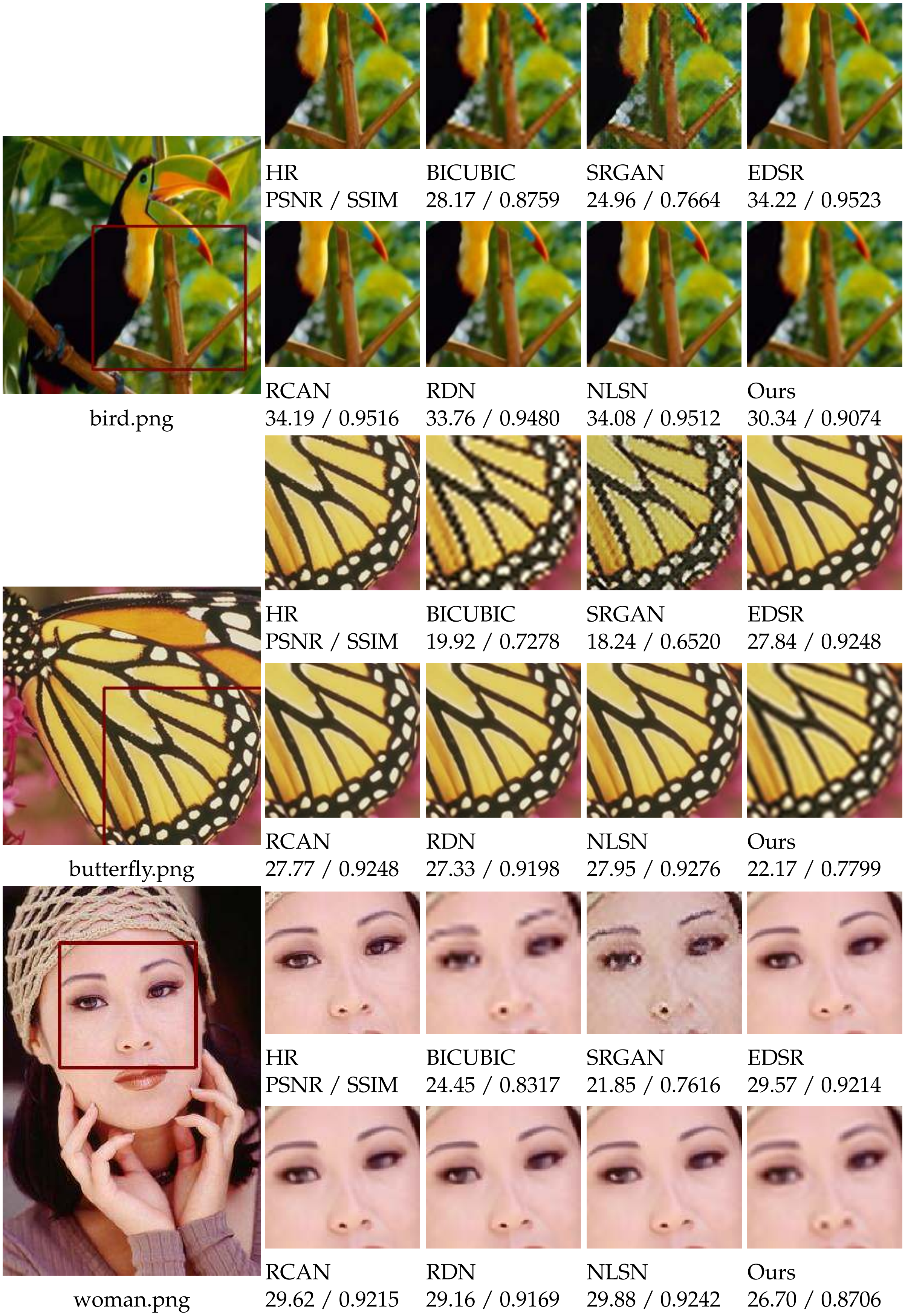


Figure 4.16: Qualitative Results on Scale-4 (SET5) (TraNLSN)

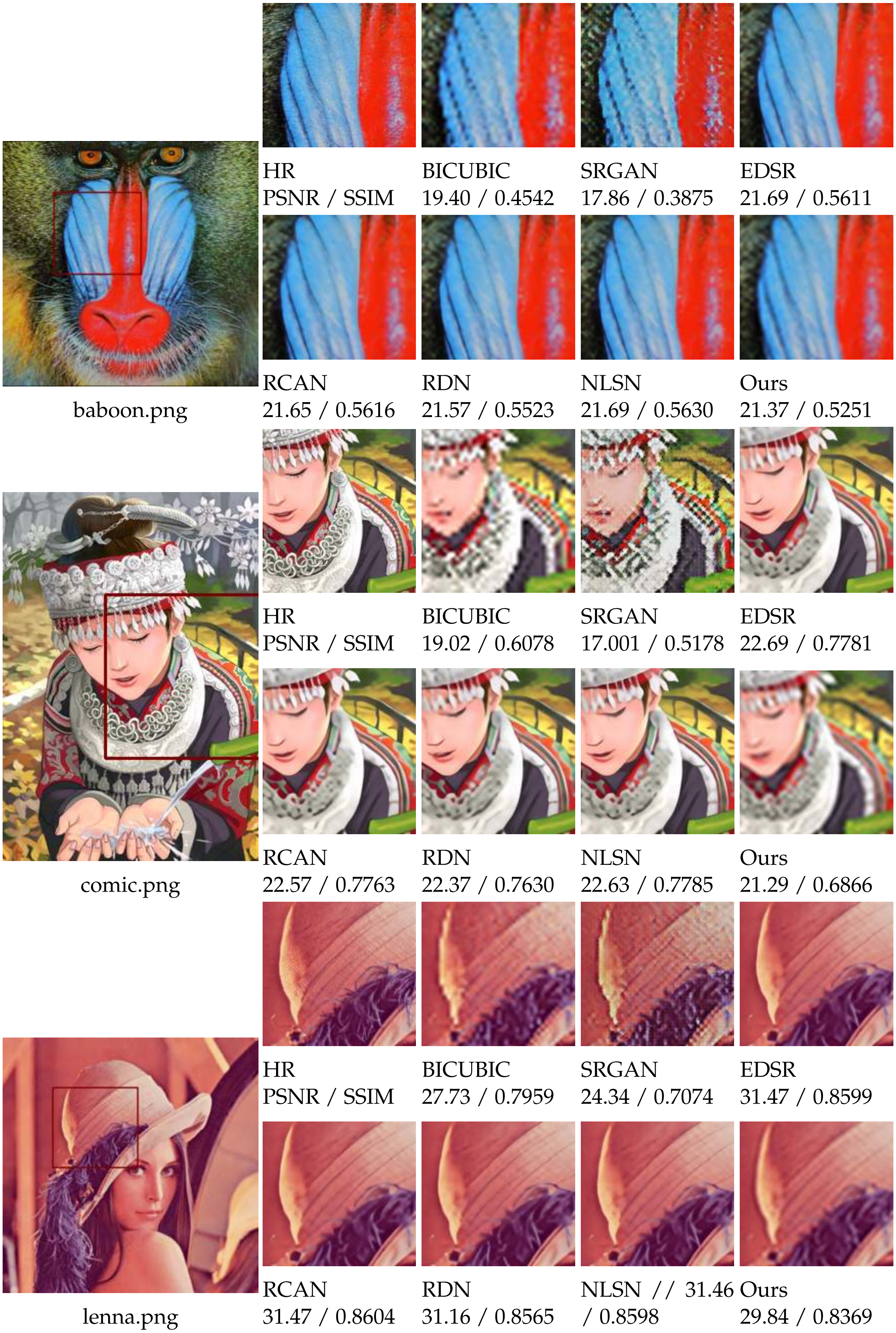


Figure 4.17: Qualitative Results on Scale-4 (SET14) (TraNLSN)

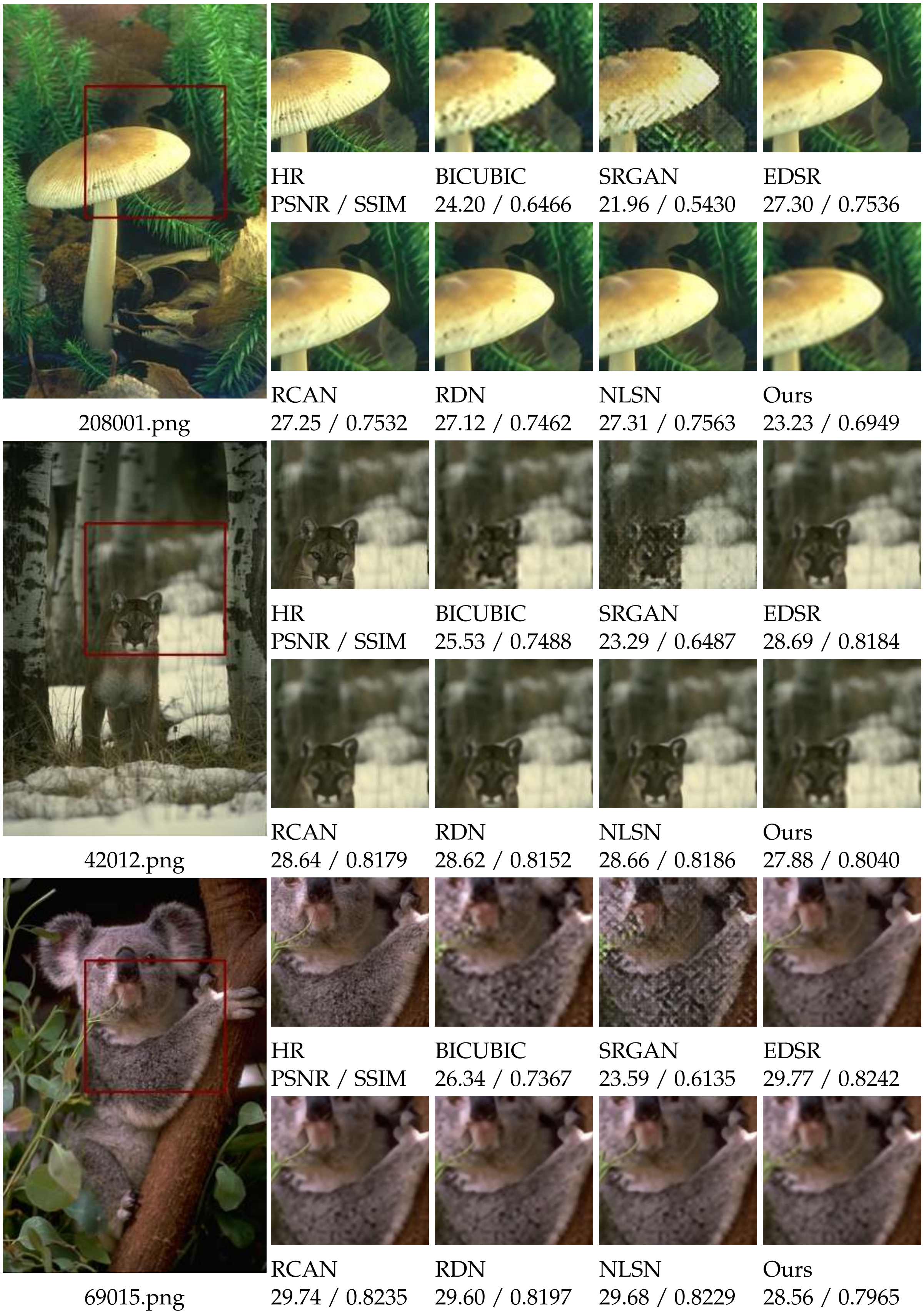


Figure 4.18: Qualitative Results on Scale-4 (BSDS100) (TraNLSN)

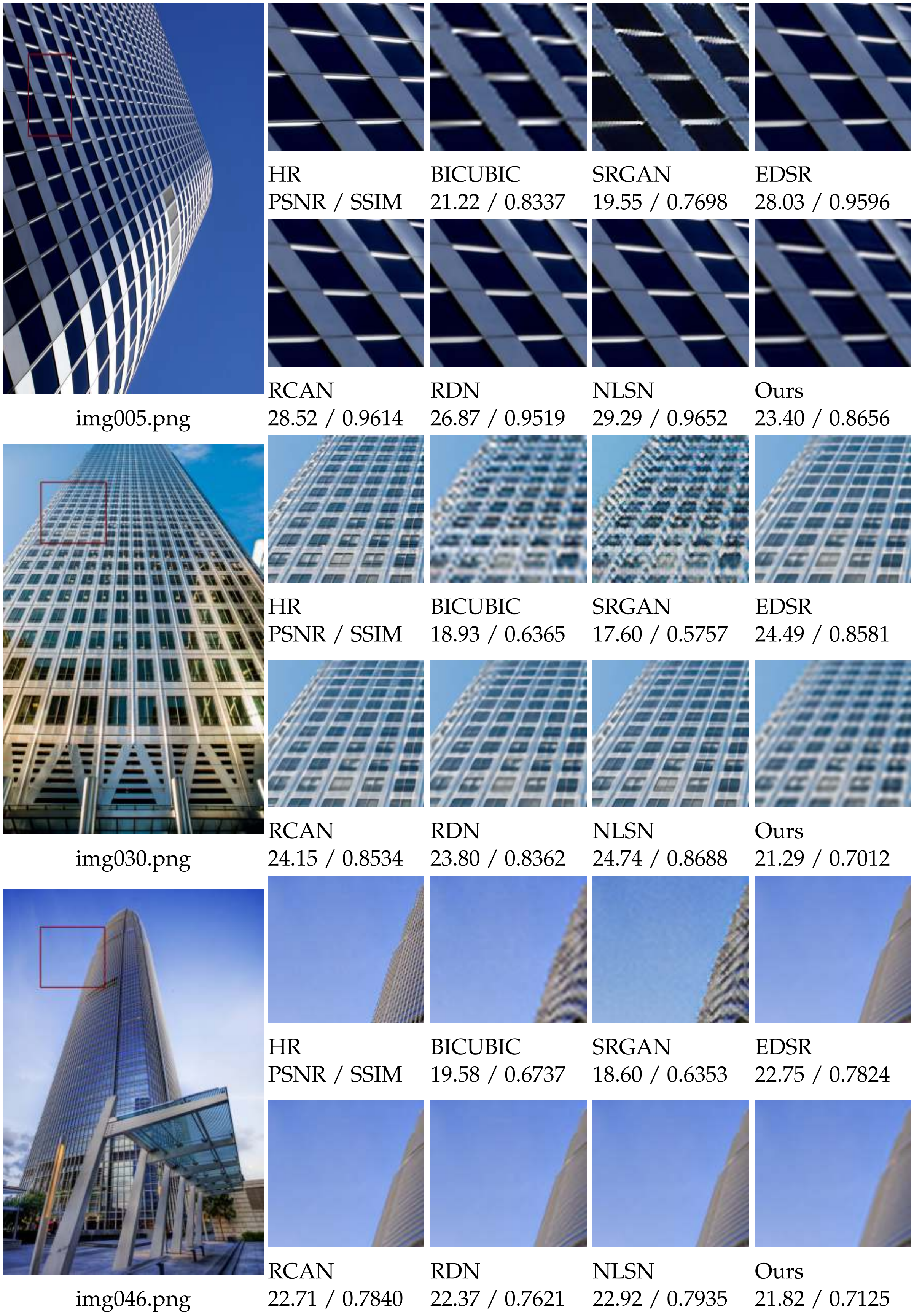


Figure 4.19: Qualitative Results on Scale-4 (Urban100) (TraNLSN)

CHAPTER 5
Conclusion

In this study, we explored three different methods for salient object super-resolution. Each method employed a different approach to enhancing the resolution of salient image objects.

The first method utilized the scratch-trained SRGAN (Super-Resolution Generative Adversarial Network) model for the super-resolution step. SRGAN is a state-of-the-art deep learning model specifically designed for image super-resolution. By training the SRGAN model on a large dataset of low-resolution and high-resolution image pairs, it learns to generate high-resolution versions of low-resolution images. Our method passed the segmented salient objects through the trained SRGAN model to obtain the enhanced super-resolution output.

The second method leveraged a pre-trained NLSN (Non-Local Sparse Attention Networks) model for the super-resolution step. The NLSN is a deep learning model incorporating non-local sparse attention mechanisms to capture long-range dependencies and enhance image restoration tasks. In our approach, we used a pre-trained NLSN model that was originally trained for a different task and repurposed for salient object super-resolution. We fed the segmented salient objects through the pre-trained NLSN model to obtain the super-resolved output.

The third method involved training a DRT (Deraining Recursive Transformer) model from scratch for the super-resolution step. The DRT model is a transformer-based architecture originally designed for deraining tasks. However, we adapted the DRT model for salient object super-resolution by training it on a dataset specifically curated for this task. The DRT model learns to extract relevant features and restore fine details within salient objects during training. We used the trained DRT model to process the segmented salient objects and generate the super-resolution output.

We used these three different methods to compare their performance and assess their effectiveness for salient object super-resolution. Experimental evaluation, including quantitative metrics such as PSNR and SSIM, as well as visual comparisons, will be conducted to determine the strengths and limitations of each method.

The utilization of multiple methods provides a comprehensive analysis and allows us to identify the most suitable approach for salient object super-resolution. This study contributes to the existing body of research by offering insights into different techniques and their potential applications in enhancing the resolution of salient objects in images.

In addition to the three aforementioned methods, we also propose a novel TraNLSN (Transformer with Non-Local Sparse Attention) model for super-resolution. The TraNLSN model combines the strengths of the Deraining Recursive Transformer (DRT) model and the Non-Local Sparse Attention (NLSN) mechanism to enhance the resolution of low-resolution images.

The TraNLSN model architecture consists of several components, including Recursive Transformer Blocks (RTBs), Spatial Transformer Blocks (STBs), and the newly introduced NLSA (Non-Local Sparse Attention) Block. The RTBs, inspired by the DRT model, capture long-range dependencies and hierarchically learn feature representations. Within each RTB, the STBs focus on local context and employ attention mechanisms to extract relevant spatial and channel information.

The key innovation of the TraNLSN model lies in the NLSA Block, which allows the fusion of the NLSN attention modules with the transformer architecture. This block enables selective attention to relevant spatial and channel information, improving feature preservation while suppressing noise and artefacts.

By integrating the DRT model and the NLSN attention mechanism, the TraNLSN model aims to leverage the benefits of both approaches. The DRT model excels in capturing fine details and textures within salient objects, while the NLSN attention mechanism enhances feature preservation and noise suppression. The fusion of these two components in the TraNLSN model provides a comprehensive and effective solution for super-resolution.

Through extensive experiments and evaluations, we will assess the performance of the TraNLSN model and compare it with existing methods. We anticipate that the TraNLSN model will demonstrate superior performance in enhancing low-resolution image resolution and producing visually appealing results.

The introduction of the TraNLSN model contributes to the field of salient object super-resolution by proposing a novel architecture that combines transformer-based modelling and non-local attention mechanisms. This model has the potential to advance the state-of-the-art in super-resolution and pave the way for further research and development in this area.

CHAPTER 6
Future Works

In the future, there are several avenues for further research and development in the field of salient object super-resolution. Some potential areas for future work include:

1. **Specific Salient Object Super-Resolution Model:** Develop a specialized deep learning model designed specifically for salient object super-resolution. This model could leverage the unique characteristics of salient objects, such as their distinct shapes, textures, and structures, to improve the super-resolution process. Tailoring the model to salient object characteristics can effectively capture and enhance the details within salient regions while preserving their perceptual quality.

2. **Refinement of Existing Methods:** Continuously improving and refining the existing methods, such as SRGAN, NLSN, and the scratch-trained DRT model, can lead to enhanced performance in salient object super-resolution. Fine-tuning the models, exploring different training strategies, or incorporating additional components can help achieve better results.

3. **Exploration of Hybrid Approaches:** Investigating hybrid approaches that combine multiple techniques or models can be a promising direction. For example, combining the strengths of SRGAN, NLSN, and the DRT model, or exploring the fusion of different deep learning architectures, could potentially yield more powerful and effective solutions for salient object super-resolution.

4. **Incorporating Attention Mechanisms:** Further exploration and development of attention mechanisms can enhance the capability of models to focus on salient regions and prioritize important features during the super-resolution process. Investigating different attention mechanisms, such as self-attention or attention mechanisms specifically designed for salient object detection, can lead to improved performance.

5. **Real-Time Applications:** Adapting the proposed methods and models for real-time applications can be an interesting area of future work. Optimizing the models' computational efficiency and memory requirements to enable real-time super-resolution processing on resource-constrained devices can open up new possibilities for practical applications.

By focusing on these future research directions, we can further advance the salient object super-resolution field and continue improving the quality and resolution of salient object images, opening up new opportunities for various applications such as image enhancement, computer vision, and multimedia content creation.

References

[1] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk. Frequency-tuned salient region detection. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1597–1604, 2009.

[2] E. Agustsson and R. Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1122–1131, 2017.

[3] B. Alexe, T. Deselaers, and V. Ferrari. What is an object? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 73–80, 2010.

[4] M. Bevilacqua, A. Roumy, C. Guillemot, and M. line Alberi Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In *Proceedings of the British Machine Vision Conference*, pages 135.1–135.10. BMVA Press, 2012.

[5] G. Bradski and A. Kaehler. Opencv. *Dr. Dobb's journal of software tools*, 3:120, 2000.

[6] K.-Y. Chang, T.-L. Liu, H.-T. Chen, and S.-H. Lai. Fusing generic objectness and visual saliency for salient object detection. 2011.

[7] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. S. Torr, and S.-M. Hu. Global contrast based salient region detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(3):569–582, 2015.

[8] S. Goferman, L. Zelnik-Manor, and A. Tal. Context-aware saliency detection. In *CVPR*, volume 2, page 3, 2010.

[9] S. Goferman, L. Zelnik-Manor, and A. Tal. Context-aware saliency detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(10):1915–1926, 2012.

[10] A. K. Gupta, A. Seal, M. Prasad, and P. Khanna. Salient object detection techniques in computer vision-a survey. *Entropy*, 22(10):1174, Oct 2020.

[11] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. In *Advances in Neural Information Processing Systems*, pages 545–552, 2007.

[12] K. He, J. Sun, and X. Tang. Guided image filtering. *IEEE transactions on pattern analysis and machine intelligence*, 35:1397–1409, 06 2013.

[13] J.-B. Huang, A. Singh, and N. Ahuja. Single image super-resolution from transformed self-exemplars. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5197–5206, 2015.

[14] M. J. Huiskes and M. S. Lew. The mir flickr retrieval evaluation. New York, NY, USA, 2008. Association for Computing Machinery.

[15] N. Imamoglu, W. Lin, and Y. Fang. A saliency detection model using low-level features based on wavelet transform. *IEEE Transactions on Multimedia*, 15(1):96–105, 2013.

[16] L. Itti and C. Koch. Comparison of feature combination strategies for saliency-based visual attention systems. In *Human Vision and Electronic Imaging IV*, pages 473–483. International Society for Optics and Photonics, 1999.

[17] Y. Jia and M. Han. Category-independent object-level saliency detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1761–1768, 2013.

[18] M. Jian, Q. Qi, J. Dong, X. Sun, Y. Sun, and K.-M. Lam. Saliency detection using quaternionic distance based weber local descriptor and level priors. *Multimedia Tools and Applications*, 77(11):14343–14360, 2018.

[19] M. Jian, R. Zhao, X. Sun, H. Luo, W. Zhang, H. Zhang, J. Dong, Y. Yin, and K.-M. Lam. Saliency detection based on background seeds by object proposals and extended random walk. *Journal of Visual Communication and Image Representation*, 57:202–211, 2018.

[20] B. Jiang, L. Zhang, H. Lu, C. Yang, and M.-H. Yang. Saliency detection via absorbing markov chain. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1665–1672, 2013.

[21] H. Jiang, J. Wang, Z. Yuan, T. Liu, N. Zheng, and S. Li. Automatic salient object segmentation based on context and shape prior. In *BMVC*, volume 7, page 9, 2011.

[22] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, and S. Li. Salient object detection: A discriminative regional feature integration approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2083–2090, 2013.

[23] P. Jiang, H. Ling, J. Yu, and J. Peng. Salient region detection by ufo: Uniqueness, focusness and objectness. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1976–1983, 2013.

[24] D. B. Johnson. Efficient algorithms for shortest paths in sparse networks. 24(1), 1977.

[25] J. Kim, J. K. Lee, and K. M. Lee. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1646–1654, 2016.

[26] 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*, pages 4681–4690, 2017.

[27] C. Li, Y. Yuan, W. Cai, Y. Xia, and D. Feng. Robust saliency detection via regularized random walks ranking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2710–2717, 2015.

[28] H. Li, H. Lu, Z. Lin, X. Shen, and B. Price. Inner and inter label propagation: Salient object detection in the wild. *IEEE Transactions on Image Processing*, 24(10):3176–3186, 2015.

[29] X. Li, Y. Li, C. Shen, A. Dick, and A. Van Den Hengel. Contextual hypergraph modeling for salient object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3328–3335, 2013.

[30] X. Li, H. Lu, L. Zhang, X. Ruan, and M.-H. Yang. Saliency detection via dense and sparse reconstruction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2976–2983, 2013.

[31] Z. Li, J. Yang, Z. Liu, X. Yang, G. Jeon, and W. Wu. Feedback network for image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3867–3876, 2019.

[32] Y. Liang, S. Anwar, and Y. Liu. Drt: A lightweight single image deraining recursive transformer. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 588–597, 2022.

[33] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee. Enhanced deep residual networks for single image super-resolution. pages 1132–1140, 07 2017.

[34] D. Liu, B. Wen, Y. Fan, C. C. Loy, and T. S. Huang. Non-local recurrent network for image restoration. In *Advances in Neural Information Processing Systems*, pages 1673–1682, 2018.

[35] R. Liu, J. Cao, Z. Lin, and S. Shan. Adaptive partial differential equation learning for visual saliency detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3866–3873, 2014.

[36] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H.-Y. Shum. Learning to detect a salient object. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(2):353–367, 2011.

[37] Y.-F. Ma and H.-J. Zhang. Contrast-based image attention analysis by using fuzzy growing. In *Proceedings of the eleventh ACM international conference on Multimedia*, pages 374–381, 2003.

[38] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, volume 2, pages 416–423 vol.2, 2001.

[39] Y. Mei, Y. Fan, and Y. Zhou. Image super-resolution with non-local sparse attention. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3516–3525, 2021.

[40] F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung. Saliency filters: Contrast based filtering for salient region detection. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 733–740, 2012.

[41] W. Qi, M.-M. Cheng, A. Borji, H. Lu, and L.-F. Bai. Saliencyrank: Two-stage manifold ranking for salient object detection. *Computational Visual Media*, 1(4):309–320, 2015.

[42] X. Shen and Y. Wu. A unified approach to salient object detection via low rank matrix recovery. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 853–860, 2012.

[43] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition.

[44] I. Ullah, M. Jian, S. Hussain, J. Guo, H. Yu, X. Wang, and Y. Yin. A brief survey of visual saliency detection. *Multimedia Tools and Applications*, 79(45):34605–34645, 2020.

[45] T. N. Vikram, M. Tscherepanow, and S. Wrede. A saliency map based on sampling an image into random rectangular regions of interest. *Pattern Recognition*, 45(9):3114–3124, 2012.

[46] C. M. Ward, J. Harguess, B. Crabb, and S. Parameswaran. Image quality assessment for determining efficacy and limitations of super-resolution convolutional neural network (srcnn). In *Applications of Digital Image Processing XL*, volume 10396, pages 19–30, 2017.

[47] Y. Wei, F. Wen, W. Zhu, and J. Sun. Geodesic saliency using background priors. In *European Conference on Computer Vision*, pages 29–42, 2012.

[48] Y. Xie, H. Lu, and M.-H. Yang. Bayesian saliency via low and mid-level cues. *IEEE Transactions on Image Processing*, 22(5):1689–1698, 2013.

[49] Q. Yan, L. Xu, J. Shi, and J. Jia. Hierarchical saliency detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1155–1162, 2013.

[50] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang. Saliency detection via graph-based manifold ranking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3166–3173, 2013.

[51] W. Yang, X. Zhang, Y. Tian, W. Wang, J.-H. Xue, and Q. Liao. Deep learning for single image super-resolution: A brief review. *IEEE Transactions on Multimedia*, 21(12):3106–3121, 2019.

[52] J. Yousefi. Image binarization using otsu thresholding algorithm, 05 2015.

[53] R. Zeyde, M. Elad, and M. Protter. On single image scale-up using sparse-representations. In *Curves and Surfaces*, pages 711–730. Springer Berlin Heidelberg, 2012.

[54] Y. Zhai and M. Shah. Visual attention detection in video sequences using spatiotemporal cues. In *Proceedings of the 14th ACM International Conference on Multimedia*, pages 815–824, 2006.

[55] 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.

[56] 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*, pages 2472–2481, 2018.

[57] Z. Zhou, Y. Wang, Q. J. Wu, C.-N. Yang, and X. Sun. Effective and efficient global context verification for image copy detection. *IEEE Transactions on Information Forensics and Security*, 12(1):48–63, 2017.

[58] W. Zhu, S. Liang, Y. Wei, and J. Sun. Saliency optimization from robust background detection. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2814–2821, 2014.

[59] W. Zhu, S. Liang, Y. Wei, and J. Sun. Saliency optimization from robust background detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2814–2821, 2014.