Design and Development of an Inception-Based Multiscale Algorithm for Single Image Super-Resolution

Nashra Babar

A Thesis

in

The Department

of

Electrical and Computer Engineering

Presented in Partial Fulfillment of the Requirements

for the Degree of

Master of Applied Science (Electrical and Computer Engineering) at

Concordia University

Montréal, Québec, Canada

August 2025

© Nashra Babar, 2025

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: **Nashra Babar**

Entitled: **Design and Development of an Inception-Based Multiscale Algorithm for Single Image Super-Resolution**

and submitted in partial fulfilment of the requirements for the degree of

Master of Applied Science (Electrical and Computer Engineering)

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the Final Examining Committee:

_____ Chair
Dr. Wei-Ping Zhu

_____ Examiner
Dr. Wei-Ping Zhu

_____ Examiner
Dr. Chunyan Wang

_____ Supervisor
Dr. M. Omair Ahmad

Approved by _____
Dr. Abdelwahab Hamou-Lhadj, Chair
Department of Electrical and Computer Engineering

August 14, 2025 _____
Dr. Mourad Debabbi, Dean
Gina Cody School of Engineering and Computer Science

Abstract

Design and Development of an Inception-Based Multiscale Algorithm for Single Image Super-Resolution

Nashra Babar

The field of single-image super-resolution (SISR) has made considerable progress with the emergence of deep convolutional neural networks, where residual learning techniques have played a crucial role in enhancing reconstruction quality. Among these methods, SwinIR, a transformer based network, demonstrates remarkable performance by leveraging hierarchical self-attention mechanisms to effectively capture both fine-grained local structures and broader global contextual dependencies. However, enhancing image quality while maintaining computational efficiency remains a key challenge.

To address the limitations in capturing diverse spatial features without increasing architectural overhead, we propose EMS network, an enhanced multiscale SISR framework that draws inspiration from the inception module to refine feature extraction across multiple scales. Our network design adopts the underlying principle of parallel multi-scale feature extraction from the inception module, where several convolutional layers with different receptive fields operate to capture spatial features at multiple scales. Our method maintains the lightweight design of the network while broadening its receptive field, allowing it to surpass existing state-of-the-art methods in both efficiency and reconstruction quality. Quantitative and qualitative evaluations demonstrate that enhanced multiscale network consistently outperforms state-of-the-art SISR models on standard benchmark datasets, delivering improved visual quality and superior quantitative performance with minimal impact on computational complexity.

Acknowledgments

I am deeply grateful to Allah, God Almighty, for bestowing upon me the resolve, endurance, and determination to bring this thesis to completion.

I extend my sincere gratitude to my supervisor, Dr. M. Omair Ahmad, for granting me the opportunity to pursue my thesis under his guidance. His unwavering support, thoughtful insights, and continuous encouragement have been instrumental throughout my MASc. journey. I am truly thankful for his mentorship and the privilege of conducting my research under his supervision.

I also extend my appreciation to the Natural Sciences and Engineering Research Council (NSERC) for funding my research.

My heartfelt thanks go to my loving and supportive parents, especially my younger sister and brothers, and my fiancé for their unwavering prayers, belief in me, and constant emotional support. Their presence has been my greatest strength through every challenge.

Finally, I am thankful to all the wonderful people I had the privilege of meeting during my MASc. journey — especially my friends Khushboo and Safwan. Their kindness, support, and camaraderie made this experience truly memorable.

Contents

List of Figures

List of Tables

List of Abbreviations

Adam	Adaptive moment estimation
CNN	Convolutional neural network
DL	Deep learning
EDSR	Enhanced deep super-resolution
EMS	Enhanced multi-scale network
FC	Fully connected
JPEG	Joint photographic experts group
MLP	Multi layer perceptron
PSNR	Peak signal-to-noise ratio
ReLU	Rectified linear unit
RSTB	Residual swin transformer block
SISR	Single image super-resolution
SRCNN	Super-resolution convolutional neural network
SSIM	Structural similarity index measure
STL	Shallow transformer layer
SwinIR	Swin transformer for image restoration
VDSR	Very deep super-resolution
W-MSA	Window-based multi-head self attention

List of Symbols

I_x	Low-resolution image
I_y	High-resolution image
I_{SR}	Output super-resolved image
κ	Blur kernel
σ	Standard deviation
λ	Regularization coeffecient
θ_D	Degradation parameters
$\phi(\theta)$	Regularization term
$\mathcal{L}$	Loss function
F_{SF}	Shallow features
F_{DF}	Deep features
$f_{SFE}(\cdot)$	Shallow feature extraction function
μ_x	Local mean of image x
μ_y	Local mean of image y
σ_x^2	Variance of image x
σ_y^2	Variance of image y
σ_{xy}^2	Covariance between x and y images

Chapter 1

Introduction

1.1 General

Single image super-resolution (SISR) is a fundamental problem in computer vision that aims to reconstruct a high-resolution (HR) image from a corresponding low-resolution (LR) input image. Single-image super resolution has drawn a lot of attention for decades, and several state-of-the-art methods give outstanding results. SISR is responsible for predicting and generating multiple high-resolution pixels based on a limited number of pixels in a low-quality remote sensing image. This is a core task in the computer vision field due to its multitude of applications, ranging from helping current-day issues of storage and transfer of data to restoration of low-resolution images. The SISR goal is to reconstruct a high-resolution image from a single low-resolution image.

SISR has emerged as a critical application area. Super-resolution (SR) aims to produce the edge and texture information of images, improve the clarity of images and visual perception, so it is widely used in real life, for example, medical diagnosis, video surveillance, remote sensing images, satellite imagery, and face recognition. In addition to all this, SR also helps to improve other computer vision tasks such as object detection and image synthesis.

SISR is about creating a high-resolution image out of a low-resolution one, but it has many possible solutions, making it an inherently ill-posed problem. To solve this problem, strong prior information is used, often through example-based methods that learn from similar images or pairs of LR and HR images. These methods can extract image patterns or use pairs of images to learn how to improve the resolution of the images. The degradation process, which typically involves down-sampling, blurring, and noise corruption, discards a significant amount of spatial frequency information, resulting in a loss of fine textures and high-frequency details. Due to this information loss, the mapping from LR to HR is inherently non-unique, leading to an underdetermined system where multiple HR images can correspond to the same LR input images [1].

Traditionally, SISR has relied on interpolation-based methods or sparse coding approaches. While these conventional techniques offer simplicity and low computational cost, they often struggle to recover fine details and sharp textures of images, especially under large upscaling factors. Like many image transformation tasks, SISR has three core steps: feature extraction, non-linear mapping and reconstruction [1]. Traditional methods often treat these stages separately, making the design process time-consuming and computationally inefficient. Deep learning, however, offers an end-to-end solution that unifies these steps within a single framework, significantly reducing manual and computational costs.

The emergence of deep learning has significantly reshaped the landscape of SISR research. The rapid growth of large-scale datasets and the increased computational capabilities of modern GPUs have facilitated the development of data-driven models capable of learning complex mappings between LR and HR image pairs.

The advent of deep learning, particularly convolutional neural networks (CNNs), has revolutionised the field of image super-resolution by enabling end-to-end optimisation and

hierarchical feature learning. A convolutional neural network typically has multiple convolution layers. Conceptually, each convolutional layer extracts spatial features from its inputs. In the initial layers, convolutional networks extract low-level features such as colour, texture, edges, and simple shapes. As the network deepens, the feature maps evolve to represent increasingly high-level patterns, capturing broader contextual and structural information. Deep neural networks are capable of learning complex non-linear mappings from LR to HR image domains by leveraging vast amounts of annotated data [1] and multi-scale feature abstraction [2]. Unlike the traditional methods like interpolation-based methods [3] or sparse coding [4], [5], deep networks unify the SR pipeline into a single learnable system, often yielding superior performance in terms of both quantitative metrics (e.g., PSNR, SSIM) and perceptual quality.

Deep learning has significantly advanced the field of single-image super-resolution (SISR), with CNN-based methods like SRCNN [1], VDSR [6], and EDSR [7] showing early promise through their ability to learn spatial features and residual mappings. Generative models such as SRGAN [8] and ESRGAN [9] further improved visual quality by incorporating perceptual and adversarial loss functions. More recent transformer-based [10], [11], [12], [13], and attention-guided approaches [13], [14], including SwinIR [13], have enhanced the modeling of long-range dependencies and global context. These developments highlight the evolving focus on architectural efficiency, feature diversity, and adaptability to varying degradation types, motivating the need for more balanced and scalable solutions. Deep convolutional neural networks (CNNs) have demonstrated remarkable success by automatically learning hierarchical features, leading to substantial improvements in reconstruction fidelity.

1.2 Overview of Deep Learning for Image Super Resolution

Convolutional Neural Networks (CNNs) are among the first deep learning architectures applied to SISR. The pioneering work of SRCNN [1] has introduced an end-to-end learning framework that laid the foundation for direct LR-to-HR mapping. However, its shallow structure limited its ability to capture complex patterns and fine textures. Subsequent deeper CNN-based methods [7], [11], [15] have addressed these limitations, progressively improving accuracy through architectural and training strategies.

EDSR [7] has enhanced the SRCNN [1] architecture by removing batch normalisation layers and deepening the residual blocks, leading to significant gains in PSNR. However, the increase in network depth has also resulted in greater computational demands and memory consumption. RCAN [15] has tackled this issue by incorporating channel attention modules that dynamically recalibrate features, allowing the network to focus on informative components. However, its reliance on local receptive fields restricted its ability to capture long-range dependencies.

To overcome these challenges, architectures such as MemNet [16] has used recursive units and memory mechanisms, which has improved learning capacity without excessively increasing parameters. Models like MAN [17] and CVHSSR [18] have introduced multi-scale and hierarchical attention strategies to balance performance and efficiency, although they still struggled with modeling global contextual information.

Several deep learning SISR architectures have drawn inspiration from the Inception architecture originally introduced as GoogLeNet (Inception v1) by Szegedy et al. [19] for image classification. Designed to enhance classification accuracy via multi-branch, multi-scale convolutional operations while maintaining computational efficiency, the inception

module enabled deep hierarchical feature extraction. Though its primary use was for classification, it has since influenced various goals, including object detection, segmentation, and SISR.

Recent SISR methods [2], [16], [20] have adapted inception-inspired strategies, particularly parallel convolutions of varying kernel sizes to enhance multi-scale representation learning. This adaptation supports efficient capturing of both local textures and broader structures, addressing key challenges in SISR. While some architectures adopt inception-like modules directly, others integrate these ideas into residual or attention-based designs. However, balancing computational cost and effective feature aggregation remains a challenge.

Recently, transformer-based architectures have emerged to address these challenges. SwinIR [13] has introduced a novel approach by combining the hierarchical design of CNNs with the self-attention mechanism of transformers. It employs window-based attention and shifted window strategies to model long-range dependencies efficiently while maintaining manageable computational overhead. SwinIR overcomes the local limitation of CNNs and achieves superior performance across standard SISR benchmarks, making it a compelling baseline for further architectural exploration.

1.3 Motivation and Objectives

While SwinIR [13] has established itself as a strong performer in single-image super-resolution, it still faces challenges in modeling fine textures and global context due to its uniform convolution operations and fixed receptive fields. As a result, performance tends to degrade on complex or severely degraded images. To address this, models can incorporate mechanisms that can extract features across multiple spatial scales and adaptively learn both local and global representations. One potential avenue to address this shortcoming

is the use of inception-inspired modules [19], which perform parallel convolutional operations of different kernel sizes, thereby enabling effective multi-scale feature extraction without significantly increasing computational cost.

The objective of this thesis is to propose a lightweight SISR network, referred to as an enhanced multi-scale (EMS) network, that is capable of learning both local and global features effectively to overcome the problem of network performance across different scales. The proposed EMS network consists of a combination of two main components. The first one is inspired by the inception network [19], enabling efficient multi-scale feature extraction within a single layer by combining convolutions of different kernel sizes. The second one is derived from the SwinIR [13], a transformer-based hierarchical network, to model long-range dependencies with minimal computational burden. While both approaches offer unique advantages in feature representation, their combined integration for SISR remains largely unexplored. The approach aims to address the limitations of SwinIR by integrating the two previously mentioned modules into a computationally efficient framework that improves reconstruction quality across standard SISR benchmarks.

1.4 Organisation of the Thesis

The thesis is organised as follows. Chapter 2 provides essential background material, such as convolutional neural networks, transformer architectures, and the SwinIR framework, relevant to the development of the scheme presented in this thesis. This material forms the foundation for our proposed method. In Chapter 3, we present our proposed Enhanced Multi-Scale network for SISR. In this chapter, our design is introduced and evaluated through a series of ablation studies to demonstrate the effectiveness of its design. Various strategies employed during the network design and training processes are analysed to ensure a robust and efficient image reconstruction across multiple scales. The performance of the proposed scheme is compared against leading state-of-the-art methods using

standard metrics such as PSNR and SSIM. In Chapter 4, we conclude the thesis by summarising the proposed scheme and highlighting its key contributions. The chapter also includes a brief discussion on the scope for further investigation based on the findings in this thesis.

Chapter 2

Background Material

2.1 Introduction

In this chapter, we introduce key background material supporting the methodology to be presented in this thesis. This chapter introduces the theoretical underpinnings and design discussion behind the network components that helps to understand our proposed work. We begin the chapter with an overview of convolutional neural networks and their relevance to image processing tasks. Then, we outline the basic operations involved in single image super-resolution. The core stages such as, patch extraction, non-linear mappings, and reconstruction, of a CNN-based super-resolution network, are discussed to establish a clear understanding of the single image super-resolution network design. A brief introduction to transformer, an attention-based network is provided, followed by a discussion of SwinIR [13], a transformer-based image restoration model. Lastly, the inception module is introduced for its role in enabling multi-scale feature learning within deep networks.

2.2 Convolutional Neural Networks

The initial contributions to convolutional neural networks were fundamental to the evolution of contemporary deep learning. LeCun et al. introduced CNNs in 1990 [21] with their groundbreaking research on recognising handwritten digits through backpropagation, the network is referred to as LeNet. LeNet demonstrates the power of learning hierarchical features. As time has progressed, the architecture of convolutional neural networks (CNNs) has evolved significantly, leading to the creation of more profound and deeper models such as VGGNet [22] and ResNet [2], which have established new standards in image recognition tasks. The capability of CNNs to effectively capture spatial and contextual features has contributed to their extensive application across various computer vision tasks. In a CNN, each layer receives input represented as a three-dimensional tensor with dimensions H X W X D, where D denotes the depth, representing the number of channels (which is 3 for an RGB image), while H and W indicate the height and width of the tensor respectively. Convolutional layers in a CNN utilizes several kernels (or filters), which are also represented in three dimensions but are smaller in size than the input image.

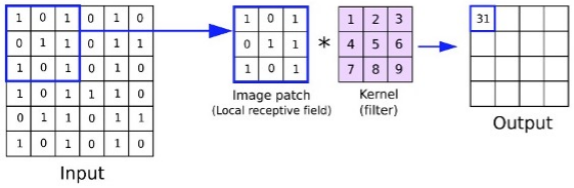

Figure 2.1: Convolution dot product

These kernels slide over the input to perform an element-wise multiplication followed by summation, known as the dot product, which creates feature maps, as demonstrated in Figure 2.1.

As shown in Figure 2.2, a typical CNN consists of multiple layers: an input layer, several intermediate layers each consisting of convolutional layers, an activation layer, a pooling layer and an output layer consisting of a fully connected layer. The input layer

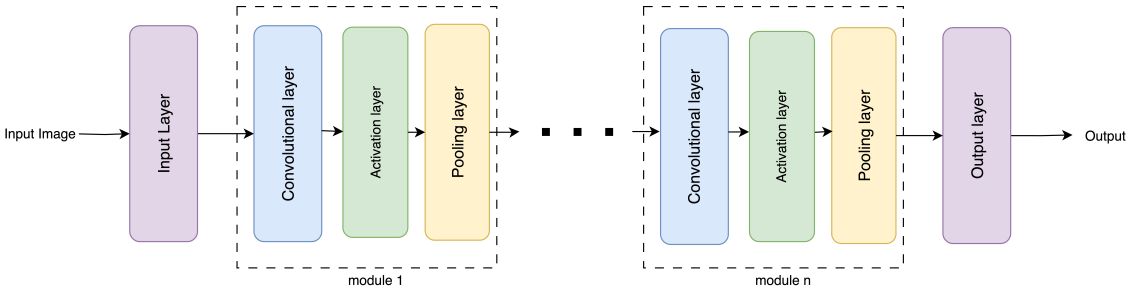

Figure 2.2: A typical architecture of a convolutional neural network.

serves as the first layer for an input image, representing pixel values in a multi-dimensional array format. The output of this layer is then processed by convolutional layers, which apply learnable filters or kernels that convolve over the input to extract local features such as edges and textures. These kernels are initialised with random weights and are updated through backpropagation during training. By sharing weights across spatial positions, convolutional layers reduce the number of parameters and computational complexity while effectively learning spatial hierarchies within the image. Following convolution, a non-linear activation function commonly used rectified linear unit (ReLU)—is applied to introduce non-linearity by preserving positive values and suppressing negative ones, enabling the network to model complex patterns beyond linear transformations. The pooled feature maps then undergo downsampling in the pooling layers, which reduce their spatial dimensions through operations like max pooling or average pooling. This step not only reduces the learnable parameters but also helps in overcoming overfitting. The final layer of a typical convolutional neural network constitutes the fully connected layer(FC), where the input is reduced feature maps. Fully connected layers are predominantly utilised in tasks where the primary goal is to map the learned features to specific categories or output classes, such as in image classification, image clustering, and speech recognition. However, for dense prediction tasks such as image super resolution, denoising, and artefact removal, the final layers are typically a 1 X 1 or 3 X 3 convolutional layer. In classification networks, the output from the fully connected (FC) layer is processed using a SoftMax function to generate

the probability distribution across the target classes.

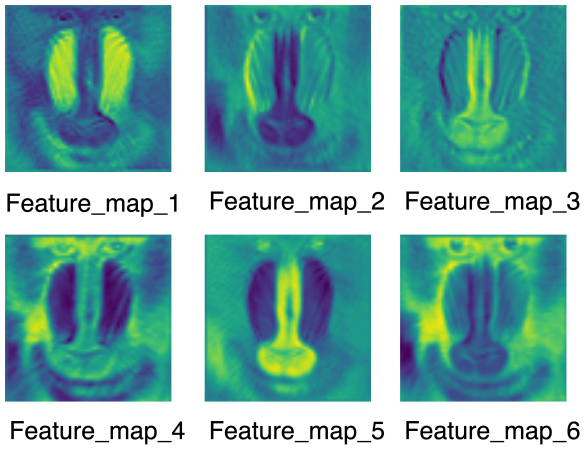

Figure 2.3: Visualisation of feature maps extracted from different convolutional blocks within a CNN network.

The structure and components of convolutional neural networks are highly adaptable and evolve based on the nature of the task. While the core architecture includes the components discussed above, additional elements such as batch normalisation, dropout, and regularisation techniques are included to enhance performance. For example, a network may have six distinct convolutional layers, each responsible for extracting progressively more complex features from the input image as shown in the Figure 2.3. The initial convolutional blocks focus on capturing basic local patterns such as edges, corners, and textures. As the weights propagates through successive blocks, the network develops more abstract and high-level feature representations, such as shapes, object parts, or semantic information relevant to the specific task. This hierarchical feature extraction process, distributed across multiple convolutional layers, allows the CNN to effectively learn rich and discriminative representations that improve performance in tasks like image classification, segmentation, or super-resolution. The choice of the final layers, whether fully connected or convolutional, varies depending on the objective of the problem at hand, which could

be, for example, classification [16], or super resolution [3], [7], [10], [12], [23], or segmentation [24]. Henceforth, a CNN architecture is highly efficient and flexible, with its configurations tailored to meet the specific tasks.

2.3 Basic Operations in a Single Image Super Resolution using Deep Convolutional Neural Network

Single image super-resolution (SISR) focuses on reconstructing a high-resolution (HR) image from a single low-resolution (LR) input. Early SISR methods relied on interpolation such as, bicubic [25] or Lanczos resampling [26] while later methods used statistical priors methods such as, sparse coding [5], [20] and neighbour embedding [25], [26], but all of the above methods had some limitations in feature extraction, scalability, and generalization.

The advent of deep learning (DL) has led to a paradigm shift to SISR, offering the ability to model complex mappings directly from data through end-to-end learning pipelines. Methods based on deep learning combine feature extraction, non-linear transformation, and reconstruction into cohesive frameworks, which greatly lessen the need for architecture design and enhance the accuracy of reconstruction. Convolutional neural networks and, more recently, transformer-based models [11], [12], [13], [17] have demonstrated state-of-the-art results across standard datasets. Furthermore, modern hardware accelerators have made it feasible to train deeper and more expressive models efficiently. This section presents an overview of the basic operations involved in SISR.

The common degradation process illustrated in Figure 2.4 involves blurring, downsampling, and the addition of noise to simulate the low-resolution (LR) image. This degradation from HR to LR can be modelled as

$$I_x = D(I_y; \theta_D) = (I_y \otimes \kappa) \downarrow_s + n, \tag{1}$$

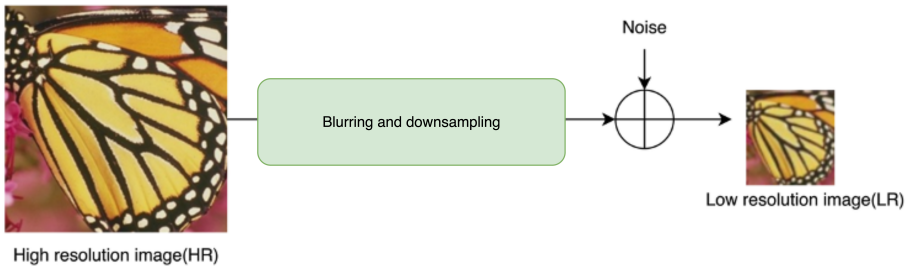

Figure 2.4: Example of degradation process for HR to LR in SISR framework.

where D is the degradation function, θ_D is the degradation parameter, I_y and I_x denote the HR and LR images respectively, $I_y \otimes \kappa$ is the convolution between the blurry kernel κ and I_y, HR image, $\downarrow_s$ denotes downsampling by scale factor s, and n is additive white Gaussian noise (AWGN) with standard deviation σ. The degradation parameters θ_D, including κ and σ, are typically unknown in practical scenarios.

Consequently, SISR aims to learn a mapping $F(I_x; \theta_F)$ that reconstructs a super-resolved image in as

$$\hat{\theta} = \arg\min \mathcal{L}(I_{SR}, I_y) + \lambda\phi(\theta), \tag{2}$$

where I_{SR} is the predicted super-resolved image predicted by model F with parameters θ_F, $\mathcal{L}$ denotes a loss function (typically mean absolute error loss or mean squared error loss) between the resultant SR image I_{SR} and the HR image I_y, $\phi(\theta)$ is a regularization term, and λ controls the weight of the regularization term.

Modern deep learning-based methods, particularly convolutional neural networks, have made this end-to-end mapping highly effective by leveraging large-scale datasets and hierarchical feature extraction capabilities. Architectural advancements such as, residual blocks, multiscale designs, and attention mechanisms further enhance reconstruction fidelity and network efficiency. In convolutional neural network-based SISR, the central objective is to reconstruct a high-resolution image from a low-resolution input by learning

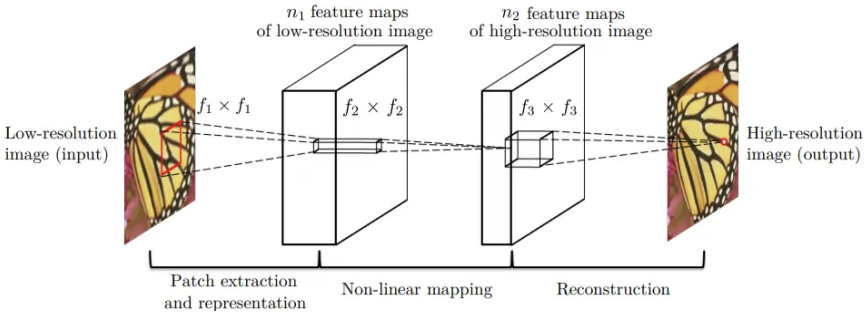

Figure 2.5: Architecture of SRCNN for image super resolution

an effective mapping between the two. CNNs achieve this by leveraging deep, hierarchical feature extraction and end-to-end learning, resulting in substantially improved accuracy over traditional methods.

As shown in Figure 2.5 CNN-based SISR methods [1], [7], [15] typically follow a pipeline comprising three main stages: feature extraction, non-linear mapping, and image reconstruction.

The process begins with an LR image, often preprocessed through normalisation layer [27] and, in some cases, conversion to the luminance (Y) channel to align with standard evaluation practices. The network first extracts low-level spatial features using shallow convolutional layers, which are designed to capture basic textures and patterns. These features are then passed through a sequence of deeper convolutional layers or residual blocks that perform complex nonlinear transformations. These intermediate layers are crucial for capturing hierarchical representations of image content, such as edges, contours, and fine-grained structures that are typically lost during the downsampling process.

The network subsequently reconstructs the HR image using an upsampling module. Upsampling can be achieved using several techniques, such as transposed convolution, interpolation followed by convolution, or more efficiently via sub-pixel convolution (e.g., pixel shuffle) [28]. The output of this stage is an HR, which aims to match the spatial resolution of the ground truth while preserving perceptual quality.

2.4 Transformer

The transformer is a deep learning model designed to capture contextual relationships within sequential data using self-attention mechanisms. Unlike traditional recurrent models, it processes all input elements simultaneously, allowing for greater parallelisation and scalability [14].

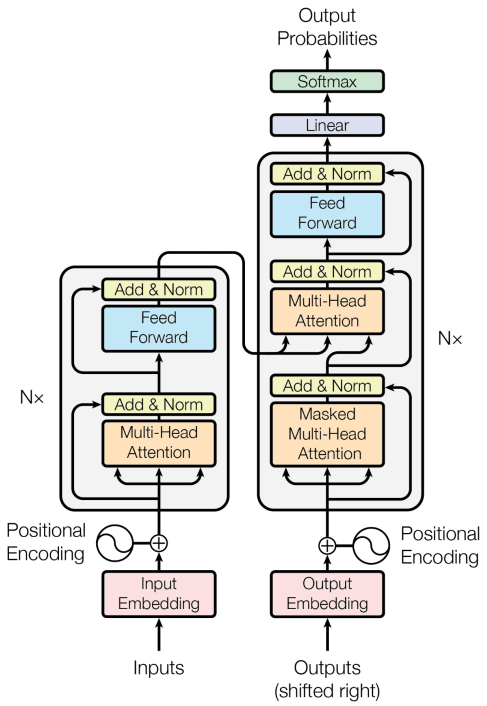

Figure 2.6: Model architecture of the transformer.

At its core, the transformer [14] is composed of an encoder-decoder structure as shown in Figure 2.6. Each encoder layer consists of two primary sublayers: a multi-head self-attention mechanism that enables the model to weigh the importance of each token relative to others in the sequence, and a feed-forward network with non-linear activation (typically ReLU). These sublayers are followed by residual connections and layer normalisation to ensure stable training. Self-attention mechanisms help the network to focus on

relevant spatial regions, regardless of their proximity in the image grid. Unlike convolutions, which are inherently local and require deeper networks or dilation to model wider context, self-attention can dynamically relate distant pixels within an image. This is particularly beneficial for restoring fine textures and edges, where context from far-apart regions improves reconstruction quality. When integrated into transformer-based models like SwinIR [12], [13], self-attention enables hierarchical learning with both local and global dependencies, enhancing the ability of the network to reconstruct high-fidelity images from low-resolution inputs. The decoder mirrors this structure but, it includes an additional attention layer that allows it to focus on the encoder's output.

Specifically, each decoder layer contains: (1) masked multi-head self-attention block, which ensures autoregressive generation by attending only to preceding tokens; (2) encoder-decoder attention block, which performs alignment with the input sequence; and (3) a feed-forward network, which applies position-wise non-linear transformations independently to each token representation within the sequence. Like the encoder, residual connections [21] and layer normalization [27] follow each sublayer. Both encoders and decoders use positional encoding to retain order information since the architecture lacks recurrence or convolution. This mechanism, combined with the multi-head attention structure, allows transformers [14] to effectively capture long-range dependencies.

Swin transformer [12] architecture as shown in Figure 2.7 enhances the traditional transformer model by incorporating a hierarchical structure that utilises a multi-head self-attention mechanism within shifted windows. The Swin transformer architecture [12] introduces a novel approach by segmenting the input image into distinct patches and applying self-attention locally within fixed-size windows.

To facilitate interaction among various regions and capture a wider context, the model utilizes a shifted window strategy, where the windows in successive layers are displaced by a given value. This technique facilitates cross-window information exchange, overcoming

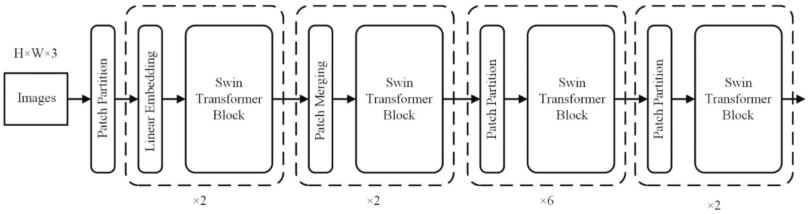

Figure 2.7: Model architecture of the Swin transformer.

limitations of local attention. Internally, the architecture is organised into multiple stages, each consisting of Swin transformer blocks that progressively process the image at increasing levels of abstraction. Between stages, a patch merging layer reduces the spatial resolution while increasing the number of feature channels, similar to downsampling in CNNs. This hierarchical structure enables the model to build multi-scale feature representations that are both rich in semantic content and efficient in terms of memory and computation.

By limiting attention computations to within windows and implementing the shifted windows mechanism, the Swin transformer reduces the computational complexity of global attention from quadratic to linear, thus making it feasible for tasks that require detailed and contextual understanding, such as image classification, object detection, and segmentation.

2.5 SwinIR: Image Restoration using Swin Transformer

SwinIR [13] is a state-of-the-art image restoration framework that integrates the Swin transformer architecture [12] with traditional convolutional layers to enhance degraded images. Designed to address super-resolution, denoising, and JPEG artifact removal, SwinIR leverages both the local detail-capturing ability of CNNs and the global context modeling of transformers [13].

As shown in Figure 2.8, the network [13] is composed of three primary modules: shallow feature extraction, deep feature extraction, and high-quality image reconstruction. The shallow feature extraction module begins the pipeline by applying a 3×3 convolutional

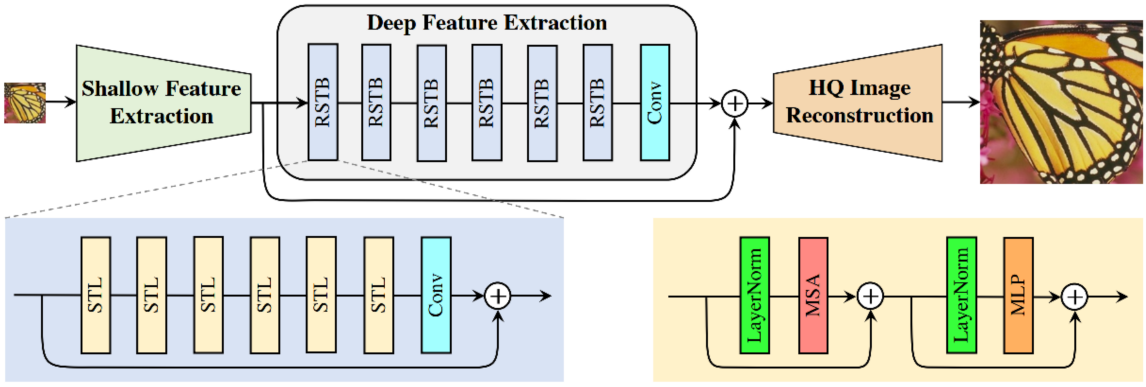

Figure 2.8: The architecture of the SwinIR for image restoration.

layer to the input image. This stage focuses on extracting low-frequency components crucial for preserving basic structure and stability during training [12]. These features are not only processed further but, are also passed directly to the reconstruction module via skip connections to help retain original image details throughout the network. The deep feature extraction module consists of several residual swin tansformer blocks (RSTBs) [13], each comprising multiple Swin transformer layers (STLs) [13]. These STLs operate through local self-attention within shifted non-overlapping windows, enabling efficient learning of both local and long-range dependencies. RSTBs are structured with residual connections and are followed by a convolutional layer, allowing a seamless blend of deep and shallow features. This stage captures high-frequency details and complex spatial dependencies in the image. In the final stage, the reconstruction module combines features from both earlier modules to generate the enhanced high-resolution output. A concluding convolutional layer further refines the output, and the residual skip connections improve the convergence and learning stability. For tasks like JPEG deblocking [29], [30], where up sampling is not required, SwinIR simplifies the architecture by using a single convolution layer in the reconstruction phase [13]. The model's flexible architecture and superior performance across datasets demonstrate its effectiveness for various restoration challenges.

2.6 Inception Module for the Multiscale Network

The inception architecture, as proposed by Szegedy et al. in the Google Net model [19], was developed to efficiently capture features across multiple receptive fields without significantly increasing computational cost. It accomplishes this by executing parallel convolution operations using different kernel sizes (usually 1×1, 3×3, and 5×5). As illustrated in Figure 2.9, this architecture enables the network to analyze both fine-grained and coarse features simultaneously.

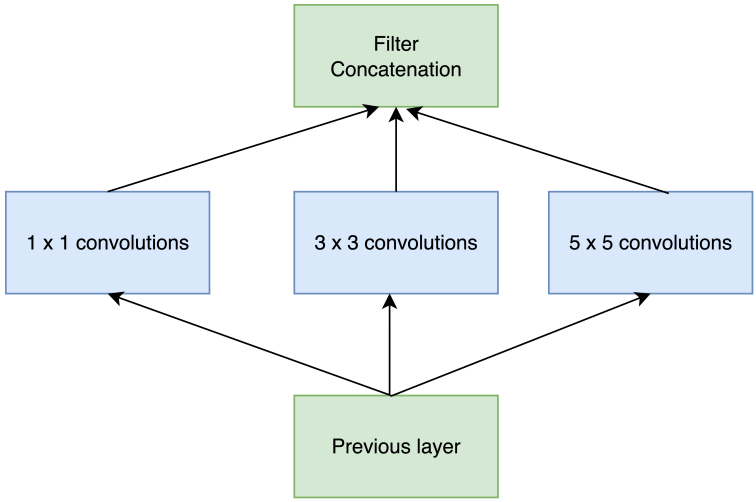

Figure 2.9: An example of an inception module.

To manage the increase in computational load, the architecture incorporates 1×1 convolutions for dimensionality reduction before the larger convolutions, thereby preserving essential information while reducing input size. This parallel multi-scale design enables more robust feature extraction by capturing information at multiple levels of abstraction [19], [20]. As the network progresses through deeper layers, the spatial concentration of features typically decreases, making wider filters more beneficial for learning higher-level representations. The modularity and efficiency of the inception block have led to its wide adoption in vision tasks requiring high representational power with manageable complexity [17], [20]. By incorporating parallel convolutional paths with different receptive fields, a

multi-scale architecture inspired by the inception module can effectively capture both local textures and global structures. This diversity of feature representation contributes to enhanced restoration performance, enabling the network to better reconstruct high-frequency details while maintaining contextual coherence. Such architectural design aligns well with the challenges of SISR, where handling diverse degradation patterns and spatial variations is essential for producing high-quality outputs.

2.7 Summary

This chapter has presented the essential background material foundational to this thesis. The chapter begins with an overview of convolutional neural network followed by the fundamental operations of single image super-resolution. The discussion then covers transformer architecture, including the hierarchical Swin Transformer and the SwinIR image restoration network. The chapter concludes with a description of the inception network module, which is integral to the multiscale network developed in this work.

Chapter 3

Single Image Super Resolution using Enhanced Shallow Feature Extraction Inception-Based Module

3.1 Introduction

SISR remains a key challenge in low-level vision, aiming to reconstruct high-quality images from their degraded low-resolution counterparts. Existing deep learning methods, particularly those based on CNNs and transformers, demonstrate promising results, but they often struggle to capture both local detail and broad contextual information efficiently. The fundamental approach and design of our proposed network have been outlined in our paper, enhanced multiscale network for single image super-resolution [31]. In this chapter, we propose EMS network, an enhanced multiscale inception-based architecture that integrates inception-style convolutional design into the SwinIR architecture as its core framework. This work emphasises the impact of multiscale feature extraction in improving reconstruction quality while maintaining architectural efficiency. The motivation behind EMS Net

is twofold: (i) to improve shallow feature representation through multi-kernel processing, and (ii) to enhance overall reconstruction quality with minimal parameter overhead. By incorporating inception-style multiscale processing, our model enhances the feature extraction efficiency, leveraging both fine-grained and contextual information to improve super-resolution performance. In contrast to conventional methods that rely solely on fixed kernel sizes, our approach dynamically captures features across multiple spatial scales, ensuring robust detail preservation and structural consistency.

In this chapter, we have discussed the architecture of our enhanced multi-scale (EMS) network. Following architecture, we have described each component of the network, including the shallow feature extraction, deep feature extraction, and reconstruction module. At last, we have included the experimental results, ablation studies, comparative evaluations, and qualitative performance analysis.

3.2 Proposed Scheme of EMS Network

In this section, we present the enhanced multiscale (EMS) network, a novel architecture designed to advance SISR by integrating inception-based modules within a Swin transformer framework. By adapting its receptive fields in response to the varying structures present within local image regions, the network concurrently extracts subtle fine details and broader contextual information, thereby enhancing reconstruction quality.

Traditional SISR models often rely on fixed convolutional kernel sizes that restrict the range of spatial features extracted at each layer. In contrast, the EMS network introduces a parallel convolutional design within its shallow feature extraction stage—drawing inspiration from the inception [19] architecture to capture both fine-grained and broad contextual information across multiple scales. This adaptive multiscale feature aggregation, combined with the hierarchical and efficient modeling capabilities of the Swin Transformer, enables the EMS network to overcome limitations inherent in fixed-kernel convolutional

approaches. Additionally, it allows the network to respond more effectively to varied texture patterns and object boundaries within the image, thereby improving the features of the restored output.

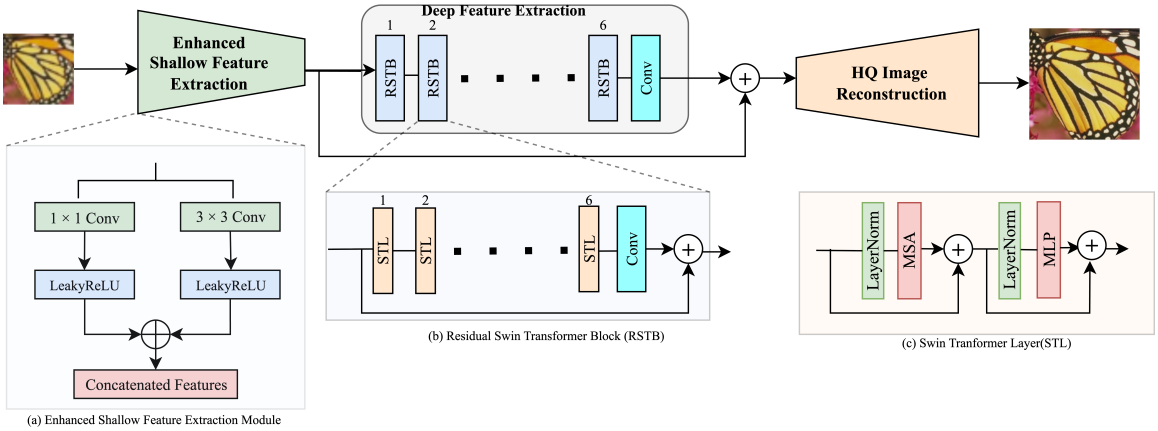

Figure 3.1: Network architecture of our proposed enhanced multiscale inception network for SISR.

The architecture of our EMS network is depicted in the Figure 3.1 and it is composed of three primary elements: an enhanced shallow feature extraction module, a deep feature refinement module using residual Swin transformer blocks (RSTBs), and a high-quality image reconstruction module. The shallow feature block integrates multiple convolution kernels (1×1 and 3×3) to concurrently extract localized details and broader structural cues. These are concatenated to form an enriched representation passed to the deep extraction block. Within the RSTBs, transformer layers enable non-local attention across patches, supporting hierarchical refinement of features without the computational burden of global attention. Finally, the reconstruction module combines the refined features and produces a high-resolution image using an efficient combination of a 3×3 convolution and pixel-shuffle operation.

By leveraging the strengths of both CNN and transformer architectures, our model achieves a strong balance between reconstruction quality and parameter efficiency, demonstrated by its competitive PSNR/SSIM scores across standard benchmarks.

3.2.1 Enhanced shallow feature extraction module

The shallow feature extraction module is designed to efficiently capture both fine-grained and broad contextual information utilizing a parallel convolutional architecture. To achieve this, we designed a parallel convolutional architecture inspired by the inception design introduced by Szegedy et al. [19], which captures multiscale features using filters of varying receptive fields.

Given a low-resolution input image $I_Q \in R^{H \times W \times C_{in}}$, the enhanced shallow feature extraction layer is designed to capture the multiscale spatial features. Specifically, we apply two parallel convolutional layers:

1×1 convolutional layer $f_{SF1}(\cdot)$ and 3×3 convolutional layer $f_{SF2}(\cdot)$ to extract the shallow features. Each layer is followed by a non-linear ReLU activation to introduce non-linearity and suppress negative activations

$$F_{SF1} = \text{ReLU}(f_{SF1}(I_Q)), \tag{3}$$

$$F_{SF2} = \text{ReLU}(f_{SF2}(I_Q)), \tag{4}$$

where $f_{SF1}(\cdot)$ and $f_{SF2}(\cdot)$ focus on the fine details and capture broader contextual structures, respectively. The extracted features are then concatenated to form an enriched multiscale representation

$$F_{SF} = \text{Concat}(F_{SF1}, F_{SF2}), \tag{5}$$

where Concat$(\cdot)$ is the function which concatenates the result from both kernels channel-wise. As illustrated in Figure 3.2 and described by Equations (1) and (2), the shallow feature extraction module processes input patches in parallel by applying convolutional operations. The extracted features are then pixel-wise concatenated as shown in Equation (3),

leading to an increase in channels from 60 to 180, while maintaining a spatial resolution of 64×64. To mitigate this channel expansion, a 1×1 convolution is applied, serving as a bottleneck layer to reduce the number of channels while preserving essential features.

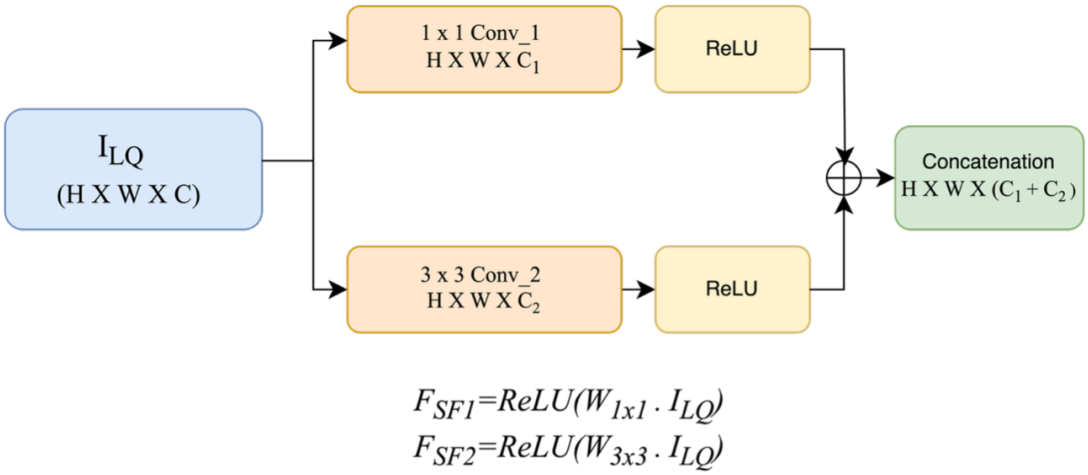

Figure 3.2: Architecture of the inception-inspired enhanced shallow feature module.

$$F_{SF} = \text{ReLU}(f_{\text{bottleneck}}(F_{SF})) \tag{6}$$

where $f_{\text{bottleneck}}(\cdot)$ compresses the concatenated feature map to a manageable channel dimension, allowing efficient downstream processing. This depth-wise fusion enhances the representation quality before passing the features to the deep extraction module.

This design follows the inception module concept. The 1×1 convolution in the initial stage captures fine-grained details of the input, while the 3×3 convolution extracts broader contextual information. The depth-wise concatenation of these feature maps enables a more effective representation for the deep feature extraction layer, enhancing the capture of high-frequency details necessary for subsequent layers.

3.2.2 Deep feature extraction module

The deep feature extraction stage in our network builds upon the architecture proposed in SwinIR, leveraging multiple residual Swin transformer blocks (RSTBs) for hierarchical representation learning. The purpose of this module is to capture high-frequency and non-local contextual information from the enriched shallow features.

The process begins by embedding the input into a tokenized patch representation using a patch embedding module. Given an enriched feature map $F_{SF} \in R^{H \times W \times C}$, the patch embedding layer reshapes it into a sequence of flattened non-overlapping patches. This process is denoted as

$$\mathbf{X} = \text{PatchEmbed}(F_{SF})$$

If absolute positional encoding is enabled, positional information is added

$$\mathbf{X} = \mathbf{X} + \mathbf{P}_{\text{abs}}$$

where $\mathbf{P}_{\text{abs}}$ is the learnable positional embedding matrix. Following this, a dropout operation is applied for regularization.

The core of the deep feature extraction involves 6 stacked RSTBs. Each RSTB contains multiple Swin transformer layers, alternating between window-based multi-head self-attention (W-MSA) and shifted W-MSA (SW-MSA). These layers refine the representation locally while still modeling long-range dependencies. Let $f_{DF}(\cdot)$ represent the deep feature extractor

$$F_{DF} = f_{DF}(F_{SF}) \tag{7}$$

To enhance spatial priors, a 3×3 convolution is applied at the end of the RSTB pipeline. This convolution integrates the spatial consistency back into the feature space, bridging the

transformer output with convolutional semantics. Finally, patch unembedding restores the tensor back into its 2D spatial structure.

Residual Swin Transformer Block (RSTB)

Each RSTB integrates several Swin transformer layers (STLs) to model long-range dependencies while preserving spatial structure. A convolutional layer is placed at the end of each RSTB to reintroduce locality priors into the transformer-based pipeline, ensuring spatial refinement.The output of the RSTB is added to its input via a residual connection.

Swin Transformer Layer (STL)

The STL is the fundamental unit that introduces hierarchical attention-based learning. It works by computing multi-head self-attention within non-overlapping local windows, significantly reducing the computational complexity compared to global self-attention. Each STL consists of two main sub-layers:

(1) Window-based multi-head self-attention (W-MSA) or shifted window attention (SW-MSA)

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}} + B\right)V$$

where B represents the relative positional bias, introduced to capture spatial relationships, while Q, K, and V denote the query, key, and value matrices, respectively.

(2) Feed-Forward Network, multi-layer perceptron (MLP) with GELU activation

$$\text{MLP}(X) = \text{Linear}_2\left(\text{GELU}(\text{Linear}_1(X))\right)$$

Each sub-layer is followed by residual connections and layer normalization:

$$\text{Output} = X + \text{DropPath}(\text{Module}(\text{LayerNorm}(X)))$$

To support cross-window feature interaction, alternating STLs apply a shift to the window partitions. This mechanism, known as SW-MSA, allows information to propagate across window boundaries over layers.

Dimensionality Handling

To manage the dimensional growth due to windowed attention and patch interactions, patch embedding and patch unembedding operations are employed at the start and end of the deep module, respectively. The embedding flattens the spatial structure into sequences, while the unembedding restores the spatial layout for convolutional reconstruction. The final deep output F_{DF} is obtained by passing through a convolutional layer, which ensures compatibility with the upsampling module and reintroduces spatial inductive bias.

3.2.3 Reconstruction module

The image reconstruction stage aggregates the low-frequency and high-frequency information captured from the shallow and deep feature extraction modules, respectively. This fusion is critical for producing a high-quality super-resolved image $I_{HQ} \in R^{H \times W \times C}$. The reconstruction function is expressed as

$$I_{HQ} = f_{\text{IR}}(F_{SF} + F_{DF}) \tag{8}$$

where $f_{\text{IR}}(\cdot)$ denotes the reconstruction block. It consists of a 3×3 convolution layer followed by a pixel shuffle layer [32] to upscale the spatial resolution by a factor s. This ensures that the final image matches the original high-resolution size while maintaining the feature integrity.

For tasks involving standard SR, we adopt a *pixel shuffle* [32] strategy for efficient and artifact-free upscaling. Additionally, the fused feature is passed through a LeakyReLU-activated 1×1 convolution prior to upsampling. This operation reduces computational

load and aids in dimensional alignment for residual learning.

3.2.4 Operational Algorithm of the EMS Network

The EMS network module of the proposed scheme consists of multiple deep networks, each specifically designed for different upscaling factors $s = \{2, 3, 4\}$, addressing single image super-resolution (SISR) across varying scales. The architectural blueprint for each of these networks is illustrated in Figure 3.1, sharing a common structure inspired by the SwinIR framework [13].

Similar to SwinIR, EMS network has been adapted for multiscale image reconstruction tasks, with key modifications introduced to enhance its feature representation capability. Significantly, the network includes a refined shallow feature extraction module that is architecturally inspired by the inception module. This module analyses input images via two concurrent convolutional branches. The first branch applies a 1×1 convolution with 180 filters to focus on fine-grained spatial details. The second parallel branch uses a convolution, also with 180 filters, to capture broader contextual features. The outputs from these branches are concatenated along the channel dimension, forming a 360-dimensional tensor. A subsequent 1×1 convolution is then applied to project this tensor back to the original embedding size of 180 channels, ensuring uniformity for the downstream stages while maintaining efficiency.

Unlike traditional fully connected layers that tend to collapse spatial structure, EMS network—like other modern SISR networks uses point-wise linear layers within the MLPs of its transformer blocks. These layers act independently on each token (i.e., flattened image patch), thus preserving spatial correspondence and allowing the network to retain fine image structures vital for high-quality reconstruction.

Following the shallow module, the features are passed into the deep feature extraction

stage, which comprises 6 residual Swin transformer blocks (RSTBs). Each RSTB contains 6 Swin transformer layers (STLs), operating on 180-dimensional embeddings. As discussed above, within each STL, attention mechanisms alternate between window-based multi-Head self-attention (W-MSA) and shifted window attention (SW-MSA) to model both local and global dependencies. To reintroduce spatial priors lost in the flattened attention operations, a final 3 x 3 convolution with 180 filters is appended to the output of the last RSTB. This deep architecture ensures that the network effectively balances representational capacity and computational complexity. Finally, in the reconstruction module, the refined features from the deep and shallow modules are fused and passed through. Each deep network employed in the EMS module is trained with pairs of degraded low-resolution images y[m,n] and their ground truth version x[m,n] using the L1 pixel loss function for a scale of 4. The loss function can be represented as

$$L_{\text{REC}}(\theta) = \frac{1}{N} \left\| f_{\text{ESN}}(I^{LQ}) - I^{HQ} \right\|_1 \tag{9}$$

where $i = 1, 2, 3, \ldots, I$ (depth of the network), and ε is empirically set to 10^{-9} for numerical stability. The algorithm for the EMS network is summarized below:

Algorithm 1 Learning Algorithm of EMS network

Input: Low-resolution images $I_{LQ} \in R^{H \times W \times C}$ and their corresponding high-resolution ground truth image I_{HQ} from the training dataset.

Output: Trained model parameters θ^* of the proposed EMS network.

Initialize the EMS network consisting of three modules: shallow feature extraction, deep feature extraction, and high-quality image reconstruction.

Train the network $f_{SFE}(\cdot)$ (enhanced shallow feature extraction) using the loss function L_{REC}.

for $i = 1, 2, \ldots, I$ **do**

    Train the networks. Apply 1×1 and 3×3 convolution layers on I_{LQ}, which is obtained by downsampling input high-resolution images using scale factors $s = \{2, 3, 4\}$, followed by ReLU activation.

    Concatenate the outputs channel-wise and fuse them using a 1×1 convolution layer to generate shallow feature F_{SF}.

    Generate deep features F_{DF} by: applying patch embedding to F_{SF}, adding absolute positional encodings, passing through residual swin transformer blocks (RSTBs), and applying patch unembedding.

    Fuse F_{SF} and F_{DF}, then reconstruct the high-quality image I_{HQ} via a 3×3 convolution and pixel-shuffle upsampling.

end for

Return: optimized parameters θ of the proposed EMS network.

3.3 Experimental Results

This section presents the experimental results of the proposed EMS network architecture. A comprehensive ablation study was conducted to evaluate the effectiveness and contribution of each design component. Furthermore, a comparative analysis was carried

out against several state-of-the-art single-image super-resolution (SISR) networks to assess the performance of EMS network.

The EMS network was trained using the DIV2K dataset [33], which comprises 800 high-resolution (HR) images. For training purposes, degraded versions of these images were generated using bicubic downsampling with scaling factors $s \in \{2, 3, 4\}$, consistent with the multiscale design of EMS network. Downsampling was performed using MATLAB, and distinct folders were created for each scaling factor to organise the training data accordingly.

During training, image patches of size 64×64 were extracted from the low-resolution (LR) images along with their HR counterparts. Data augmentation techniques such as random rotation and horizontal flipping were used to improve model invariance to spatial transformations. The model was trained for multiple scales to support scale-agnostic super-resolution within a unified architecture. For evaluation, widely used benchmark datasets were used, including Set5 [34], Set14 [35], BSD100 [36], and Urban100 [37]. Each dataset consists of diverse scenes and structural complexities, facilitating a thorough performance assessment across various contexts.

The metrics used to measure the performance of the proposed scheme are peak signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM). PSNR is defined as

$$\text{PSNR} = 10 \log_{10} \left(\frac{\text{MAX}^2}{\text{MSE}} \right), \tag{10}$$

where MAX is the maximum possible pixel value (typically 255 for 8-bit images) and MSE is the mean squared error between the predicted and ground truth images. SSIM measures perceptual similarity between the predicted and ground truth images and considers luminance, contrast, and structural information in the images for the computation of this metric. It is defined as

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}, \tag{11}$$

where μ_x and μ_y are the local means, σ_x^2 and σ_y^2 are the variances of the two images x and y, respectively, σ_{xy} is the covariance between x and y, and C_1 and C_2 are constants to avoid numerical instability in the computation of the metric.

Most existing SISR networks are trained for a single scale factor, requiring separate models for each resolution task. Unlike conventional approaches, the EMS network integrates an inception-inspired shallow module with parallel 1×1 and 3×3 convolutions, enabling simultaneous extraction of local and contextual features. This multiscale representation enhances scale adaptability while maintaining low computational complexity within a unified framework. Each EMS network variant (for scales ×2, ×3, ×4) was trained independently using the Adam optimiser with an initial learning rate of 10^{-4}. The total number of training iterations was 18K.The learning rate was decayed by a factor of 0.5 after every 60K iterations. All training was conducted on a system equipped with an NVIDIA A6000 GPU. The performance of the proposed model was evaluated using standard image quality metrics PSNR and SSIM. Additionally, PSNR-Y, which computes PSNR on the luminance (Y) channel in the YCbCr colour space, was used to align with conventional SISR evaluation. The results, detailed in the subsequent tables, report PSNR and SSIM values for different scales across benchmark datasets. The ablation studies demonstrate the utility of EMS network's architectural components, and the comparative analysis confirms that EMS network consistently outperforms or matches the performance of leading SISR models across all scales.

3.3.1 Ablation Study Results

To evaluate the robustness and efficiency of the proposed EMS network, we performed an ablation study by reducing the number of RSTB blocks and decreasing the embedding

dimension. Specifically, the number of RSTBs was reduced to 4, and the embedding dimension was lowered to 60 channels. For a scaling factor of 4, this configuration achieved a PSNR of 32.33 and an SSIM of 0.8961 on the Set5 [34] benchmark dataset. We refer to this variant of the proposed scheme as Variant 1. Table 3.1 shows the results of proposed EMS network and its Variant 1 on the images of Set5 datasets for the scale of 4.While this represents a slight performance drop, it demonstrates that EMS network maintains competitive results even under reduced capacity, highlighting the effectiveness of the overall architecture. To assess the efficiency of the suggested approach, we analyze its performance alongside a baseline variant under the same training conditions.

Table 3.1: Impact of training a single deep network with a downsampling scale of 4 reduced RSTB blocks. Performances are in terms of PSNR-Y/SSIM.

Method	**RSTBs**	**Channels**	**PSNR (set5, x4)**	**SSIM (set5, x4)**
Proposed Scheme	6	180	32.88	0.9041
Variant 1	4	60	32.33	0.8961

Additionally, to assess the influence of convolutional kernel size in the shallow feature extraction module, we experimented with a combination of 1×1, 3×3, and 5×5 convolutional kernels for our Variant 2. The model integrating all three kernel sizes yielded a PSNR of 32.28 on Set5 with the ×4 scale, as shown in Table 3.2. The observation suggests that including larger kernel sizes introduces redundancy and does not significantly benefit performance in the presence of the multi-branch inception design already employed in EMS network.

Table 3.2: Impact of training a single deep network with a downsampling scale of 4, and RSTBs as 6 and 3 parallel kernels in the enhanced feature extraction module.

Method	**Kernels (feature extraction layer)**	**PSNR (set5, x4)**	**SSIM (set5, x4)**
Proposed Scheme	$1 \times 1 \oplus 3 \times 3$	32.88	0.9041
Variant 2	$1 \times 1 \oplus 3 \times 3 \oplus 5 \times 5$	32.28	0.8950

Table 3.3: Impact of integrating a deep-shallow feature extraction module with three parallel convolutional branches and 6 RSTBs.

Method	RSTBs	Channels	PSNR (set5, x4)	SSIM (set5, x4)
Proposed Scheme	6	120	32.88	0.9041
Variant 3	4	60	32.35	0.8955

In this 3rd Variant as shown in Table 3.3, we extended the shallow feature extraction block by integrating an additional convolutional refinement step. Specifically, parallel 1×1 and 3×3 convolutional layers were first applied to the input image to capture fine-grained and contextual features, respectively. These outputs were concatenated channel-wise and further processed through a point-wise convolution (1×1) to project them back to the base embedding dimension. To deepen this initial representation, an additional concatenation was introduced by fusing the processed features with another convolutional output. This enhanced feature map—formed through multi-level fusion was then forwarded to the deep feature extraction module composed of RSTBs. We referred to this variant as the deep-shallow feature extraction block. Despite the more elaborate design, this configuration did not yield performance improvements over the baseline. Training plateaued with a PSNR of 32.35 and SSIM of 0.8955 on the Set5 test dataset for a scale factor of 4. The model exhibited signs of overfitting and convergence stagnation, indicating that the added complexity did not translate into meaningful gains in reconstruction quality.

Table 3.4: Table showing PSNR-Y and SSIM comparisons between our proposed EMS network and its 4^{th} variant on set-5 with $\times 2$ downsampled images.

Method	Kernels (feature extraction layer)	PSNR (set5, x2)	SSIM (set5, x2)
Proposed Scheme	$1 \times 1 \oplus 3 \times 3$	38.41	0.9622
Variant 4	$3 \times 3 \oplus 5 \times 5$	38.29	0.9617

In our Variant 4, the shallow feature extraction module was modified to use 3×3 and

5×5 convolutional kernels instead of the proposed 1×1 and 3×3 combination. The network was trained on bicubic downsampled images at scale 2 using the Set5 dataset for evaluation. This modification aimed to explore the impact of removing the fine-grained feature mapping typically provided by the 1×1 convolution layer.While the model achieved a strong PSNR of 38.29 dB and SSIM of 0.9617, it slightly underperformed compared to the EMS network's original configuration (PSNR: 38.72 dB, SSIM: 0.9648). This performance drop can be attributed to the absence of the 1×1 convolution, which plays a crucial role in capturing fine-level spatial correlations and enabling efficient channel mixing. Without this localized mapping, the broader 5×5 kernel may introduce redundant contextual information, thereby diluting fine details necessary for accurate super-resolution.

The results confirm that the inclusion of both fine (1×1) and broad (3×3) kernels, as adopted in the proposed EMS network, yields better high-fidelity reconstruction by balancing detail preservation and contextual abstraction.

Table 3.5: Table showing PSNR-Y and SSIM comparisons between our proposed EMS network and its 5^{th} variant on Set5 with $\times 2$ downsampled images.

Method	**Kernels (Shallow feature extraction layer)**	**PSNR (set5, x2)**	**SSIM (set5, x2)**
Proposed Scheme	$1 \times 1 \oplus 3 \times 3$	38.41	0.9622
	$1 \times 1 \oplus 3 \times 3 \oplus 5 \times 5 \oplus 7 \times 7$	38.27	0.9618

For our final variant experiment, we extended the shallow feature extraction module to incorporate three parallel convolutional paths with kernel sizes of 1×1, 3×3, 5×5 and 7x7, aiming to capture fine, medium, and coarse spatial features respectively. The design was motivated by the assumption that combining diverse receptive fields would enhance multiscale representation learning. The network was trained on bicubic downsampled images at scale ×2 and evaluated on the Set5 dataset. We can observe from Table 3.5, that the performance plateaued early in training and failed to outperform the proposed 1×1 and 3×3 configuration. One possible explanation is that, while the 5×5 and 7x7 convolution

introduces a broader receptive field, it may also introduce redundant or overly smoothed features, which can be detrimental in low-scale (×2) restoration tasks where finer structural details are crucial. In contrast, the 1×1 convolution captures pixel-wise local dependencies and facilitates channel-wise interactions without spatial blurring, and the 3×3 convolution effectively models local textures. Their synergy appears optimal for reconstructing high-resolution details from modestly degraded inputs. After evaluating multiple configurations, the results consistently indicate that the 1×1 and 3×3 parallel convolution setup in EMS network provides the best trade-off between feature richness and computational efficiency for ×2 scale SISR.

3.3.2 Comparative Study Results

In this section, we present a comprehensive comparison of the proposed EMS network with several state-of-the-art SISR methods, across multiple benchmark datasets—Set5 [34], Set14 [35], Urban100 [37], and BSD100 [36]—for scale factors of 2, 3, and 4. All methods included in the comparison are based on deep learning architectures, with particular emphasis on recent transformer-based networks that have shown strong performance in image restoration tasks.

While early SISR models such as SRCNN [1] introduced the idea of learning end-to-end mappings from low-resolution to high-resolution space, recent advancements, especially those based on the transformer architecture (e.g., SwinIR [13]), have demonstrated significant improvements by enabling non-local attention and hierarchical feature refinement. EMS network builds upon this foundation by incorporating inception-inspired multiscale convolutional modules for enhanced feature extraction while maintaining the spatial structure via the Swin transformer [12] backbone.

For consistency and fairness, we have utilised the official implementations provided by the respective authors for all comparative methods. All models were trained on the

DIV2K [33] dataset and tested using the standard benchmarks. Performance was evaluated using PSNR and SSIM metrics, calculated on the Y-channel of the YCbCr colour space, following conventional practices.

The Table 3.6 represents the average values of PSNR and SSIM. The best results for each benchmark and scale are highlighted in red, while the second-best are marked in blue. As evident, EMS network consistently outperforms existing state-of-the-art methods like RCAN [15], SAN [38], HAN [39] across most scales and datasets. Notably, for scale ×2, EMS network achieves a PSNR of 38.41 on Set5, surpassing SwinIR by a margin of 0.06 dB. Similar trends are observed for scales ×3 and ×4, where EMS network maintains superior or competitive performance while using a more efficient and structurally aware design.

These results validate the effectiveness of EMS network's inception-based multi-scale shallow feature module and its seamless integration with the transformer architecture. The observed improvements in reconstruction quality reinforce our network's capability to generalize across diverse degradation scales and image types.

In addition to qualitative comparisons, we also evaluate the proposed EMS network on computational efficiency and parameter complexity. Table 3.7 presents a quantitative comparison of various state-of-the-art SISR models with respect to model size (number of parameters), inference time, FLOPs, and corresponding PSNR/SSIM values for a scaling factor of ×4 on the Set5 dataset.

All experiments were conducted on a machine equipped with an NVIDIA A6000 GPU and 45 GiB of RAM, with inference time averaged over input tensors of size 64×64. Despite introducing only a slight parameter overhead compared to SwinIR (11.96M vs. 11.90M), EMS achieves a PSNR/SSIM of 32.88/0.9041, outperforming SwinIR by +0.16 dB in PSNR and +0.0020 in SSIM. These gains signify notable improvements in both perceptual quality and structural fidelity of the reconstructed images, especially at high

Table 3.6: Quantitative comparison (average **PSNR/SSIM**) with state-of-the-art methods for **SISR** for classical image SR. All methods are trained on the DIV2K dataset. The quantiles representing the best and second-best performances are indicated in red and blue coloured fonts, respectively.

Method	Scale	Set5		Set14		Urban 100		BSD100	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
RCAN [15]	x2	38.27	0.9614	34.12	0.9216	33.34	0.9384	32.41	0.9027
SAN [38]	x2	38.31	0.9620	34.07	0.9213	33.10	0.9370	32.42	0.9028
IGNN [40]	x2	38.24	0.9613	34.07	0.9217	33.23	0.9383	32.41	0.9025
HAN [39]	x2	38.27	0.9614	34.16	0.9217	33.35	0.9385	32.41	0.9027
NLSA [41]	x2	38.34	0.9618	34.08	0.9231	33.42	0.9394	32.43	0.9027
SwinIR [13]	x2	38.35	0.9620	34.14	0.9227	33.40	0.9393	32.44	0.9030
EMS [31]	x2	38.41	0.9622	34.41	0.9251	33.76	0.9423	32.65	0.9033
RCAN [15]	x3	34.74	0.9299	30.65	0.8482	29.09	0.8702	29.32	0.8111
SAN [38]	x3	34.75	0.9300	30.59	0.8476	28.93	0.8671	29.33	0.8112
IGNN [40]	x3	34.72	0.9298	30.66	0.8484	29.03	0.8696	29.31	0.8105
HAN [39]	x3	34.75	0.9299	30.67	0.8483	29.10	0.8705	29.32	0.8110
NLSA [41]	x3	34.85	0.9306	30.70	0.8485	29.25	0.8726	29.34	0.8117
SwinIR [13]	x3	34.89	0.9312	30.77	0.8503	29.29	0.8744	29.37	0.8124
EMS [31]	x3	34.96	0.9316	30.87	0.8531	29.70	0.8816	29.45	0.8127
RCAN [15]	x4	32.63	0.9002	28.87	0.7889	26.82	0.8087	27.77	0.7436
SAN [38]	x4	32.64	0.9003	28.92	0.7888	26.79	0.8068	27.78	0.7436
IGNN [40]	x4	32.57	0.8998	28.85	0.7891	26.84	0.8090	27.77	0.7434
HAN [39]	x4	32.64	0.9002	28.90	0.7890	26.85	0.8094	27.80	0.7442
NLSA [41]	x4	32.59	0.9000	28.87	0.7891	26.96	0.8109	27.78	0.7444
SwinIR [13]	x4	32.72	0.9021	28.94	0.7914	27.07	0.8164	27.83	0.7459
EMS [31]	x4	32.88	0.9041	29.08	0.7951	27.45	0.8244	27.92	0.7463

scaling factors. More importantly, EMS matches SwinIR in inference time (22.32 ms vs.

Table 3.7: Quantitative Comparison of SISR Methods for Performance and Efficiency

Network	**RCAN**	**SAN**	**HAN**	**SwinIR**	**EMS(Net)**
#params	15.59M	15.86M	25.92M	11.90M	11.96M
Inference Time	57.86	46.47	53.24	22.26	22.32
Flops	65.25	66.61	64.78	53.83	54.10
PSNR/SSIM	32.63/0.9002	32.64/0.9003	32.64/0.9002	32.72/0.9021	32.88/0.9041

22.26 ms) while maintaining lower FLOPs (54.10 vs. 53.83), indicating that the architectural enhancements do not introduce significant computational overhead. Compared to other heavier networks like RCAN [15], SAN [38], and HAN [39]—which exhibit longer inference times and higher FLOPs—EMS demonstrates superior accuracy and efficiency.

These observations affirm that the proposed inception-inspired multi-scale shallow feature extraction in EMS, combined with transformer-based deep refinements, offers an optimal trade-off between performance and computational cost. The findings validate that our architectural choices enhance high-resolution reconstruction without compromising efficiency, making EMS a practical and effective solution for real-world SISR applications.

3.3.3 Qualitative Performance and Comparison

To further evaluate the efficacy of the proposed EMS network, this section presents a qualitative comparison with several state-of-the-art single image super-resolution (SISR) methods. The visual outputs generated by EMS network are compared against leading multiscale transformer-based networks, including RCAN [15], SAN [38], HAN [39], and SwinIR [13].

Figures 3.3, 3.4 and 3.5 showcase representative image samples from the benchmark datasets, each downsampled with a scale factor of 4 using bicubic degradation. These visual results provide a comparative understanding of the perceptual quality delivered by

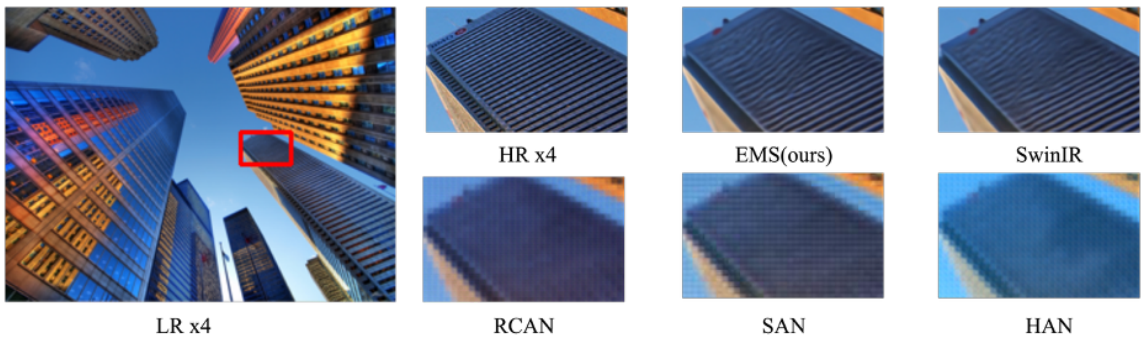

Figure 3.3: Visual comparison of bicubic image SR (Scale 4) methods. Compared images are derived from the Urban100 test dataset.

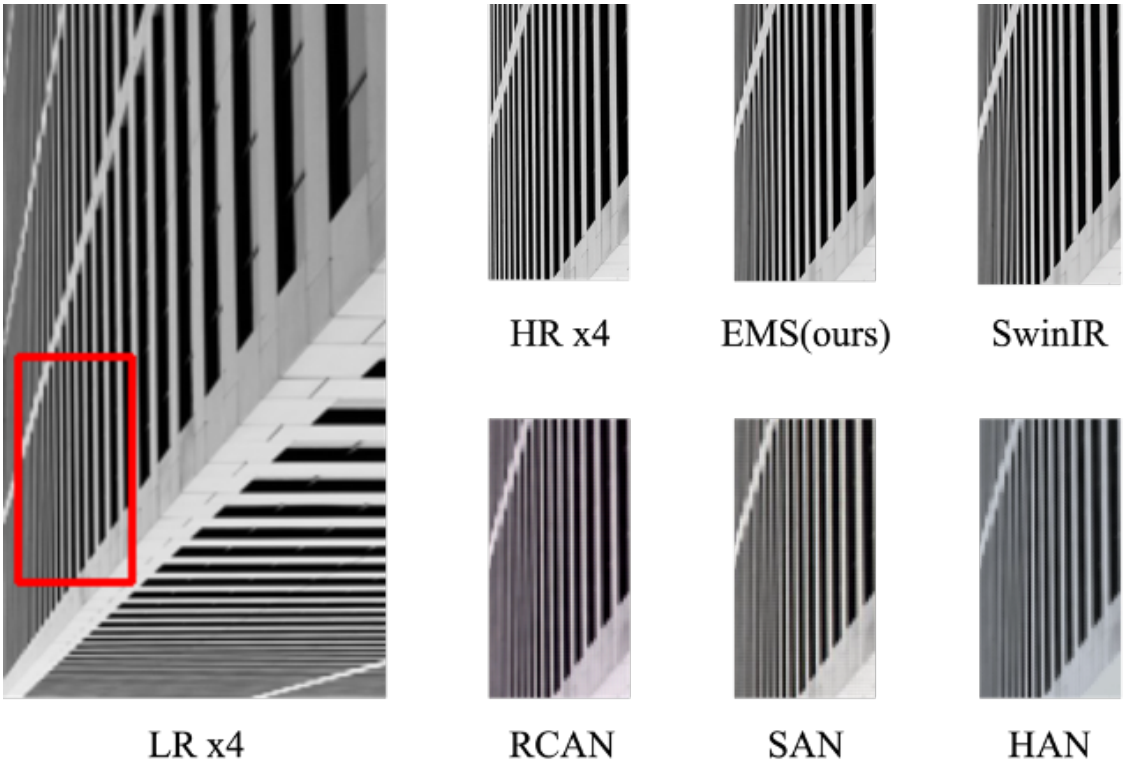

Figure 3.4: Qualitative comparison

the respective networks.

From the qualitative evaluation, it can be observed that EMS network consistently reconstructs sharper edges, finer textures, and visually more coherent structures than the other methods. The enhanced spatial detail recovery and reduced visual artefacts validate

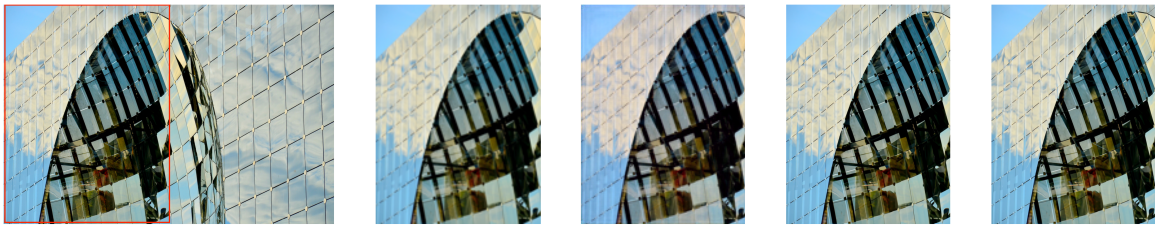

Figure 3.5: Qualitative comparison of super-resolved images generated by RCAN, SwinIR, and the proposed EMS network on a cropped region of Urban100 dataset.

the superiority of EMS network in high-scale image reconstruction tasks. This reinforces the findings of our quantitative analysis and highlights the ability of EMS network to effectively synthesise high-quality super-resolved images in various image conditions.

3.4 Summary

In this chapter, we presented the architectural design and implementation details of the proposed enhanced multi-scale network (EMS network) for SISR. The chapter began with a comprehensive explanation of the overall EMS network framework, highlighting its modular structure composed of three core components: the enhanced shallow feature extraction module, the deep feature extraction module based on RSTBs, and the reconstruction module. We discussed the motivation for incorporating inception-inspired multi-kernel convolutions (1×1 and 3×3) in the shallow feature stage, enabling the network to capture both fine-grained and broad contextual features. Deep feature extraction module was detailed with its use of Swin transformer layers (STLs) operating within RSTBs to model long-range dependencies efficiently while preserving spatial locality. The reconstruction module was introduced as a lightweight yet effective upsampling pipeline that combines learned features through pixel shuffle operations.We also introduced the training methodology, loss functions, and evaluation metrics used, including the L1-Pixel loss and

Charbonnier loss for robust gradient flow. A detailed algorithm summarizing the EMS network training pipeline was presented. The latter part of the chapter focused on experimental results, ablation studies, and comparisons. Several ablation experiments were conducted to examine the impact of architectural variants—including different kernel configurations and fusion strategies—on the network's performance. The results demonstrated that the 1×1 and 3×3 convolution combination consistently outperformed other variants in both PSNR and SSIM, with all the scale factors of x2,x3 and x4.

Chapter 4

Conclusion

4.1 Concluding Remarks

Though numerous state-of-the-art networks have demonstrated commendable performance on the task of SISR, they tend to exhibit a consistent degradation in performance as the scaling factor increases. Most traditional and deep learning-based approaches, while effective for fixed-scale restoration, often struggle to preserve finer spatial structures under large scaling factors due to their limited capacity to capture diverse feature representations without substantially increasing model complexity. In contrast, the proposed EMS network introduces an efficient and scalable architecture that addresses this limitation. The EMS network is composed of three core modules: an enhanced shallow feature extraction module, a deep feature extraction module based on Swin transformer blocks, and a reconstruction module. As elaborated in Chapter 3, the design leverages a dual-branch convolution strategy in the shallow feature extractor, utilizing 1×1 and 3×3 kernels to effectively capture both fine-grained and coarse image structures. This design choice enables the network to extract features at multiple receptive fields without incurring a significant computational burden.

Our ablation studies have validated that alternative kernel configurations, such as replacing the 1×1 convolution with larger kernels (e.g., 5×5), did not yield performance improvements and often plateaued at lower PSNR and SSIM scores. These experiments reinforce the idea that our choice of kernel sizes was optimal for the task, particularly with higher scaling factors. In addition, EMS network [31] demonstrates robustness across multiple degradation levels and consistently outperforms baseline models in benchmark datasets, as seen in our comparative evaluations.

Additionally, we examined the computational efficiency of the network through detailed FLOPs analysis and inference time comparisons. Despite the minimal increase in model parameters, 0.5% over SwinIR, EMS network achieves higher reconstruction quality—demonstrated by consistent PSNR gains—especially under 4 × scaling conditions. This efficiency-performance balance is crucial for real-world deployment scenarios, where both accuracy and latency are critical.

In summary, EMS network successfully addresses the core challenges of SISR by combining a lightweight multi-scale feature extraction strategy with transformer-based context modeling. The empirical results and ablation studies presented in this thesis validate the architectural decisions made, showcasing EMS network as a scalable and computationally efficient solution for high-quality single image super-resolution.

4.2 Scope for Further Investigations

While EMS has demonstrated competitive performance across multiple SISR tasks and scaling factors, there remain several promising directions for extending and enhancing this research. One such direction involves the integration of dilated convolutions within the shallow or deep feature extraction stages. These convolutions can expand the receptive field without increasing the number of parameters, potentially improving contextual understanding and feature aggregation.

Another avenue worth exploring is the incorporation of multi-layer residual connections, which could improve gradient flow and facilitate deeper network training, further enhancing the network's ability to reconstruct high-frequency image details. Additionally, the use of gated recurrent units (GRUs) or other recurrent elements could be investigated to refine spatial feature dependencies and improve consistency in sequential image reconstruction tasks such as video super-resolution.

Developments such as ghost convolutions, which aim to reduce redundancy in feature maps and accelerate computation, also present an opportunity to build more lightweight architectures without compromising performance. Future versions of EMS could benefit from adopting such components, especially for deployment in resource-constrained environments.

Moreover, training on real-world degraded datasets or adopting real-image super-resolution paradigms (beyond bicubic down sampling) would enhance the network's practical applicability. This includes expanding the dataset to better model realistic degradation patterns, improving generalisability across diverse imaging conditions.

Finally, further comparative analysis against state-of-the-art models using additional evaluation metrics (e.g., perceptual quality indices, FLOPs, and inference latency) will help benchmark the network's efficiency and effectiveness. These enhancements would not only strengthen the robustness of the EMS network but also make it more adaptable to real-world deployment scenarios in mobile and edge devices.

References

[1] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 2, pp. 295–307, February 2016.

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

[3] R. Keys, "Cubic convolution interpolation for digital image processing," *IEEE transactions on Acoustics, Speech, and Signal Processing*, vol. 29, no. 6, pp. 1153–1160, December 1981.

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

[5] Y. Mei, Y. Fan, and Y. Zhou, "Image super-resolution with non-local sparse attention," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, June 2021, pp. 3517–3526.

[6] 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*, June 2016, pp. 1646–1654.

[7] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee, "Enhanced deep residual networks for single image super-resolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 136–144.

[8] 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 International Conference on Computer Vision and Pattern Recognition*, October 2017, pp. 4681–4690.

[9] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. Change Loy, "Esrgan: Enhanced super-resolution generative adversarial networks," in *Proceedings of the European Conference on Computer Vision ECCV Workshops*, 2018, pp. 1–16.

[10] H. Chen, Y. Wang, T. Guo, C. Xu, Y. Deng, Z. Liu, S. Ma, C. Xu, C. Xu, and W. Gao, "Pre-trained image processing transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12 299–12 310.

[11] Z. Lu, J. Li, H. Liu, C. Huang, L. Zhang, and T. Zeng, "Transformer for single image super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops(CVPRW)*, June 2022, pp. 457–466.

[12] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE International Conference on Computer Vision*, 2021, pp. 10 012–10 022.

[13] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "Swinir: Image restoration using swin transformer," in *Proceedings of the IEEE International Conference on Computer Vision*, 2021, pp. 1833–1844.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the 31st International*

Conference on Neural Information Processing Systems, Long Beach, USA, 2017, pp. 6000–6010.

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

[16] Y. Tai, J. Yang, X. Liu, and C. Xu, "Memnet: A persistent memory network for image restoration," in *Proceedings of the IEEE Conference on Computer Vision*, 2017, pp. 4539–4547.

[17] Y. Wang, Y. Li, G. Wang, and X. Liu, "Multi-scale attention network for single image super-resolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2024, pp. 5950–5960.

[18] W. Zou, H. Gao, L. Chen, Y. Zhang, M. Jiang, Z. Yu, and M. Tan, "Cross-view hierarchy network for stereo image super-resolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops(CVPRW)*, June 2023, pp. 1396–1405.

[19] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, June 2015, pp. 1–9.

[20] W. Muhammad and S. Aramvith, "Multi-scale inception based super-resolution using deep learning approach," *Electronics*, vol. 8, no. 8, p. 892, 2019.

[21] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," in *Proceedings of the IEEE*, vol. 86, no. 11, November 1998, pp. 2278–2324.

[22] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[23] K. Zhang, Y. Li, W. Zuo, L. Zhang, L. Van Gool, and R. Timofte, "Plug-and-play image restoration with deep denoiser prior," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 10, pp. 6360–6376, 2021.

[24] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image segmentation using deep learning: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 7, pp. 3523–3542, July 2022.

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

[26] C. E. Duchon, "Lanczos filtering in one and two dimensions," *Journal of Applied Meteorology*, pp. 1016–1022, 1979.

[27] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, July 2016.

[28] J. Li, Z. Pei, W. Li, G. Gao, L. Wang, Y. Wang, and T. Zeng, "A systematic survey of deep learning-based single-image super-resolution," *ACM Computing Surveys*, vol. 56, no. 10, pp. 1–40, October 2024.

[29] C. Dong, Y. Deng, C. C. Loy, and X. Tang, "Compression artifacts reduction by a deep convolutional network," in *Proceedings of the IEEE International Conference on Computer Vision(ICCV)*, December 2015, pp. 576–584.

[30] Y. Kim, J. W. Soh, J. Park, B. Ahn, H.-S. Lee, Y.-S. Moon, and N. I. Cho, "A pseudo-blind convolutional neural network for the reduction of compression artifacts," *IEEE*

Transactions on Circuits and Systems for Video Technology, vol. 30, no. 4, pp. 1121–1135, April 2020.

[31] N. Babar and M. O. Ahmad, "Enhanced multiscale network for single image super-resolution," in *Proceedings of the IEEE International Conference on Image Processing(ICIP)*, September 2025.

[32] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient subpixel convolutional neural network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1874–1883.

[33] E. Agustsson and R. Timofte, "Ntire 2017 challenge on single image super-resolution: Dataset and study," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop*, 2017, pp. 126–135.

[34] M. Bevilacqua, A. Roumy, C. Guillemot, and M. L. Alberi-Morel, "Low-complexity single-image super-resolution based on nonnegative neighbor embedding," in *Proceedings of the British Machine Vision Conference*, 2012, pp. 1–10.

[35] R. Zeyde, M. Elad, and M. Protter, "On single image scale-up using sparse-representations," in *Proceedings of the International Conference on Curves and Surfaces*, 2010, pp. 711–730.

[36] 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 of the 8th IEEE International Conference on Computer Vision*, 2001, pp. 416–423.

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

[38] T. Dai, J. Cai, Y. Zhang, S.-T. Xia, and L. Zhang, "Second-order attention network for single image super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, June 2019, pp. 11 065–11 074.

[39] B. Niu, W. Wen, W. Ren, X. Zhang, L. Yang, S. Wang, K. Zhang, X. Cao, and H. Shen, "Single image super-resolution via a holistic attention network," in *Proceedings European Conference on Computer Vision*, January 2020, pp. 191–207.

[40] S. Zhou, J. Zhang, W. Zuo, and C. C. Loy, "Cross-scale internal graph neural network for image super-resolution," in *Proceedings of the International Conference Neural Information Processing Systems*, 2020, pp. 3499–3509.

[41] Y. Mei, Y. Fan, and Y. Zhou, "Image super-resolution with non-local sparse attention," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, June 2021, pp. 3517–3526.