Modern Trends in Deep Learning Based Super- Resolution

Thesis Report Submitted to
Indian Institute of Technology Ropar
in partial fulfilment of the requirements for the
Degree of

DOCTOR OF PHILOSOPHY

By

Nancy Mehta
(Reg.No. 2018eez0017)

Under the guidance of

Dr. Subrahmanyam Murala

Department of Electrical Engineering,

Indian Institute of Technology Ropar

Rupnagar-140001, Punjab, India

2022-23

January-2023

Nancy Mehta: *Modern Trends in Deep Learning Based Super-Resolution*

Dedicated to My Beloved Family

- Who are the inspiration and power behind success of this work

Declaration of Originality

I hereby declare that the work which is being presented in the thesis entitled **MODERN TRENDS IN DEEP LEARNING BASED SUPER-RESOLUTION** has been solely authored by me. It presents the result of my own independent research conducted during the time period from JANUARY-2019 to JANUARY-2023 under the supervision of Dr. Subrahmanyam Murala, Associate Professor, Department of Electrical Engineering. To the best of my knowledge, it is an original work, both in terms of research content and narrative, and has not been submitted or accepted elsewhere, in part or in full, for the award of any degree, diploma, fellowship, associateship, or similar title of any university or institution. Further, due credit has been attributed with appropriate citations and acknowledgments, in line with established ethical norms and practices. I also declare that any idea/data/fact/source stated in my thesis has not been fabricated/ falsified/ misrepresented. All the principles of academic honesty and integrity have been followed. I fully understand that if the thesis is found to be unoriginal, fabricated, or plagiarized, the Institute reserves the right to withdraw the thesis from its archive and revoke the associated Degree conferred. Additionally, the Institute also reserves the right to appraise all concerned sections of society of the matter for their information and necessary action (if any). If accepted, I hereby consent for my thesis to be available online in the Institute's Open Access repository, inter-library loan, and the title & abstract to be made available to outside organizations.

Nancy Mehta

Signature

Name: Nancy Mehta
Entry Number: 2018eez0017
Program: Doctor of Philosophy (Ph.D.)
Department: Electrical Engineering
Indian Institute of Technology Ropar
Rupnagar, Punjab 140001

Date: 23 January 2023

Acknowledgement

No words are sufficient to express my sincere gratitude to my guide, **Dr. Subrahmanyam Murala**, for continuous encouragement, technical discussions, and invaluable guidance during my Ph.D. It has been a great honour and privilege to work under his supervision and to be his student.

I take this opportunity to convey my sincere thanks to **Prof. Rajeev Ahuja**, *Director of Indian Institute of Technology Ropar*, for his all round support, whole-hearted cooperation during the Ph.D journey. I am grateful to the **Head of the Electrical Engineering Department**, for his zealous and encouraging attitude.

I also extend my genuine thanks to doctoral committee members **Dr. Sam Darshi**, **Dr. Puneet Goyal**, **Dr. Deepti Bathula**, and **Dr. Suman Kumar** for their insightful comments and encouragement and also for the hard questions which incented me to widen my research from various perspectives.

I also sincerely thank my collaborators **Dr. Salman Khan**, **Dr. Waqas Zamir**, **and Dr. Fahad Khan** from MBZUAI, Dubai for their valuable suggestions in my work.

My special thanks and appreciation goes to **Dr. Santosh Kumar Vipparthi**, **Prof. Jyotindra Singh Sahambi** from IIT Ropar, and **Dr. Ravibabu Mulaveesala** from IIT Delhi for providing valuable guidance and support.

A special thank to **Dr. Akshay Dudhane**, MBZUAI Dubai for valuable discussions, and more importantly for a constant support during the hard times I had gone through.

I also want to specially thank my seniors **Dr. Sachin Chaudhary**, NTU Singapore, **Dr. Prashant Patil**, Deakin University Australia, **Dr. Praful Hambarde**, Deakin University, Australia, and **Dr. Kirandeep Kaur**, Swinburne University, Australia, for their valuable suggestions during my entire Ph.D. journey.

I wholeheartedly thank my cooperating juniors and colleagues **Mr. Jasdeep Singh**, **Mr. Ashutosh Kulkarni**, **Mr. Chetan Gupta**, **Mr. Omkar Thawakar**, **Mr. Devendera Deshmukh**, **Mr. Nishant Gupta**, and **Mr. Hemant Sharma** for constant encouragement.

I want to sincerely thank my close friends **Ms. Shruti Phutke Patil**, and **Miss Surbhi Gautam** for their motivational and emotional support from the beginning of my journey. My special appreciation to **Miss Akanksha Gupta**, **Miss Sahibia Vohra**, and **Miss Muskan Sharma** for their persistent support.

Finally, my deep and sincere gratitude to my family for their continuous love, prayers, encouragement, blind trust and for understanding and cooperating me throughout my journey. Above all, I praise and thank God for blessing me with success in this research work.

January 2023.
Place : IIT Ropar, Punjab, India. Nancy Mehta

DEPARTMENT OF ELECTRICAL ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY ROPAR
RUPNAGAR-140001, INDIA

Certificate

This is to certify that the thesis entitled **MODERN TRENDS IN DEEP LEARNING BASED SUPER-RESOLUTION**, submitted by **Nancy Mehta (2018eez0017)** for the award of the degree of **Doctor of Philosophy** of Indian Institute of Technology Ropar, Punjab, INDIA, is a record of bonafide research work carried out under my guidance and supervision during 2019-23. To the best of my knowledge and belief, the work presented in this thesis is original and has not been submitted, either in part or full, for the award of any other degree, diploma, fellowship, associateship or similar title of any university or institution.

In my opinion, the thesis has reached the standard fulfilling the requirements of the regulations relating to the degree.

Signature

Dr. Subrahmanyam Murala
Associate Professor
Department of Electrical Engineering
Indian Institute of Technology Ropar
Rupnagar, Punjab 140001
Date: 23 January 2023

Lay Summary

Image Super-Resolution (SR) is the task of reconstructing a low resolution image to generate a plausible looking high-resolution image. For example, we have a image that is taken from a smartphone, and it is corrupted on account of hand tremor. Image super-resolution, specifically burst super-resolution helps to recover this corrupted image by merging information from multiple frames. Also, if we have some old photo that is degraded on account of low-quality cameras or time issues, we can use SR to restore the quality of image.

The existing deep-learning based SR models do not distinguish much about the frequency of image (*e.g.* low and high-frequency) equally across the channels and often lack the discriminative learning ability that limits their ability to implicitly recover high-frequency details. Along with yielding photo-realistic outputs, network size is equally a challenging problem in image SR. The existing models are quite heavy with millions of parameters, and inference through such a complex network demands billions of floating point operations. Recently, multi-frame super-resolution approaches are gaining popularity. They seek the reconstruction of HR images by employing numerous degraded LR images of a scene. But capturing LR images under the burst mode results in sub-pixel shifts among the multiple LR burst images and thereby, generates different LR samplings of the underlying scene. Additionally, in general the existing CNN-based super-resolution (SR) methods are often based on the assumption that the degradations are fixed and known (*e.g.* bicubic downsampling). However, these approaches suffer a severe performance drop when the real degradation is different from the predefined assumption.

In this work, we propose different methods for tackling different problem in image SR. All our approaches make use of novel state-of-the-art concepts for proposing different solutions for each problem. Our proposed methods give visually plausible results with less time as compared to the existing methods.

Abstract

Digital images, an extension of human memory is one of the most important information carrier for human activities. It plays a critical role in many day-to-day applications, from online social networking to commercial advertising to medical images. On account of the constraints in the physical characteristics of the digital sensor, *e.g.* size and density, the resolution of the captured image is limited. In many cases, the limited resolution becomes a barrier for fast and accurate analysis. Thus, it is highly desirable to breach the resolution limitation and acquire high-resolution (HR) digital images. One of the most promising approach is to utilize signal processing techniques for obtaining an HR image from Low-resolution (LR) image, and this resolution enhancement approach is called super-resolution (SR). The major advantage of this software approach is that it costs much less than upgrading hardware and existing camera systems. Over the past decades, many pioneers have developed various algorithms to improve the quality of reconstructed images. More critically, low-resolution images have lesser number of pixels representing an object of interest, thus making it difficult to find the details. SR targets to solve this problem, whereby a given LR image is upscaled to retrieve an image with higher resolution and thus obtain more discernible details that can be employed in downstream tasks like face recognition, and object classification. The common goal of these techniques is to provide finer details than the given low-resolution (LR) image by increasing the number of pixels per unit of space. Additionally, in comparison to DSLR cameras, low-quality images are generally output from portable devices on account of their physical limitations. The synthesized low-quality images usually have multiple degradations - low-resolution owing to small camera sensors, mosaic patterns on account of camera filter array and sub-pixel shifts on account of camera motions. These degradations generally refrain the performance of single image super-resolution for retrieving high-resolution (HR) image from a single low-resolution (LR) image. Considering the above points, the current prevailing deep-learning based super-resolution algorithms often lack in some aspects: they are highly dependent on designing heavy-weight architectures to achieve state-of-the-art (SoTA) results and generally do not take into consideration the real-world degradations. They generally fail in maintaining the balance between spatial details and contextual information, that is the basic requirement for exhibiting superior performance in super-resolution task. We also observe that the recent approaches focus more on feature extraction, without paying much attention to the up-sampling strategies involved. Moreover, the current approaches fail to leverage the advantages of capturing abundant information from multiple LR images. Our work focuses on analysing and designing different solutions for super-resolution task in the context of providing solution to the above mentioned challenges.

The significant contributions of this work are : (1) A novel approach for generating contextually enriched outputs by preserving the required information without any sort of prior information, (2) A novel lightweight approach capable of generating contextually enriched features for image super-resolution and other applications, (3) A novel framework for efficiently merging multiple burst LR RAW images in a coherent and effective way to generate HR RGB outputs with realistic textures and additional high-frequency details, and (4) A novel transformer based blind approach for resolving the real-world degradations. The proposed super-resolution approaches are evaluated on the current SoTA single image SR databases such as Set5 [1], Set14 [2], BSD100 [3], Urban100 [4], DIV2K [5], Flickr2K [6], and burst SR databases such as BurstSR [7], and SyntheticBurst dataset [7] and animated SR database, Manga109 [8]. Also, we evaluate our proposed modules for DND [9], SIDD [10] for the case of single image denoising and color [11] and gray-scale [12] datasets for burst denoising. We utilize LoL [13] and MIT [14] datasets for single image low-light enhancement and SONY dataset for burst low-light enhancement. The qualitative and quantitative results of proposed methods are examined and compared with SoTA hand-crafted and learning based methods. Standard quantitative evaluation parameters such as Structural Similarity Index (SSIM), Peak-to-Signal Ratio (PSNR) and Learned Perceptual Image Patch Similarity (LPIPS) are used to evaluate the proposed super-resolution approaches.

Keywords: Image Super-Resolution, Burst Super-Resolution, Blind Super-Resolution, Multi-Scale Feature Learning, Frequency Extraction, Denoising, and Low-light enhancement.

Abbreviations

CNN	:	Convolutional Neural Network.
FLOPs	:	Floating point operations.
HR	:	High-Resolution.
LR	:	Low-Resolution.
LPIPS	:	Learned Perceptual Image Patch Similarity.
PSNR	:	Peal Signal to Noise Ratio.
SISR	:	Single Image Super-Resolution.
SSIM	:	Structural Similarity Index.
SR	:	Super-Resolution.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In this chapter, the topic of image super-resolution is introduced. The introduction about super-resolution is given in Section 1.1 and the motivation of this work is discussed in Section 1.2. Section 1.3 introduces a few applications of super-resolution and its usability in the real world. The common research challenges in the field of super-resolution are discussed in Section 1.4. The major problems identified in the task of super-resolution are discussed in Section 1.5. In Section 1.6, the main aim and objectives of our research work are discussed. Section 1.7 defines the major contributions of our work. Finally, Section 1.8 provides the overall structure and outline of the thesis.

1.1 Introduction

We all have seen old monochrome pictures (most often grayscale) that have several artefacts, which are then colourised and thereafter made to look like they were recently acquired with a modern camera. It is an example of *image restoration*, that can be generally defined as the process of retrieving the underlying high-quality original image from a corrupted image. Several factors affect the output quality of an image, and the most common are the suboptimal photography conditions (*e.g.*, on account of motion blur, poor lightning conditions), lens properties (*e.g.*, noise, blur, lens flare), and post-processing artefacts (*e.g.*, lossy compression schemes, that perform compression in such a manner that is irreversible and consequently leads to loss of information).

Another factor that affects the image quality is the resolution. More critically, low-resolution (LR) images have lesser number of pixels representing an object of interest, thus making it difficult to find the details. It can be either because the image itself is very small, or because an object is too far from the camera thus causing it to cover a small area within the image. Super-Resolution (SR) targets to solve this problem, whereby a given LR image is up-scaled to retrieve an image with higher resolution for more discernible details that can then be employed in downstream tasks like face recognition, object classification, and so on.

Additionally, extensive use of high-resolution displays, high-definition televisions and hand-held portable devices in our day-to-day life is the primary reason for the recent explosive attention of image super-resolution in the research field and industrial applications. This high-resolution technology is also exploited much for broadcasting of the image content. However, the images while data transmission are often contaminated on account of the medium and dynamics between the cameras, scene elements and various

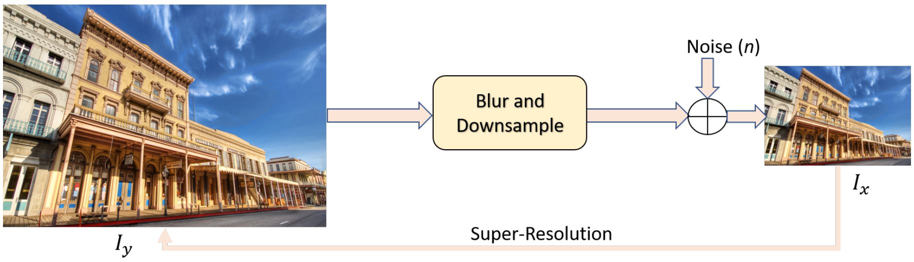


Figure 1.1: Overall framework of SR.

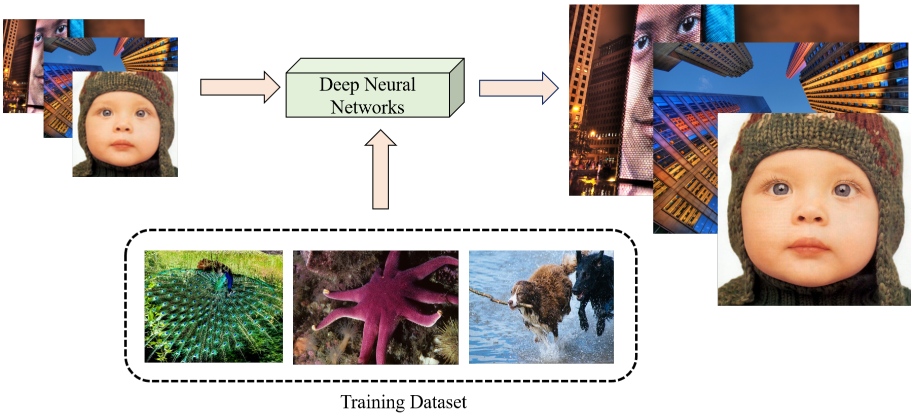


Figure 1.2: The training process of data-driven based deep SR networks.

dependent or independent noises. Therefore, restoration of images is vital for improving their aesthetic quality.

As shown in Figure 1.1, let us consider a low-resolution image, $I_x \in \mathbb{R}^{h\times w}$ and the corresponding ground-truth high-resolution image as $I_y \in \mathbb{R}^{H\times W}$, where $H > h$ and $W > w$. Generally, in a single image super-resolution (SISR) framework as shown in Fig. 1.2, any LR image is modelled as $I_x = D(I_y; \theta_D)$, where D denotes a degradation map $\mathbb{R}^{\mathbb{H}\times\mathbb{W}} \rightarrow \mathbb{R}^{h\times w}$ and θ_D represents the degradation factor. Typically, the researchers try to model the unknown degradation process. The most popular degradation model is:

$$D(I_y; \theta_D) = (I_y \otimes k)\downarrow_s + n \tag{1.1}$$

where, $I_y \otimes k$ denotes the convolution operation between the blur kernel k and HR image I_y. $\downarrow_s$ is a subsequent downsampling operation with scale factor s, and n represents the additive Gaussian noise with standard deviation σ. For the SISR task, it is required to recover an SR image, I_{SR} from the LR image I_X. Thus, the overall task can be formulated as $I_{SR} = F(I_x; \theta_F)$, where F denotes the SR algorithm and θ_F represents the parameter set of SR process.

Recently, the researchers have resorted for SISR as an end-to-end learning task, that relies on massive training data and effective loss functions. Thus, the overall SISR task can be transformed into the following optimization goal:

$$\hat{\theta}_F = \arg_{\theta_F} \min L(I_{SR}, I_y) + \lambda\phi(\theta) \tag{1.2}$$

where, L denotes the loss function between the generated SR image I_{SR} and the HR image I_y, $\phi(\theta)$ is used to represent the regularization term, and λ is the trade-off parameter for controlling the percentage of regularization term. Depending upon the degradation, especially the downsampling operation used to generate the low-resolution images, image super-resolution can be classified into single image and blind image super-resolution.

Additionally, depending upon the number of input LR images used for improving the resolution of images, image super-resolution can be classified into single image super-resolution and multi-frame super-resolution.

1.2 Motivation

For most electronic imaging applications, images with high-resolution are desired and often required. At this point, one natural question is ”what is the need to get into the trouble of developing algorithms for obtaining the results, if we could use better quality cameras?” The answer to this question lies in some practical considerations, for instance though mobile phone cameras capture fairly good quality images, but they yield several imperfections resulting primarily from the need to utilize lenses and sensors which are compact enough for fitting on a phone without making it bulky, while also being comparatively cheap. Though, the most feasible solution for increasing the spatial resolution is to reduce the pixel size by utilizing sensor manufacturing techniques. But the decrease in pixel size, also reduces the amount of light available. It results in shot noise that decreases the quality of image. Thus for reducing the pixel size, without suffering the effects of shot noise, there exists a limitation in pixel size reduction, and the optimal size of pixel is 40 um^2 for a 0.35 um CMOS process. Presently, the current sensor technology has almost touched this level. Another approach to increase the spatial resolution is to reduce the chip size or increase the focal length. But, the increase in focal length will lead to an increase in the size and weight of cameras and reducing the chip size leads to an increase in capacitance. Therefore to overcome the limitations of hardware technology, a new software based approach is introduced towards increasing the spatial resolution.

One of the most promising approach is to utilize signal processing techniques for obtaining an HR image from LR image, and this resolution enhancement approach is called super-resolution. The major advantage of this being that it costs less and the existing LR imaging systems can be utilized.

Even for the case of CCTVs, it is very interpretable that the cost of cameras increase with higher quality, additionally higher-quality footage also demands more storage space, leading to additional costs. Furthermore, there exist several images that are acquired from low quality cameras but they contain necessary information. Hence, it is important to improve the quality of these images.

Though, we can increase the size of an image by using basic editing software programs, like Microsoft Paint. But, these software programs often employ simple algorithms like bicubic and bilinear interpolation that are fast, but not much capable of producing high-fidelity images. In fact, the generated output images tend to contain much pixelations, and it is not easier to discern details. Since, increasing the resolution means increasing the number of pixels, that also means that we need to infer the missing information. This is the main reason due to which the simple interpolation techniques like bicubic, nearest and bilinear interpolation are not able to yield satisfactory results as they do not leverage any knowledge by looking at other similar samples to infer the missing data and fail to generate high-quality images. Thus the overall motivation for image SR can be summed as:

- Down-sampled images make the data more manageable by reducing the dimension of data and thus resulting in fast processing of images.
- The possible hardware based solutions (designing better quality cameras) for increasing the resolution are generally costly compared to the algorithmic based approaches.
- There are plenty of smartphone manufacturers deploying super-resolution in their products (Samsung, GooglePixel, Apple) to enhance their zooming capabilities.
- Most digital videos are generated by complex pipeline mapping the RAW sensor data to low-resolution frames resulting in loss of high-frequency details.
- Degradations arising due to poor imaging equipments, erosion in the air and time can degrade the quality of image.

1.3 Applications of Image Super-Resolution

Image SR is an important class of image processing techniques in computer vision and image processing. Besides improving image perceptual quality, it enjoys a wide range of real-world applications, like medical imaging, surveillance and security, amongst others. Below, we have shortly described the usability of the SR in real-world applications:

1.3.1 Multimedia Industry and Video Enhancement

In today's scenarios, movies, animations, and visual effects all require HD data, so SR can be a useful technique for video enhancement as shown in Fig. 1.3. Additionally, the resolution of TV displays has increased over the past few years. Zooming has recently

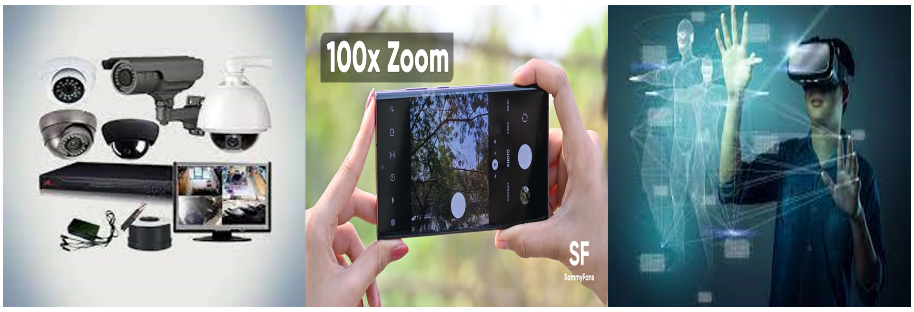


Figure 1.3: Zooming and security based applications of super-resolution.

been the most important feature of smartphone cameras with the leading manufacturers advertising their devices to achieve high level of picture quality at sometimes high zoom level. Thus, SR is now used to provide high quality pictures at high zoom levels in Huawei's flagship models and several emerging devices of leading mobile phone makers. In contrast to the noise filtering and edge enhancement techniques, SR provides real increase of effective resolution thus making visible the indistinguishable details. It captures more details in zoomed area, and replaces the optical zoom lens for a fraction of cost of such lens, without adding much weight and size to the device.

Nowadays demand for high-resolution security cameras is increasing day by day. The higher the resolution of the security cameras, better is the image quality and it will generate vivid details. For instance, 4K (3840×2160) security cameras generate more sharp and clean images than 1440p (2560×1440), 1080p (1920×1080), and 720p (1280×720) security cameras. But to ensure the long-term stable operation of recording devices, as well as the appropriate frame rate for dynamic scenes, these surveillance products tend to sacrifice resolution to some extent. Thus, SR techniques could be deployed to adapt images for displaying on devices with different resolutions and for generating visually pleasing outputs from surveillance cameras.

Additionally image down-scaling and compression techniques are widely employed for meeting the limits of hardware storage and network capacity, that sometimes sacrifice the visual effects as well as causes trouble in visual detection and recognition. Thus SR techniques are widely deployed to reduce the cost as media could be sent at low-resolution and up-scaled on the fly.

Latest developments in machine learning (ML) when combined with increasingly powerful Internet of Things (IoT) devices through efficient processors, are resulting in near real-time object detection and classification for augmented reality (AR) and virtual reality (VR) applications. Thus opening the opportunities for exploring new object detection and classification technologies leveraging super-resolution (SR), that have the potential to be integrated into small, mobile and low-power PSNR devices.

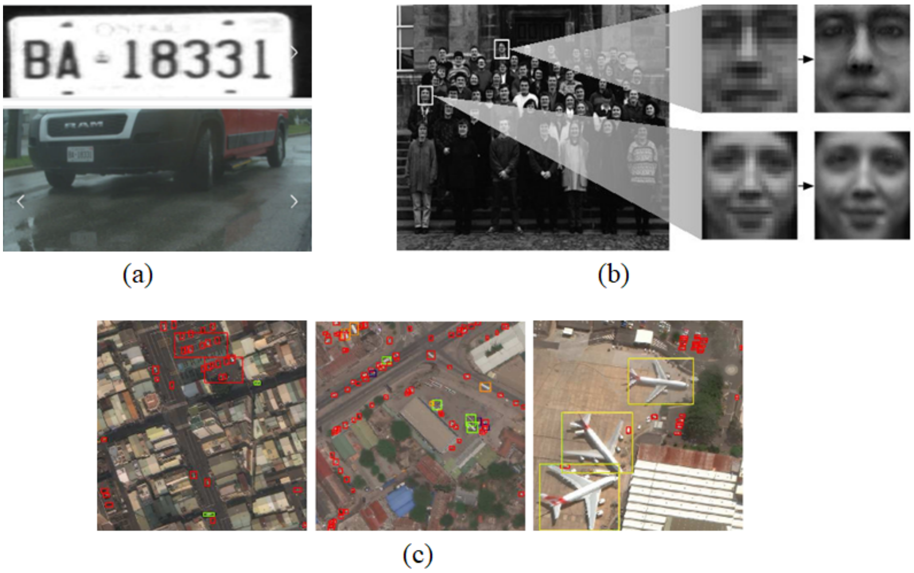


Figure 1.4: (a) Low-Resolution vehicle plate recognition [15], (b) Face recognition [16], and (c) Object detection in satellite images [17].

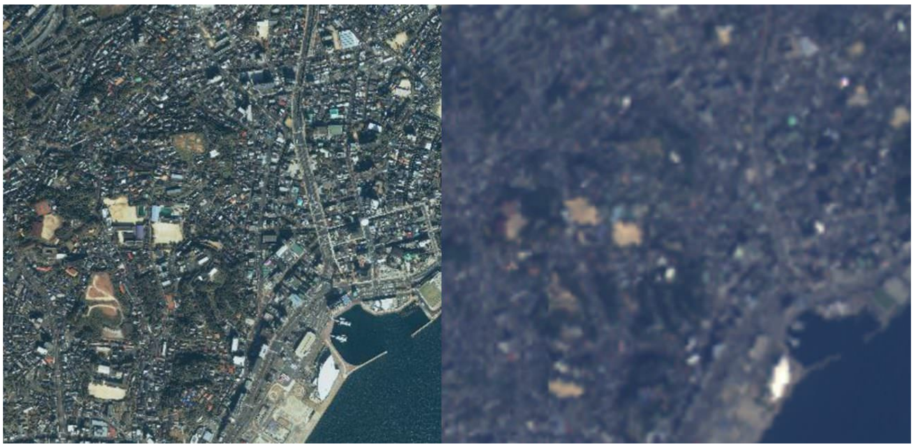


Figure 1.5: Application of super-resolution in remote sensing.

1.3.2 Performance Improvement in High-Level Vision Problems

Super-resolution is a classical problem of reconstructing low-resolution images, that is naturally lost after downsampling the HR image. It is thus widely employed in applications like face recognition, object detection, vehicular plate recognition as shown in Figure 1.4. As the complexity of traffic management is becoming more and more challenging year by year, much research is going on for improving the efficiency and accuracy of vehicle plate recognition. They are highly useful for traffic monitoring and control systems like intelligent parking management, finding stolen vehicles and traffic law enforcement. However in surveillance systems, low-resolution images or videos are widely used. In low-resolution systems, the car plate text is very negligible on account of distance, illumination, and distortion. Thus super-resolution techniques come handy for improving the car plate image quality by processing a single LR image or multiple LR images into a

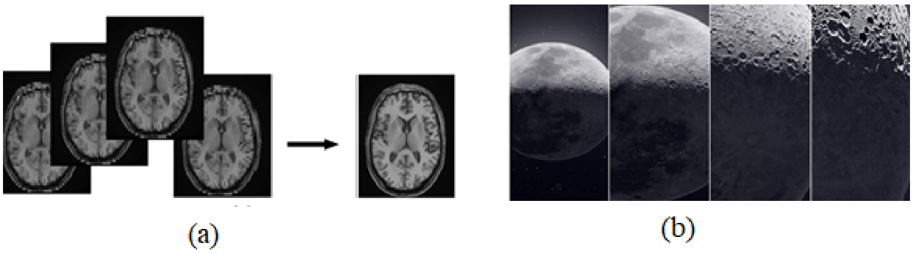


Figure 1.6: Applications in (a) Medical [18] and, (b) Astronomical images [19].

single HR image.

In remote sensing field as shown in Figure 1.5, unlike ordinary camera imaging, images acquired from space-borne equipments are greatly affected to different degrees on account of several factors: (1) ground sampling area, (2) atmospheric attenuation. Thus, the quality of existing LR remote sensing images is too low for meeting the learning requirements of models. For solving this, SR strategy can be used to reconstruct a low-resolution image.

1.3.3 Astronomical and Medical Image Processing

High-resolution telescopic imaging is quite a necessity in astronomy, especially for binary stars, and gravitational lenses. Space telescopes or ground-based telescopes are capable of reaching diffraction-limited spatial resolution. For increasing the spatial resolution and to obtain more discernible details of the celestial bodies, the astronomers are trying to build much larger space and ground-based telescopes. However, the cost of these kind of telescopes will also become more expensive. In contrast, SR technologies can break the diffraction limit of the imaging system and enhance its spatial resolution with compact set-up and low cost, thus making SR telescopic imaging more attractive and meaningful.

In the field of medical imaging, for each imaging modality, specific physical laws are in control, defining the meaning of noise and the sensitivity of the imaging process. But, how to extract 3D models of the human structure with high-resolution images while reducing the level of radiation is still an open challenge. Until now, the designers of medical device were trying to make a trade-off between the size of device and resolution. Additionally, the dimensions of camera modules and their integrated image sensors are often limited by the outer diameter of the endoscope. But, for obtaining a sufficiently bright image while limiting the heat dissipation of the LEDs, the sensors must have a relatively large pixel size. Thus SR algorithm enhances the sensor's resolution and image quality that enables the doctors and nurses to view the captured images on high-definition monitors and tablets.

Thus, the overall applications of super-resolution can be summarized as:

- *Smartphones* for enabling high-quality lossless zoom.

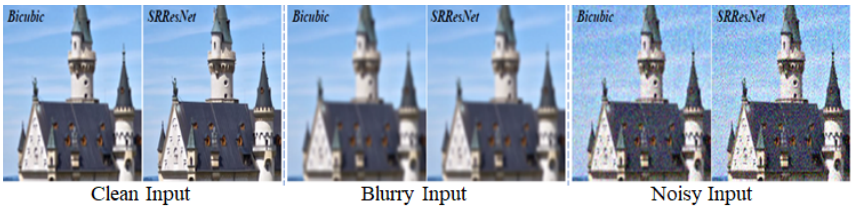


Figure 1.7: Failure cases of a non-blind SR method, that is unable to sharpen the texture for a blurry input and keeps the noise for noisy input. *Here, SRResNet is a popular non-blind SR method* [20].

- *Medical Cameras* for enabling doctors to see with the smallest endoscope.
- *Machine Vision Cameras* for achieving robust object detection during weather conditions.
- *VR/AR Head Mounted Displays* for enhancing the picture quality of see-through cameras.
- *Laptop Cameras* for achieving picture quality by keeping thin structure.

1.4 Research Challenges

There are several challenges that are yet to be resolved for developing a robust automated system. Because, these systems demand the SISR data as input for further processing like video surveillance, traffic monitoring, *etc.* The challenges for SR are discussed below:

1.4.1 Performance Degradation for Real-World Images:

Image SR is highly limited in real-world scenarios like suffering from unknown degradations. Real-world images often tend to suffer from degradations like blurring, compression artifacts, and additive noise. Therefore, the models trained on datasets that are synthesized manually perform poorly in several real-world scenes as shown in Figure 1.7. Further, such methods assume a known fixed degradation process, and thus tend to fail when approached with different and more complex degradations than those for which they were specifically trained upon. Moreover, the type of degradations afflicting an image are generally unknown.

1.4.2 Ill-posed Problem

Generally a super-resolution model is trained by using pairs of high and low-resolution images. Infinitely several high-resolution images, as shown in Figure 1.8 can be down-sampled to the same low-resolution image. It makes SR problem an ill-posed one, that cannot be inverted with a deterministic sampling. This one-to-many stochastic formulation has not been explored in literature and generally outputting multiple

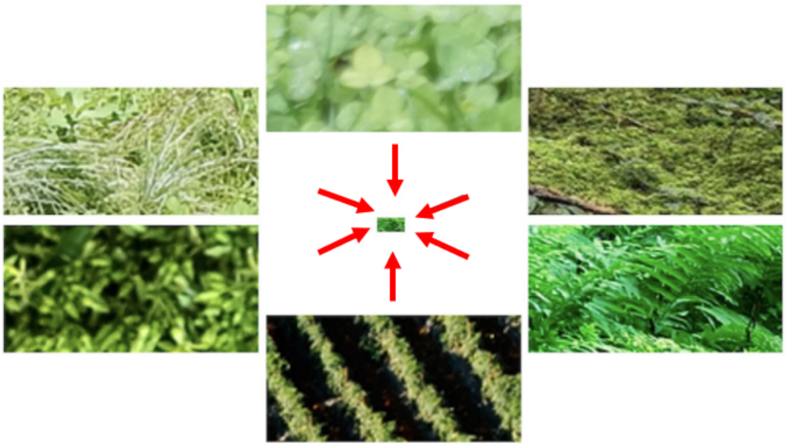


Figure 1.8: Several high-resolution images could be downsampled to a single low-resolution image, making SR an ill-posed problem.

predictions for a single image leads to indeterministic mapping, especially for real-world images.

1.4.3 Aliasing Artifacts in Existing Upsampling Methods

Upsampling plays the most indispensable role while reconstructing a low-resolution image. Despite rapid advancements in learning based super-resolution techniques, there is dearth of research being done on the upsampling techniques. One of the most popular learnable upsampling method, deconvolution can easily have uneven overlap, especially when the kernel size is not divisible by stride, this uneven overlap is more prominent for two dimensional cases, where the overlap tend to multiply together thus generating checkerboard artifacts as shown in Figure 1.9. Sub-pixel convolution (another popular learnable upsampling layer) is a specific implementation of deconvolution layer, being interpreted as a standard convolution in low-resolution space followed by periodic shuffling operations. Though, sub-pixel convolution is constrained to not allow deconvolution overlap but it generally suffers from artifacts owing to their random initializations. Specifically, the involved shifting of feature channels into the spatial domain generally results in the introduction of alignment artifacts.

1.4.4 Incapability of Single-Image Approaches in Smartphones

As shown in Figure 1.10, in some of the scenarios single image approaches fail to properly retrieve the high-frequency details when compared to multi-frame approaches. However, lesser research has been done in the field of multi-frame super-resolution as compared to single frame approaches. Additionally, with the soaring popularity of smart-phones in day-to-day life, the demand for capturing high-quality images is rapidly increasing. However, the camera in smartphone has several limitations due to the constraints placed on it in order to be integrated into smartphone's thin profile. The most prominent hardware

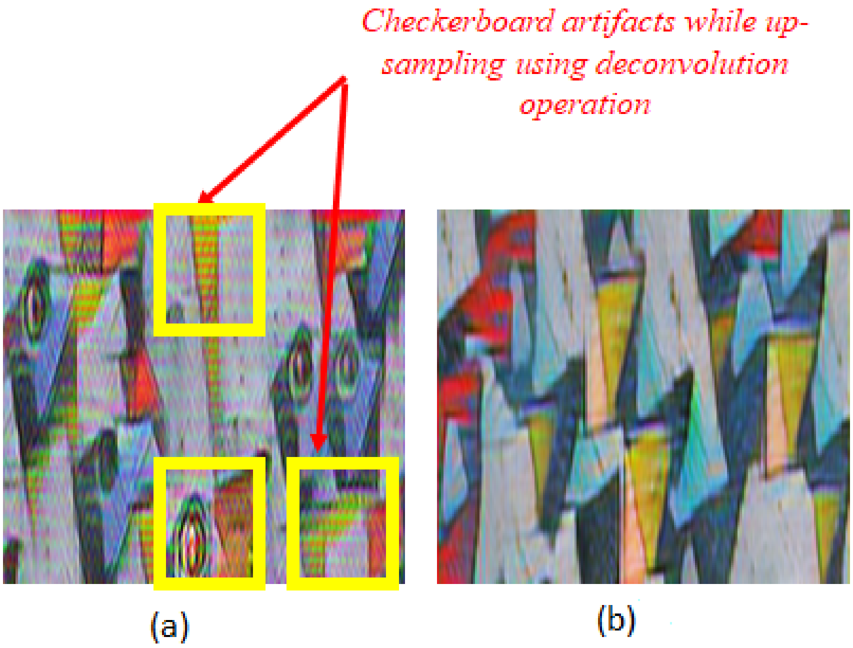

Checkerboard artifacts while up-sampling using deconvolution operation
(a)
(b)

Figure 1.9: (a) Aliasing artifacts introduced by popular Conv2d-Transpose upsampling layer, (b) Desired high-resolution image.

limitations are the small camera sensor size and the associated lens optics that reduce their spatial resolution and dynamic range [74], impeding them in reconstructing DSLR-alike images. To deal with these inherent physical limitations of mobile photography, one emerging solution is to leverage multi-frame (burst) processing instead of single-frame processing.

1.4.5 Alignment Issues while Fusing Multiple Frames

As shown in Figure 1.11, the major issue for multi-frame super-resolution is the artifacts arising on account of inaccurate alignment of the multiple frames. Generally, any burst processing approach is limited by the accuracy of alignment process due to the camera and scene motion of dynamically moving objects. Therefore, it is crucial to design a module for facilitating accurate alignment, as the subsequent fusion and reconstruction modules must be robust to misalignment for generating an artifact-free image.

1.5 Problem Statement

From the above observations, we have identified the following problems for image super-resolution:

1. The performance of the existing state-of-the-art approaches for super-resolution is highly dependent upon prior edge information.

2. Lack of computationally efficient architectures for image super-resolution.

3. More end-to-end novel solutions for the burst processing /multi-frame processing approach need to be explored.

4. The generalization capability of the existing super-resolution approaches is limited in real-world scenarios.

5. The existing up-sampling approaches generate artifacts (jaggy, and checkerboard) in the reconstructed image.

1.6 Aims and Objectives

From the identified problems in existing image super-resolution methods, we define the aim and objective of our work as:
Aim: *To propose novel solutions for resolving different problems in Image Super-Resolution.*
Objectives:

1. To design a novel SR approach for improving the desired high-frequency details and to preserve the low-frequency information without using prior information.

2. To propose a novel lightweight approach that is capable of generating spatially accurate and contextually enriched features.

3. To propose a novel burst/multi-frame processing super-resolution approach for targeting alignment and fusion problems.

4. To propose a novel blind SR approach that is generalized for different real-world degradations.

1.7 Main Contributions

This study is focused on analysing the different modalities and recent trends of image super-resolution task. The major contributions of this work are listed below:

- A novel approach is proposed to improve the desired high-frequency details and preserve the low-frequency information for SR task without any explicit prior information.

- A novel lightweight approach is proposed that is capable of generating spatially accurate and contextually enriched features by maintaining a proper trade-off between accuracy and speed.

- A novel multi-frame super-resolution based approach is proposed that targets at solving multiple degradations-low resolution owing to small camera sensors, mosaic patterns on camera filter array, and sub-pixel shifts on account of camera motion.

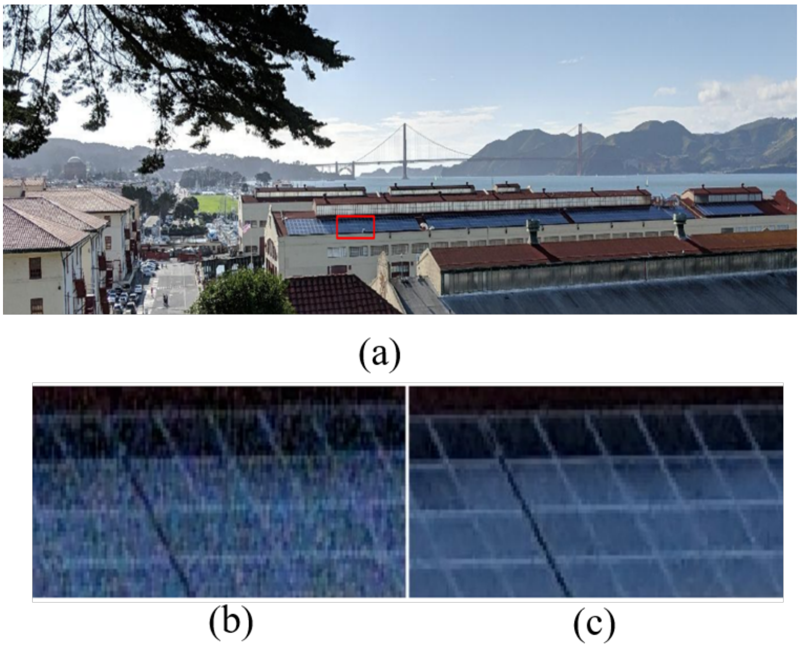


Figure 1.10: Single frame approach *vs.* multi-frame approach. (a) Original image, (b) Single-frame output, and (c) Multi-frame output

- An efficient transformer-based network, based upon the kernel-oriented adjustment of features, KOADNet is proposed that jointly learns the kernel degradation and content information for adapting to the blur characteristics in real-world images.

1.8 Thesis Structure

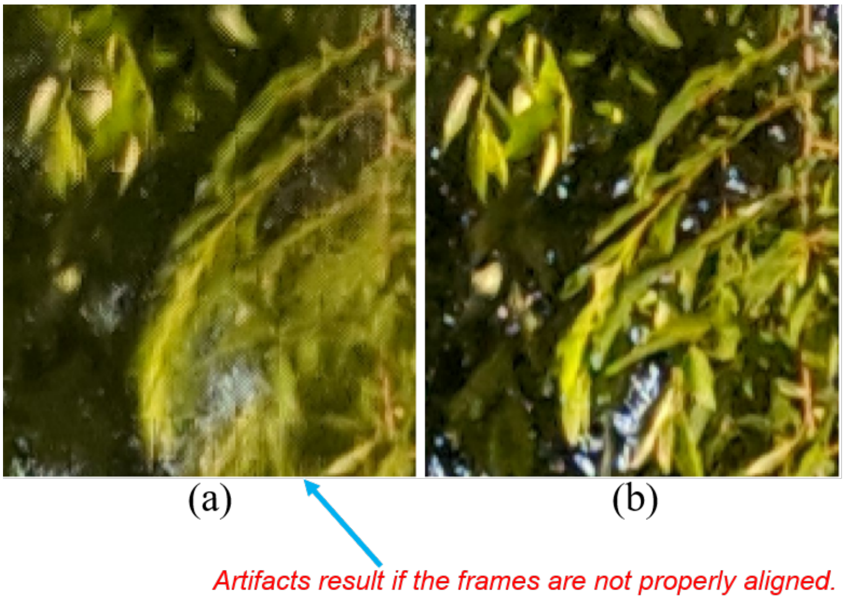
- **Chapter 1:** This chapter describes the motivation behind the present work and presents the preface of whole thesis work.

- **Chapter 2:** This chapter describes the comprehensive study on different hand-crafted and learning based approaches for image super-resolution, burst super-resolution and blind super-resolution.

- **Chapter 3:** We propose a learning-based frequency extraction approach for super-resolution without prior information.

- **Chapter 4:** In this chapter, to resolve the problems of computationally expensive architectures of SR, several lightweight approaches for SR and other restoration applications are discussed.

- **Chapter 5:** To effectively capture more information from multi-frames, a novel burst super-resolution approach is proposed that is also extensible to other applications.



Artifacts result if the frames are not properly aligned.

Figure 1.11: (a) Artifacts resulting while fusing multiple frames if not properly aligned, (b) Output for proper alignment.

- **Chapter 6:** In this chapter, we have proposed the generalization of the existing SR approaches on blind/real-world degradations.
- **Chapter 7:** This chapter lists the conclusion of thesis work and discusses the possible future scope which could further improve the usability of super-resolution for different high-level computer vision applications.

Chapter 2

Literature Survey

In this chapter, existing approaches and benchmark datasets used for experimental analysis in the field of super-resolution and evaluation measures used for the proposed networks analysis are discussed.

2.1 Existing Super-Resolution Approaches

The existing state-of-the-art super-resolution approaches are divided into four parts.

1. Prior Hand-crafted Based Single Image Super-Resolution Approaches
2. Learning-Based Single Image Super-Resolution Approaches
3. Learning-Based Multi-Frame Super-Resolution Approaches
4. Learning-Based Blind Image Super-Resolution Approaches

2.1.1 Prior Hand-crafted Based SISR Approaches

Generally SISR approaches aim at generating high-quality HR images from a single LR input after exploiting certain image priors. In accordance with the image priors, hand-crafted based SISR algorithms can be categorized into different types of approaches.

Prediction Models: SISR algorithms under this category generate HR images from LR inputs via a predefined mathematical formula without any training. For instance, Interpolation-based methods (bilinear, bicubic and Lanczos) reconstruct HR pixel intensities by weighted average of neighborhood LR pixel values. As interpolated intensities are locally quite similar to the neighboring pixels, these algorithms generate smooth regions but output insufficient large gradients along edges and for high-frequency regions. Irani *et al.* [89] generate a low-resolution image via predefined downsampling model and it compensates the difference map in LR back to the HR image.

Edge-based Methods: Edges are important primitive structures that play a vital role in visual perception. Many SISR algorithms have been proposed for learning priors from edge features to reconstruct HR images. Many edge features have been proposed like the width and depth of an edge [90] or by employing the parameter of a gradient profile [91]. As the priors are learned from edges, the reconstructed HR images have sharp edges with optimum brightness and minute artifacts. But, edge priors are not much effective to model the textural information.

Image Statistical Methods: Several image properties may be exploited as priors for predicting HR images from LR images. Shan *et al.* [92] exploited the heavy-tailed gradient distribution [93] for SISR. Additionally, the sparsity property of large gradients in generic images is further exploited in [94] for reducing the computational load and for regularizing the LR input images. To generate HR images, total variation has been employed as a regularization term to generate HR images [95].

Patch Based Methods: From a given set of paired LR and HR training images, patches can be cropped from the training images for learning the mapping functions. The exemplar patches may be generated from the external datasets [96, 97], input image [98], or from combined sources [99]. Several learning based methods of the mapping functions have been proposed like weighted average [100, 101], support vector regression [102], Gaussian process regression [103], kernel regression [104], and sparse dictionary representation [105, 106]. Additionally many methods for blending the overlapped pixels have been proposed including weighted averaging [107], conditional random fields [108], and markov random fields [109].

2.1.2 Learning-Based Single Image Super-Resolution Approaches

The pioneer work in learning based SISR was introduced by Dong *et al.* [110], who proposed SRCNN. After outperforming most of the conventional example-based methods, it paved the way for various SoTA CNN-based SISR techniques. However, on account of its shallow network (three convolutional layers), this method was limited in its learning capabilities. To address this issue, Kim *et al.* through VDSR and DRCN [111] increased the depth of the network (20 layers) and achieved noticeable improvement in the performance of SR. After the remarkable success of residual blocks in ResNet [112], deep networks like EDSR [26] and DRRN [113] were proposed using local residual learning for SISR. To overcome the inefficiency and shortcomings of the residual learning, multi-scale architectures like [29], [114],[115], [64] were proposed for improving the feature representation of residual blocks. Furthermore, for building strong relationships for information flows in each convolution layer, Zhang *et al.* [45] proposed RDN model differing from other CNN-based models, *i.e.*, it did not employ full utilization of hierarchical features from LR images. Lately, there has been an increasing trend of building lightweight and efficient models in SISR for reducing the computational cost. Mehri *et al.* [116] proposed a novel lightweight network for avoiding abundant low-level information via its efficient adaptive residual block. Other recent notable SISR network architectures employ progressive reconstruction [117, 118], generative adversarial networks [119], [120], and recursive learning [121, 122] for improving the efficacy of super-resolution. Our literature analysis reveals that above mentioned deep learning-based SR models do not show much concern regarding image frequency and network complexity, resulting in under-restoration of complex textures and over-restoration of simple textures. Henceforth, to alleviate this limitation, our work focuses on developing an architecture associating frequency information with model of appropriate computational complexity.

After achieving massive success in classification, machine translation tasks, the self-attention methods [123], have been tremendously explored for low-vision tasks. The attention concept in the field of SR was brought up by Zhang *et al.* [44], who investigated the importance of high-frequency channel-wise features for HR reconstruction. Following CBAM [124], Hu *et al.* further proposed [125], employing both channel-wise and spatial features into the residual blocks, for modulating the features globally and locally. Wang *et al.* proposed a new class of neural networks [126] for capturing long-range dependencies via non-local operations. Following which several works like, SAN [52], CSNLN [39], DRLN [41], RNAN [81] were proposed demonstrating huge benefits from non-local operations in the field of SR. Niu *et al.* proposed a new layer attention module in HAN [42] for considering the correlation between multi-scale layers. Liu *et al.* [43] proposed an enhanced spatial attention block for obtaining a more sophisticated attention map. Recently, Du *et al.* [127] proposed a deep expectation-maximization attention cross residual network for tackling the image SR reconstruction. Recently, transformer-based models [128, 129] have been incorporated in the field of SR. SwinIR [129] applied the concept of shifted window mechanism (spatially varying convolution) for modelling long-range dependency problem. Despite achieving significant progress, their major constraint resides in memory inefficient and highly computationally complex operations.

The traditional works in the field of SISR used interpolation for resizing the feature maps from LR to HR. Few example operators include nearest-neighbour, bicubic interpolation [130], edge-based methods [131] for accomplishing the task of SR. Despite being fast, these methods suffered from poor accuracy and were incapable to capture semantic information, since they focus on sub-pixel neighbourhood. The concept of learnable up-sampling in the field of SR was introduced by [132]. Following which, several variants in the up-sampling layers were introduced in the form of deconvolution layer [132], and pixel-shuffle layer [133]. Deconvolution is usually associated to perform the inverse operation of convolution and pixel shuffling was based on the concept that depth of the feature-map channels can be reshaped spatially into height and width of the feature maps. These layers, along with the suitable architectures have been constantly refreshing the results for SR. But as proposed in [134], these up-sampling layers, on account of uneven overlapping, tend to introduce checker-board effects, resulting in low-quality image. Li *et al.* [135] utilised the concept of separation of low and high-frequency components, and proposed residual deconvolution for up-sampling the residual information. Further in [136], Xiong *et al.* introduced a variant in the up-sampling technique using the concept of spatial-shuffle and channel-shuffle. Hu *et al.* [137] discussed a novel way of up-sampling for non-integer factors by predicting the weights of convolution kernel. Recently, Dai *et al.* [138] introduced the concept of learning affinity while upsampling (Affinity-Aware Upsampling), for exploiting pairwise interactions in deep networks. Following the trend of learning-based paradigm, we too intend to propose a learnable upsampling block. We show that, when compared with the other popular upsampling techniques, our proposed module achieves better performance, by maintaining a light-weight learning paradigm.

2.1.3 Learning-Based Burst SR Approaches

SISR usually exploit strong priors or require training data. However, they are often limited in the extent they can reconstruct from aliasing. In contrast to SISR approaches, Multi-Frame Super-Resolution (MFSR) aims at increasing the optical resolution. But it encounters new challenges while estimating the offsets among different images caused by camera movement and moving objects. In terms of sampling theory literature, MFSR approaches date as far back as the '70s [139]. Tsai and Huang [140] were the first to put forward a frequency domain based solution, easy in implementation and computationally cheap for MFSR problem. However, processing in frequency domain resulted in serious visual artifacts. On account of the drawbacks of frequency based approaches, algorithms that enhances image in spatial domain became increasingly popular [141]. Irani and Peleg [142] and Peleg *et al.* [143] proposed an iterative back-projection approach for sequentially estimating the HR image. Successive works [144, 145, 146] improved this approach with maximum posterior (MAP) model. Robustness to varying noise levels and outliers were further addressed in [147, 148]. But all the above mentioned approaches were based upon the assumption that motion between input frames, as well as the image formation model can be well estimated. Subsequent works addressed this issue by joint estimation of the unknown parameters [149, 150]. Farsui *et al.* [151] proposed a hybrid method for performing demosaicking and super-resolution with MAP framework. Wronski *et al.* [152] proposed a MFSR algorithm that merges burst of raw images for supplanting the requirement of demosaicking in camera pipeline.

Recently, a few works resorted to incorporating deep learning for handling the MFSR problem. Deudon *et al.* [73] presented HighRes-net, the first deep learning MFSR approach in satellite imagery, capable of learning all its sub-tasks in an end-to-end fashion. Molini *et al.* [153] designed a novel CNN based algorithm for exploiting both temporal and spatial correlations to combine multiple images. Bhat *et al.* [7] addressed the problem of multi-frame burst SR by proposing an explicit feature alignment and attention-based fusion mechanism. However, explicit use of motion estimation and image warping techniques can pose difficulty handling scenes with fast object motions. Dudhane *et al.* [75] proposed a generalised approach for processing noisy raw bursts through their edge boosting feature alignment and pseudo burst fusion modules.

2.1.4 Learning-Based Blind SR Approaches

Non Blind Super-resolution

In the recent years, various CNN-based SR methods are focused upon the restoration of HR images from LR images that are synthesized with predefined bicubic setting. Since the first CNN-based SR work by Dong *et al.* [154], several sophisticated networks based upon bicubic downsampling have been proposed. Though these methods perform favorably under the ideal bicubic-degraded setting, they tend to generate blurry results in case the degradations in test images deviate from bicubic settings on account of domain gap. Few non-blind SR approaches try to address the problem of multiple degradations by restoring the HR images with the given corresponding kernels. Zhang *et al.* [155] proposed a dimension stretching strategy using additional input in form of kernels and concatenated it with the LR input for degradation aware SR. Based upon the concept of SRMD, Gu *et al.* [86] propose SFTMD model for inputting kernels at different stages of the network via SFT layer [156]. Xu *et al.* [157] integrated the degradation information in the same manner using dynamic upsampling filters in UDVD model for raising the SR performance. But these methods require ground truth information while testing, that is unrealistic for real-world scenarios.

Blind SR

Before the advent of deep learning era, the blind SR methods estimated the HR image and the resulting kernel via edge prior or image patch information [158, 159, 160]. Recently, deep neural network based methods have become the mainstream of blind-image super-resolution research. In the blind SR setting, HR image is recovered from the LR image that is degraded with unknown kernel [160, 161, 162]. Many DNN based blind image SR methods directly perform super-resolution without explicitly estimating the kernel, that include unapired SR [163, 164, 165, 166], and zero-shot SR [167, 168]. However, these methods do not take into consideration the entire process of image degradation and largely rely upon the training dataset while learning the model. Thus, their performance deteriorates while encountering unseen degradation parameters during inference. Hence, the recent approaches solve this problem via a two stage framework: kernel estimation and kernel oriented HR image restoration. In the former, KernelGAN [169] estimate the degradation kernel by applying generative adversarial network (GAN) on a single image, and the estimated kernel is then applied to a non-blind SR approach (ZSSR) to obtain the SR result. Based upon KernelGAN [170], Liang *et al.* improved the performance of kernel estimation by incorporating a flow-based prior. However, KernelGAN and its variant are less suitable for low-resolution images as the GAN optimization brings unstable kernel estimation. Tao *et al.* [171] proposed a spectrum-to-kernel network and prove the

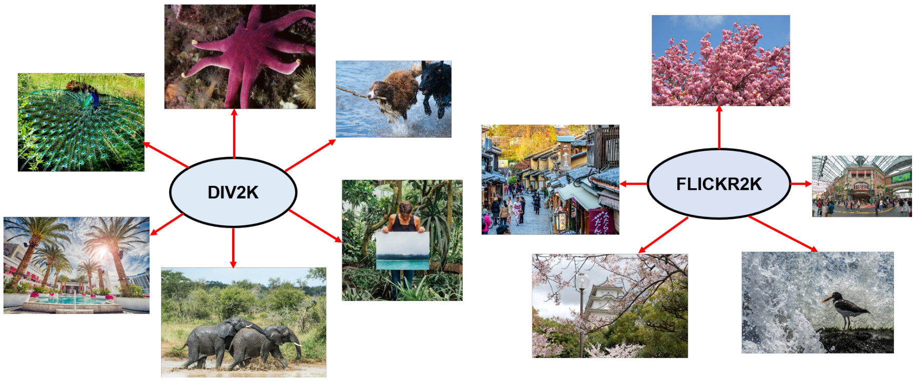


Figure 2.1: Sample images from DIV2K [5] and Flickr2K [6] database.

conducivity of kernel estimation in the frequency domain rather than the spatial domain. For the latter category, Gu *et al.* [172] applied spatial feature transform (SFT) and iterative estimation of kernel (IKC) strategy to accurately estimate kernel and refine SR. Luo *et al.* [173] develop an end-to-end deep alternating network (DAN) through reduced kernel estimation and iterative restoration of HR image. One common problem of IKC and DAN are they are time consuming and they predict the features of kernels rather than the kernel itself. Liang *et al.* [174] propsed a mutually affine transformation network for estimating the kernel in blind image super-resolution by limiting the receptive field to localize the degradation. Zheng *et al.* [175] proposed an unfolded deep kernel estimation method for jointly learning the image and kernel priors.

2.2 Existing Experimental Databases

Here, we have discussed about the existing benchmark datasets used to evaluate the performance of the proposed and the existing state-of-the-art approaches in this work.

2.2.1 DIV2K Database

DIV2K dataset [5] is a high-quality (2K resolution) image dataset for single image and blind image super-resolution task. It consists of 800 training, 100 validation, and 100 testing images.

2.2.2 Flickr2K Database

Flickr2K dataset [6] is a high-quality (2K resolution) image dataset for single image and blind image super-resolution task. It consists of 2650 training images.

2.2.3 Set5 Database

Set5 [1] dataset is a benchmark testing dataset and consists of only five images of a baby, butterfly, bird, head and a woman.

2.2.4 Set14 Database

Set14 [2] consists of some more categories as compared to Set5 [1]; however it has still less number of images, *i.e.* 14.

2.2.5 Urban100 Database

This dataset [4] is further composed of 100 images and the focus of the photographs is on human-based structures, *i.e.* urban scenes.

2.2.6 BSD100 Database

This database [3] is composed of 100 images ranging from natural images to object specific such as food, people, plants and so on.

2.2.7 Manga109

This dataset [49] is composed of 109 test images of a manga volume. These mangas were drawn professionally by Japanese artists.

2.2.8 Synthetic BurstSR Database

This dataset consists of 46,839 RAW bursts for training and 300 for validation. Each burst contains 14 LR RAW images (each of size 48×48 pixels) that are synthetically generated from a single sRGB image. Each sRGB image is first converted to the RAW space using the inverse camera pipeline [7]. Next, the burst is generated with random rotations and translations. Finally, the LR burst is obtained by applying the bilinear downsampling followed by Bayer mosaicking, sampling and random noise addition operations.

2.2.9 Real BurstSR Database

BurstSR dataset consists of 200 RAW bursts, each containing 14 images. To gather these burst sequences, the LR images and the corresponding (ground-truth) HR images are captured with a smartphone camera and a DSLR camera, respectively. From 200 bursts, 5,405 patches are cropped for training and 882 for validation. Each input crop is of size 80×80 pixels.

2.3 Evaluation Measures

Structural Similarity Index:

Structural Similarity (SSIM) index measures an image quality which is based on the hypothesis that the human visual system is highly sensitive to the structural information. Let, $\mathbf{J}^{'}$ and $\mathbf{J}$ are the predicted super-resolved and ground-truth high resolution images,

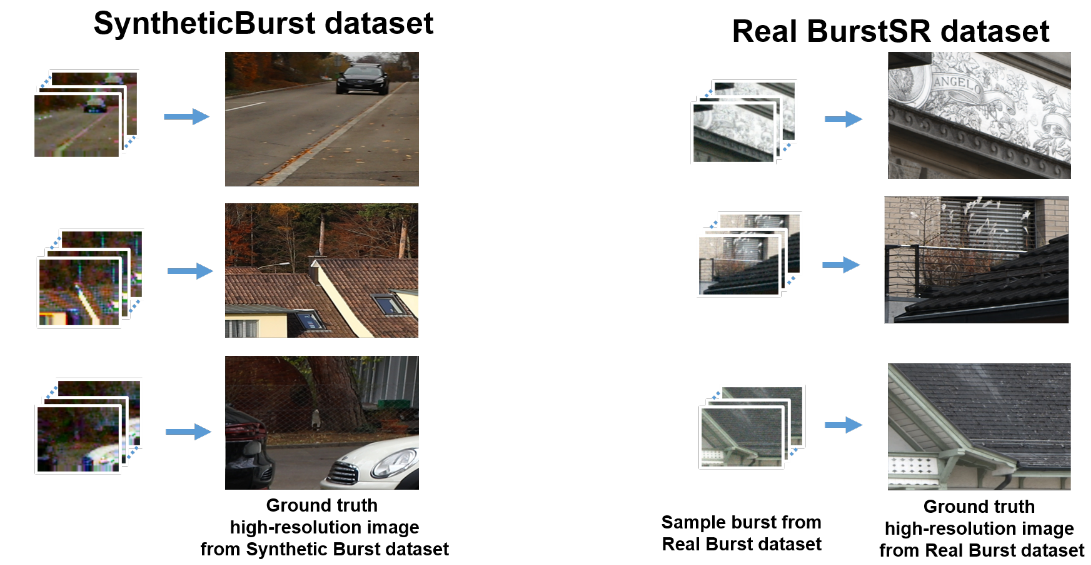


Figure 2.2: Sample images from burst super-resolution dataset [7].

respectively. Then, SSIM between $\mathbf{J}^{'}$ and $\mathbf{J}$ is given as follows:

$$SSIM = \left[lm\left(\mathbf{J}^{'},\mathbf{J}\right)\right]^{\alpha} \cdot \left[c\left(\mathbf{J}^{'},\mathbf{J}\right)\right]^{\beta} \cdot \left[s\left(\mathbf{J}^{'},\mathbf{J}\right)\right]^{\gamma} \tag{2.1}$$

where, luminance (lm), contrast (c), and structural terms (s) are the characteristics of an image having α, β and γ as the exponents respectively and given as,

$$lm\left(\mathbf{J}^{'},\mathbf{J}\right) = \frac{2\mu_{\mathbf{J}'}\mu_{\mathbf{J}} + C_1}{\mu_{\mathbf{J}'}^2 + \mu_{\mathbf{J}}^2 + C_1}$$

$$c\left(\mathbf{J}^{'},\mathbf{J}\right) = \frac{2\sigma_{\mathbf{J}'}\sigma_{\mathbf{J}} + C_2}{\sigma_{\mathbf{J}'}^2 + \sigma_{\mathbf{J}}^2 + C_2}$$

$$s\left(\mathbf{J}^{'},\mathbf{J}\right) = \frac{\sigma_{\mathbf{J}'\mathbf{J}} + C_3}{\sigma_{\mathbf{J}'}\sigma_{\mathbf{J}} + C_3}$$

If $\alpha = \beta = \gamma$ (the default exponents) and $C_3 = \dfrac{C_2}{2}$ then Eq. (2.1) reduces to,

$$SSIM = \frac{\left(2\mu_{\mathbf{J}'}\mu_{\mathbf{J}}\right)\left(2\sigma_{\mathbf{J}'\mathbf{J}} + C_2\right)}{\left(\mu_{\mathbf{J}'}^2 + \mu_{\mathbf{J}}^2 + C_1\right)\left(\sigma_{\mathbf{J}'}^2 + \sigma_{\mathbf{J}}^2 + C_2\right)} \tag{2.2}$$

where, $\mu_{\mathbf{J}'}$, $\mu_{\mathbf{J}}$, $\sigma_{\mathbf{J}'}$, $\sigma_{\mathbf{J}}$ and $\sigma_{\mathbf{J}'\mathbf{J}}$ are the local means, standard deviations, and cross-covariance for images $\mathbf{J}^{'}$, $\mathbf{J}$ respectively. C_1 and C_2 are the small constants 0.01 and 0.03 respectively are added to avoid the undefined values.

If haze-free image recovered by a certain approach is ideally matches to the ground truth haze-free image then **SSIM=1** otherwise its value stands in the range $[\mathbf{0},\ \mathbf{1}]$.

Peak Signal to Noise Ratio:

Peak Signal to Noise Ratio (PSNR) is a traditional evaluation measure for image regression tasks. Mean Square Error (MSE) is calculated by the sum of square of prediction error which is ground truth high resolution image minus predicted super-resolved image and then divide by the number of pixels in an image. It gives an absolute number on how much predicted results deviate from the actual number. Formulation of the MSE is given as follows,

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{J}_i^{'} - \mathbf{J}_i\right)^2 \tag{2.3}$$

where, N represents the number of pixels in an image.

If the predicted super-resolved image ideally matches to the target high-resolution image then **MSE=0** otherwise its value stands in the range **[0, 1]**. Peak Signal to Noise Ratio can be calculated between the super-resolved and target ground-truth image as given below,

$$PSNR = 10 \times log_{10}\left(\frac{R^2}{MSE}\right) \tag{2.4}$$

Here, R denotes the maximum fluctuation in the input image data type. For an 8-bit unsigned integer, R=255. Ideal value for **PSNR=∞** while its value stands in the range $[0, \infty]$

Chapter 3

Implicit Learning based Frequency Extraction Approach for Single Image Super-Resolution

The existing SoTA CNN-based SR method [176, 26, 177] treat different types of information (*e.g.*, low and high-frequency) equally across channels and often lack the discriminative learning ability while dealing with them, that limits their ability to implicitly recover high-frequency details. Additionally the popular SoTA methods use explicit prior (edge prior) information [178, 179] to reconstruct the lost details, thus increasing the overall complexity of the network. Moreover, the popular up-sampling methods [180, 181] introduce checker-board and aliasing artifacts in the reconstructed image, thereby resulting in relatively low performance of the overall network.

Hence, how to effectively utilize these multi-level, channel-wise features within neural networks, without introducing aliasing is crucial for HR reconstruction and need to be further explored. To practically handle the aforementioned bottlenecks, in this chapter we have proposed two different end-to-end solutions for implicitly handling the prior information without any separate training. The two contributions for implicitly extracting the frequency information are:

- Image Super-Resolution with Content-Aware Feature Processing.
- (MLE^{2}A^{2}U)-Net: Image Super-Resolution via Multi-Level Edge Embedding and Aggregated Attentive Upsampler Network.

These solutions are explained in detail in the following sections.

3.1 Image Super-Resolution with Content-Aware Feature Processing

In this chapter, we propose a novel multi-level bi-cubic up-sampler network (MBUp-Net) for high-quality image reconstruction without utilizing any prior information. Every stage of our proposed MBUp-Net consists of a stack of content-aware feature difference (CAFD) and maximum bi-cubic up-sampler (MBU) as its building modules. The proposed CAFD block is designed specifically for extracting the variable (high and low-frequency) content

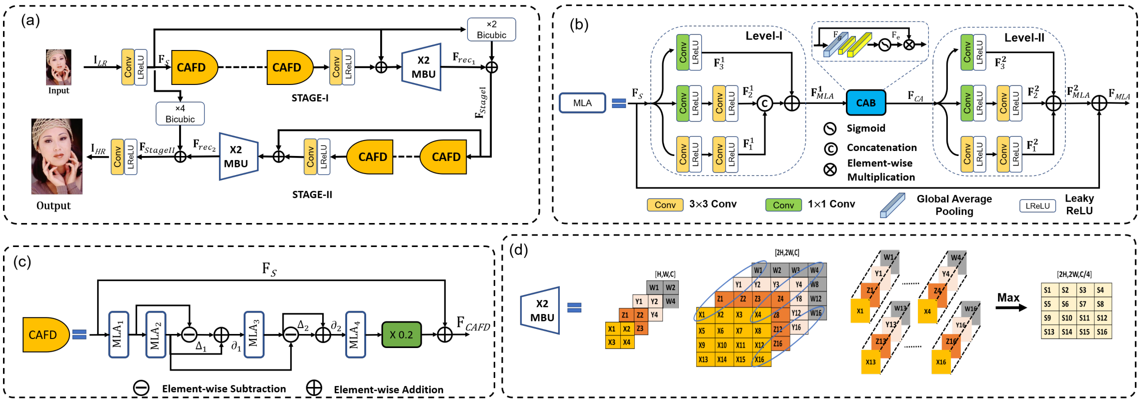


Figure 3.1: (a) Framework of the proposed two-stage MBUp-Net architecture that learns enriched features for ×4 super-resolution, (b) Schematic of MLA block. (c) CAFD block, and (d) Visual illustration of our MBU operation (Here the operation is shown for ×2).

in the image. This specific combination of stacked CAFDs effectively infers better initial LR features. Further, multi-level attention (MLA) blocks are cascaded inside each CAFD block to ensure effective utilization of contextual information from the incoming features. Additionally, unlike direct approaches that up-sample either at the beginning or at the end of the network, an MBU block is utilised progressively at the end of each stage for high-fidelity results. The main contributions of our work are summarised as follows:

- A multi-level bi-cubic up-sampler network (MBUp-Net) is proposed for generating accurate and contextually-enriched outputs.
- A novel content-aware feature difference (CAFD) block is proposed for effectively encoding the multi-scale contextual information by focusing on features of required frequency.
- A novel up-sampling layer based on bi-cubic maximum operation is designed for avoiding the artefacts introduced by other up-sampling techniques.

3.1.1 Proposed Method

The proposed multi-level bi-cubic up-sampler network (MBUp-Net), outlined in Figure 3.1 (a) is a two-stage progressive network. Every stage is a combination of stacked content-aware feature difference (CAFD) blocks and an maximum bi-cubic up-sampler (MBU) block. Let's represent $\mathbf{I}_{LR}$ and $\mathbf{I}_{HR}$ as the low-resolution input and the reconstructed high-resolution output of MBUp-Net, respectively. Given $\mathbf{I}_{LR} \in R^{H\times W\times 3}$, we first apply a convolution layer with LeakyReLU activation for exploiting the shallow features $\mathbf{F}_S \in R^{H\times W\times C}$ as:

$$\mathbf{F}_s = G_{SF}(\mathbf{I}_{LR}), \tag{3.1}$$

where, $G_{SF}(.)$ denotes the shallow feature extraction. Next, $\mathbf{F}_s$ is given as an input to Stage-I and utilized for feature restoration via content aware feature difference (CAFD) blocks. So we can further have:

$$\mathbf{F}^{i}_{CAFD} = G^{i}_{CAFD}(...G^{1}_{CAFD}(\mathbf{F}_{s})), \tag{3.2}$$

where, $G^{i}_{CAFD}(.)$ and $\mathbf{F}^{i}_{CAFD}$ denote the function of the i-th CAFD and its corresponding restoration result, $\forall\ i = 1,\ 2,\ 3,..\ 10$. The architecture of the proposed CAFD block is shown in Figure 3.1 (c), which is composed of four multi-level attention (MLA) blocks. More details about CAFD and MLA blocks are given in the following sub-sections. After extracting hierarchical features from a set of CAFD blocks in low-resolution space, we stack an MBU block in high-resolution space. It improves the reconstruction performance of the overall model and finally outputs, $\mathbf{F}_{rec_1} \in R^{2H\times 2W\times C}$. It is worth mentioning that we adopt global residual learning after the MBU block in each stage to ease the transmission process. This residual addition further ensures the adaptive fusion of features produced by the proposed up-sampling block. Thus, the overall output of Stage-I is formulated as:

$$\mathbf{F}_{StageI} = \mathbf{F}_{rec_1} + W_3(Bic_{\uparrow 2}(\mathbf{F}_S)), \tag{3.3}$$

where, W_3 denotes a convolution layer with filter size 3×3. $Bic_{\uparrow s}$ represents the simple ×s bi-cubically interpolated input followed by a convolution layer. Thereafter, the obtained features from Stage-I are imported into the second stage to learn more enriched feature representations, $\mathbf{F}_{rec_2} \in R^{4H\times 4W\times C}$. Consequently, the overall operation of Stage-II is defined as:

$$\mathbf{F}_{StageII} = \mathbf{F}_{rec_2} + W_3(Bic_{\uparrow 4}(\mathbf{F}_S)) \tag{3.4}$$

At the end of our network, we apply convolution layers with LeakyReLU activation function for transforming the output of Stage-II from feature to image domain. Thus, we obtain the output $\mathbf{I}_{HR} \in R^{4H\times 4W\times 3}$ as:

$$\mathbf{I}_{HR} = D(\mathbf{I}_{LR}), \tag{3.5}$$

where, $D(.)$ refers to the function of whole MBUp-Net. The following sections describe the individual components.

Multi-level Attention Block

One basic property of neurons in the visual cortex is to adaptively change receptive fields according to the stimulus [182]. This adjustment of receptive fields can be incorporated in CNNs via multi-scale feature generation. Thus, for efficiently harnessing multi-scale features, we impose a two-level architectural design as shown in Figure 3.1 (b). Moreover, it is certified that parallelly stacked CNN architecture [183] assists in adaptive learning of the subsequent layers to pick and choose the relevant information. This results in a larger receptive field that is associated to capture more contextual information. Considering the above factors, both the levels in our proposed MLA block have multiple parallel convolution layers for enhancing the overall ability of our proposed network. For Level-I, the feature maps obtained through 1×1 convolution layer are added to the concatenated feature maps from two parallel paths as shown in the left part of Figure 3.1 (b). These concatenated feature maps obtained from large size filters increase the contextual variation in the input and consequently improve the efficiency of next level (Level-II). Further addition with these concatenated features, as shown in Eq. (3.6) helps in preservation of the required features by avoiding the learning of redundant features. The overall function of Level-I in MLA block is defined as:

$$\mathbf{F}^1_{MLA} = [\mathbf{F}^1_1\mathbf{F}^1_2] + \mathbf{F}^1_3, \tag{3.6}$$

where, [.] represents the concatenation operation and $\mathbf{F}^1_1$, $\mathbf{F}^1_2$, and $\mathbf{F}^1_3$ represent the features extracted at Level-1, $\mathbf{F}^1_1 = \ell(W_3(\ell(W_3(\mathbf{F}_S))))$, $\mathbf{F}^1_2 = \ell(W_3(\ell(W_1(\mathbf{F}_S))))$, and $\mathbf{F}^1_3 = \ell(W_1(\mathbf{F}_s))$. Here, ℓ is used to represent LeakyReLU function and W_x represents convolution layer with filter size x.

Channel Attention Block

Multi-scale pattern targets on information from variable receptive fields, whereas attention mechanism focuses on adjustment of the distribution of feature map for efficiently exploring correlation among features. Consequently, to strengthen the effectiveness of our proposed MLA block we incorporate an attention module, channel attention block (CAB) between both the levels. CAB (middle part of Figure 3.1 (b)) generates attention maps for suppressing the less informative features of Level-I and allows only the useful features to propagate to the next level. As illustrated in Figure 3.1 (b), on the incoming features $\mathbf{F}^1_{MLA} \in R^{H\times W\times C}$ from Level-I, CAB first applies global average pooling on individual channels to obtain global feature descriptor $\mathbf{F}_g \in R^{1\times 1\times C}$. To capture the inter-channel dependencies, we pass the descriptor $\mathbf{F}_g$ through two 3×3 convolutions and sigmoid activation, resulting in new attention features $\mathbf{F}_e \in R^{1\times 1\times C}$. Finally, after getting the attention weight of all channels, each obtained attention feature is scaled by the corresponding original feature map ($\mathbf{F}^1_{MLA}$) as shown below:

$$\mathbf{F}_{CA:,i,:,:} = \mathbf{F}_{e_i} \odot \mathbf{F}^1_{MLA:,i,:,:} \forall i = \{0, 1, ...C-1\}, \tag{3.7}$$

where $\mathbf{F}_{CA}$ is used to denote the output of the channel attention block, and $\mathbf{F}_{MLA:,i,:,:}$ denotes the feature map of the ith channel of input F_{MLA}.

Further, the extracted attention-augmented features, $\mathbf{F}_{CA}$ produced by Channel attention (CA) block are passed to the next level, where they are processed again through a parallel multi-scale architecture for inheriting better contextual information among the features. Furthermore, the overall operation of Level-II in the MLA block is given as:

$$\mathbf{F}_{MLA}^{2} = \mathbf{F}_{1}^{2} + \mathbf{F}_{2}^{2} + \mathbf{F}_{3}^{2}, \tag{3.8}$$

where $\mathbf{F}_1^2$, $\mathbf{F}_2^2$, and $\mathbf{F}_3^2$ denote the features extracted at Level-II. Here, $\mathbf{F}_1^2 = \ell(W_3(\ell(W_3(\mathbf{F}_{CA}))))$, $\mathbf{F}_2^2 = \ell(W_3(\ell(W_1(\mathbf{F}_{CA}))))$ and $\mathbf{F}_3^2 = \ell(W_1(\mathbf{F}_{CA}))$. Finally, for each MLA block we adopt residual learning to improve the performance of the network and the overall output is defined as:

$$\mathbf{F}_{MLA} = \mathbf{F}_s + \mathbf{F}_{MLA}^{2} \tag{3.9}$$

This sequential exploration of the features via two-level MLA block aids the proposed network to overcome the under-utilization of local features and give visually pleasing results.

Content-Aware Feature Difference Block

Generally, the LR feature space contains abundant low-frequency and valuable high-frequency content (edges, textures) that contribute differently for recuperating high-fidelity details. It is required to exploit this variable content (low and high-frequency) for facilitating the representation power of the overall network. To efficiently extract this content, it is important to collect contextual information outside the local region [36]. In light of this and to surge the network's sensitivity towards higher contributing features, we propose a computationally effective design, content-aware feature difference (CAFD) block.

Our proposed CAFD block as shown in Figure 3.1 (c) is composed of four multi-level attention (MLA) blocks, connected through skip connections. Fundamentally, our CAFD block is inspired from the concept of high-boost filtering, which focuses on enhancing high-frequency information while preserving the information with low-frequency content. In CAFD, we initially evaluate the absolute difference between features of the second and first MLA block to obtain Δ_1. Subsequently, the obtained difference features, Δ_1 are added to the original features (from second MLA) for attaining coarse high-frequency features, ∂_1.

Thereafter, these obtained coarse features are provided as input to the third MLA block and the above procedure for getting Δ_1 (*via subtraction between features of consecutive blocks*) and ∂_1 (*via addition*) is repeated to boost and consequently generate fine high-frequency features, ∂_2. Furthermore, as shown in Figure 3.1 (c), we add the features of the initial shallow layer ($\mathbf{F}_S$) to the features processed from fourth MLA block, (F_{MLA_4}) for maintaining the diversity of overall content. Thus, the overall function of ith CAFD can be summarized as follows:

$$\mathbf{F}^{i}_{CAFD} = 0.2 * \mathbf{F}^{i}_{MLA_4} + \mathbf{F}_S, \tag{3.10}$$

where, $\mathbf{F}^{i}_{MLA_4}$ represents the extracted features of the fourth MLA block in ith CAFD. Critically as shown in Eq. (3.10), multiplication by 0.2 (residual scaling) is incorporated to avoid the amplitude magnification of the input signals in the proposed CAFD blocks, which may otherwise affect the overall training of the proposed network. To further ameliorate the network performance, the proposed network deploy 20 CAFD blocks (10 in each stage) that accomplishes significant performance gain as discussed in the ablation study.

Maximum Bi-cubic Up-sampler Block

SR requires accurate prediction of pixels present in the low-resolution image and therefore an effective up-sampling layer adept for filling in the missing details needs to be designed. Previous successful attempts for SR are subjected to novel architectures and training strategies, but little work has been done for up-sampling layer. Most commonly used non-learnable up-sampling layers like bi-cubic, nearest-neighbour, bilinear, and learnable up-sampling layers like pixel shuffle [181], deconvolution [180] introduce smoothing, aliasing, and checker-board artefacts [180]. Particularly, strided deconvolution [180] can easily have uneven overlap, especially when the kernel size is not divisible by stride, and this uneven overlap is more prominent for two-dimensional cases where the overlap tends to multiply together, thus generating checkerboard artefacts. Sub-pixel convolution (pixel-shuffle) is a specific implementation of deconvolution layer, often interpreted as standard convolution in low-resolution space followed by periodic shuffling operations. Though sub-pixel convolution is constrained to not allow deconvolution overlap, it generally suffers from checkerboard artefacts owing to their random initialization [184]. Specifically, this shifting of feature channels into the spatial domain generally results in the introduction of alignment artefacts. Unlike the above-mentioned up-sampling techniques, our proposed up-sampler (maximum bi-cubic) avoids shuffling or overlap between the channels to mitigate the artefacts with less computational complexity.

Initially as shown in Figure 3.1 (d), we first perform bi-cubic up-sampling to increase the resolution from $H \times W \times C$ to $2H \times 2W \times C$, where H, W and, C denote the height, width, and the number of channels, respectively. We then consider the maximum pixel among all the channels to preserve the high-frequency information, that is usually lost in low-resolution images. This maximization operation for all the M channels leads to a reduction in the number of channels by a factor of M, thus minimizing the number of computations. Furthermore, the effectiveness of our proposed upsampling layer in generating sharper edges can be verified from Figure 3.2. Meanwhile, embedding residual learning after every MBU module helps in proper exploitation of the relative information between high-resolution and low-resolution multi-scale features for boosting

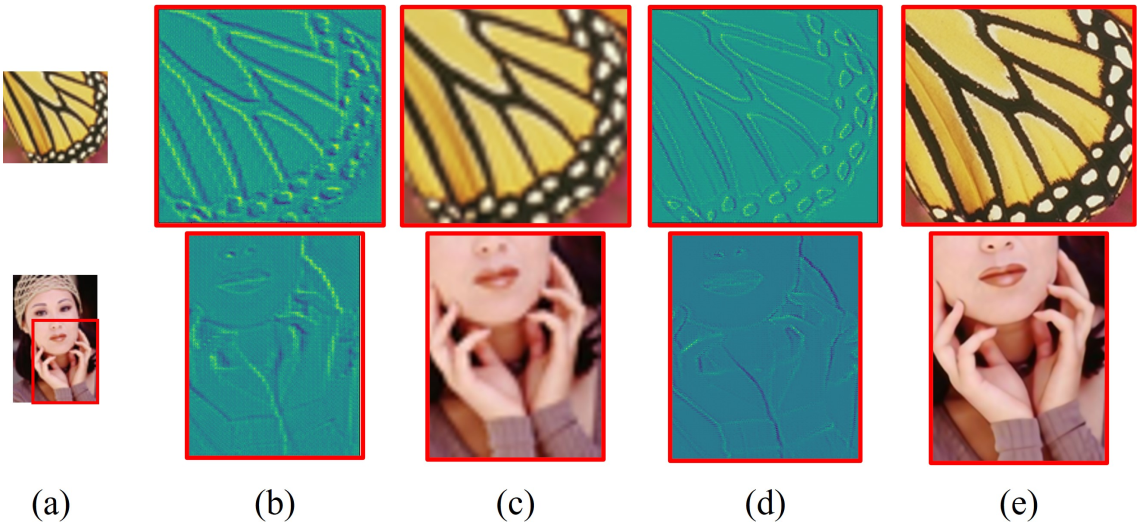


Figure 3.2: (a) Input low-resolution image, (b) cropped intermediate attention feature maps with (only bi-cubic) operation, (c) reconstructed results after bi-cubic interpolation, (d) Cropped Intermediate attention feature maps with proposed (bi-cubic+maximum) operation, and (e) the results after applying proposed maximum bi-cubic up-sampler.

the reconstruction performance.

3.1.2 Experimental Analysis

Datasets

In our experiments, we trained our network using 800 high-quality DIV2K [5] images. For high diversity, we further augmented our training dataset with random horizontal flips and 90-degree rotations. We conducted ablation studies on benchmark SR testing datasets, Set5 [1], Set14 [2], BSD100 [3], Urban100 [4], and Manga109 [49]. These datasets contain a variety of natural scenes, urban scenes and Japanese cartoon images, thus authenticating the overall performance of our proposed architecture.

Table 3.1: Training image settings for DIV2K images.

Size	×**2**	×**3**	×**4**
Sub-image size	640×640	512×512	480×480
crop size	128×128	85×85	64×64

Training Settings

The resolution of training images in DIV2K is nearly 2K and on account of memory limitations directly super-resolving 2K×2K is difficult, hence we first prepare overlapping sub-images of smaller size *i.e.* 640×640, 512×512 and 480×480, respectively for scales ×2, ×3 and ×4. From these prepared sub-images, we further crop corresponding patches of different sizes depending on the scale factors for facilitating stable training as shown in

Table 3.1. LR images of multiple scale factors are obtained by down-sampling the training images with a bi-cubic kernel. The proposed network is implemented in TensorFlow 2.0 and trained for 25,00,000 iterations using ADAM optimizer. Following the settings of [33], the initial learning rate is set at $1{\times}10^{-4}$ and halved every 10^4 iterations. The training hyper-parameters are implemented using NVIDIA DGX station with processor 2.2GHz, Intel Xenon E5-2698, NVIDIA Tesla V100 1×16 GB GPU, using peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) as the evaluation metrics.

For exhibiting the effectiveness of our proposed MBUp-Net, we choose the commonly used L_1 loss as the objective function. Given a training set $\{I_{LR}, I_{HR}\}_{b=1}^{B}$ that comprises of B corrupted low-resolution (LR) inputs and their corresponding high-resolution (HR) labels. The objective function for training MBUp-Net is defined as:

$$L(\theta) = \frac{1}{B}\sum_{b=1}^{B} ||\mathbf{I}_{HR}^{b} - D(\mathbf{I}_{LR}^{b})||_1, \tag{3.11}$$

where, θ refers to the learnable parameters and $D(.)$ refers to the overall function of the proposed MBUp-Net.

Comparison with the State-of-The-Art Methods

We compare the performance of the proposed MBUp-Net with popular CNN-based SR methods[1] that prove the efficiency of our proposed model. For a fair comparison, PSNR has been evaluated on the Y channel of the transformed YCbCr space of the reconstructed SR image.

[1] The source codes and results are downloaded from the respective authors' homepage and we have used directly the settings as recommended by the authors.

Table 3.2: Quantitative Comparison (PSNR/SSIM) of Proposed MBUp-Net on Different Benchmark Datasets for SR.

Method	Scale	Up-sampling	Set5 [38] PSNR/SSIM	Set14 [31] PSNR/SSIM	BSD100 [21] PSNR/SSIM	Urban100 [84] PSNR/SSIM	Manga109 [49] PSNR /SSIM
MSRN [32]	×2	Pixel-shuffle	38.08/0.9607	33.70/0.9186	32.23/0.9002	32.29/0.9303	38.69/0.9772
SRFBN [185]	×2	Pixel-shuffle	38.11/0.9602	33.82/0.9196	32.29/0/9010	32.62/0.9328	39.08/0.9779
RCAN [36]	×2	Pixel-shuffle	38.27/0.9614	34.12/0.9216	32.41/0.9027	**33.34/0.9384**	**39.44/0.9786**
SeaNet [178]	×2	Bi-cubic Interpolation	38.08/0.9609	33.75/0.9190	32.27/0.9008	32.50/0.9328	38.76/0.9774
SRNSSI [179]	×2	Pixel-Shuffle	38.17/0.9618	33.92/0.9205	32.38/0.9028	32.24/0.9303	-
SHSR [186]	×2	Bi-cubic Interpolation	38.22/0.9612	33.90/0.9205	32.34/0.9015	32.78/0.9342	39.15/0.9781
FM-Net [187]	×2	Pixel-Shuffle	38.16/0.9615	33.97/0.9200	32.35/0.9017	32.92/0.9338	-
CCN [188]	×2	Pixel-Shuffle	38.17/0.9616	33.92/0.9199	32.35/0.9018	32.93/0.9339	-
NLSA [177]	×2	Pixel-Shuffle	**38.39/0.9621**	**34.13/0.9219**	**32.48/0.9031**	33.22/0.9378	-
Proposed	×2	**Bi-cubic Maximum**	**38.31/0.9626**	**34.23/0.9219**	**32.52/0.9035**	**33.46/0.9391**	**39.64/0.9792**
MSRN [32]	×3	Pixel-shuffle	34.46/0.9278	30.41/0.8437	29.15/0.8064	28.33/0.8561	33.67/0.9456
SRFBN [185]	×3	Deconvolution	34.70/0.9292	30.51/0.8461	29.24/0.8084	28.73/0.8641	34.18/**0.9481**
RCAN [36]	×3	Pixel-shuffle	34.74/0.9299	30.65/0.8482	29.32/0.8111	**29.09/0.8702**	**34.44/0.9499**
SeaNet [178]	×3	Bi-cubic Interpolation	34.55/0.9282	30.42/0.8444	29.17/0.8071	28.50/0.8594	33.73/0.9463
SRNSSI [179]	×3	Pixel-Shuffle	34.56/0.9290	30.57/0.8467	29.23/0.8097	28.24/0.8557	-
SHSR [186]	×3	Deconvolution	34.65/0.9289	30.54/0.8461	29.24/0.8087	28.71/0.8630	34.10/0.9480
FM-Net [187]	×3	Pixel-Shuffle	34.66/0.9292	30.56/0.8462	29.25/0.8093	28.71/0.8631	-
CCN [188]	×3	Pixel-Shuffle	34.67/0.9290	30.58/0.8462	29.29/0.8094	28.73/0.8632	-
NLSA [177]	×3	Pixel-Shuffle	**34.83/0.9302**	**30.68/0.8488**	**29.32/0.8113**	29.02/0.8689	-
Proposed	×3	**Bi-cubic Maximum**	**34.76/0.9296**	**30.74/0.8500**	**29.46/0.8114**	**29.22/0.8713**	**34.23**/0.9455
MSRN [32]	×4	Pixel Shuffle	32.07/0.8903	28.60/0.7751	27.52/0.7273	26.04/0.7896	30.17/0.9034
NoUCSR [189]	×4	Channel Shuffle + Pixel Shuffle	32.15/0.8936	28.64/0.7824	27.57/0.7356	26.15/0.7871	30.57/0.9087
ADRD [190]	×4	Residual Deconvolution	32.45/0.8999	28.84/0.7823	27.69/0.7477	**27.26**/0.8041	-
SeaNet [178]	×4	Bi-cubic Interpolation	32.33/0.8970	28.72/0.7855	27.65/0.7388	26.32/0.7942	30.74/0.9129
SRNSSI [179]	×4	Pixel-Shuffle	32.06/0.8938	28.68/0.7845	27.58/0.7382	25.80/0.7777	-
SHSR [186]	×4	Bi-cubic Interpolation	32.47/0.8984	28.81/0.7870	27.72/0.7405	26.55/0.7995	31.07/0.9144
FM-Net [187]	×4	Pixel-Shuffle	32.47/0.8972	28.77/0.7856	27.68/0.7415	26.52/0.7988	-
CCN [188]	×4	Pixel-Shuffle	32.50/0.8974	28.77/0.7855	27.69/0.7419	26.54/0.7990	-
NLSA [177]	×4	Pixel-Shuffle	32.59/0.9000	**28.87/0.7891**	**27.78/0.7444**	26.96/**0.8109**	-
Proposed	×4	**Bi-cubic Maximum**	**32.67/0.9005**	**28.94/0.7894**	**27.88/0.7475**	**27.67/0.8113**	**31.42/0.9197**

Objective Evaluation

Table 3.2 demonstrates that our method attains SoTA results on all the benchmark SR datasets. Most of the compared CNN-based models are well-designed networks and have the best results at their time. Among the series of large SR models being proposed, EDSR [26] is the most popular with about 43 Million parameters, whereas our proposed MBUp-Net (9 Million) with just one-fourth of the total parameters of EDSR has achieved better results. Specifically, the proposed method outperforms the recently proposed SHSR [186], FM-Net [187], and CCN [188] by about 1.12, 1.15, and 1.13 dB, respectively on Urban100 dataset for ×4 scaling factor. In comparison to the conventional SR networks of VDSR [191], LapSRN [23], SRNSSI [179], and SeaNet [178], the proposed method exhibits a great performance gain on all the considered datasets.

Subjective Evaluation

We have presented some subjective results in Figure 3.3 and 3.4 for the investigation of the proposed method in terms of visual quality. Since for applications like image super-resolution, visual results are much more valuable than the quantitative comparison, we have compared selected challenging images with about eleven different methods, including Bi-cubic, SRCNN [22], VDSR [191], DRRN [192], Memnet [24], FSRCNN [35], EDSR [26], LapSRN [23], MSRN [32], RDN [30], IDN [25], CARN [28], SRMD [27], GLADSR [193], and CNF [194]. For facilitating better comparison quality, we have enlarged selected regions in the SR image, showing that the images being super-resolved by our method have shown superior results. We have focused mainly on large scaling factor (×4) for comparison of the proposed method. As shown in Figure 3.3 and 3.4, most of the compared methods are unable to recover finer details and introduce artefacts in the reconstructed image (*e.g.* Urban100img_076 in Figure 3.3, Set14-barbara in Figure 3.4). Unlike these methods, our proposed MBUp-Net generate images that are much sharper and visually faithful to the ground-truth. Additionally, as shown in Urban100img_092 and Urban100img_004 of Figure 3.3 and 3.4 respectively, majority of the compared SR methods are inept in capturing crisp edges and produce images with blotchy texture. Whereas, our proposed MBUp-Net preserves the image edges and is more coherent in gathering information from the low-level features.

Difference with prior works

In this subsection, we highlight the major differences between our proposed MBUp-Net and several popular related existing works.

For *MSRN* [32], a multi-scale approach for efficient image reconstruction is proposed. But there exist few differences that need to be addressed. Firstly, their multi-scale model

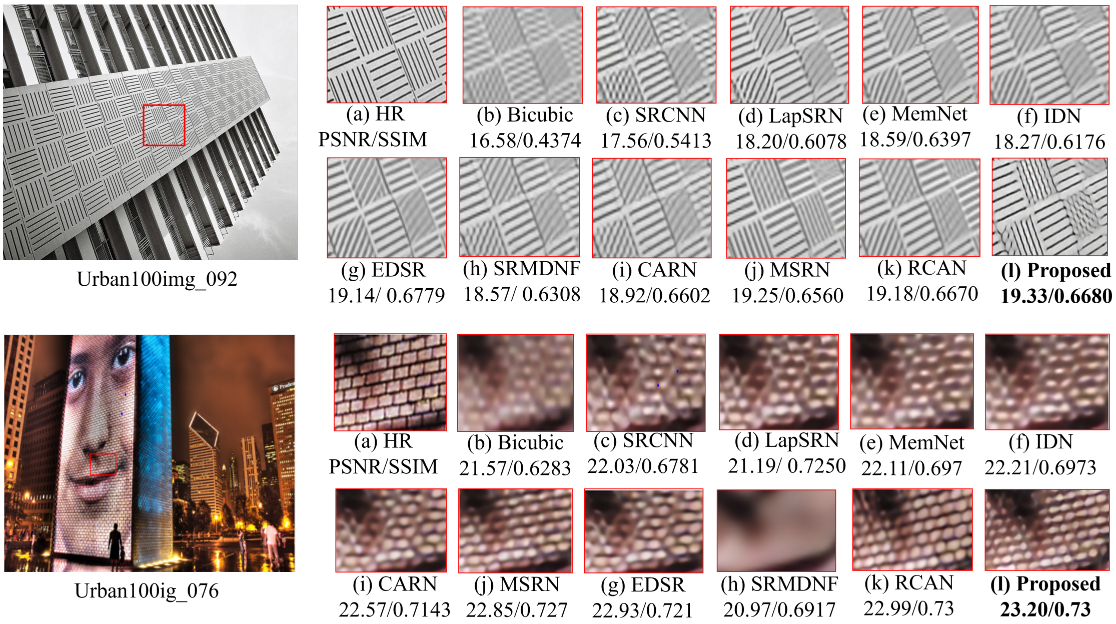


Figure 3.3: Visual comparison on ×4 images for Urban100 dataset [21] where (a) cropped ground-truth HR, (b) Bi-cubic, (c) SRCNN [22], (d) LapSRN [23], (e) MemNet [24], (f) IDN [25], (g) EDSR [26], (h) SRMDNF [27], (i) CARN [28], (j) MSRN [29], (k) RCAN [30], and (l) Proposed Method.

Table 3.3: PSNR/SSIM results achieved on Set5 [38] dataset for two scale factors (×2, and ×4).

Methods	Params	Flops	Time	×2	×4
EDSR [26]	43M	1.2G	0.05s	38.11/0.9602	32.46/0.8968
RDN [30]	22M	0.3G	0.07s	38.24/0.9614	32.47/0.8990
RCAN [36]	15M	0.5G	0.23s	38.27/0.9614	32.63/0.9002
MBUp-Net	**9M**	**0.1G**	**0.02s**	**38.31/0.9626**	**32.67/0.9005**

does not use a progressive architecture that is associated to have stable training for large-scale factors [23]. Secondly, unlike MSRN, our proposed MBUp-Net utilizes an attention mechanism in the residual block for emphasizing important channel features. Thirdly, we incorporated maximum bi-cubic up-sampling unlike their pixel shuffle for resizing the obtained features.

RCAN [36]: Our proposed model adopts the attention module similar to RCAN, but there are subtle differences in our proposed architecture. Firstly, RCAN is a much larger model (more than 400 layers) with 15M parameters as compared to the proposed MBUp-Net with 9M parameters. Secondly, for effective removal of dead features in the attention branch for deeper networks we utilized LeakyReLU, unlike ReLU used in RCAN. Thirdly, unlike RCAN our proposed model is a multi-scale progressive architecture focused on increasing the receptive field for better reconstruction.

ESRGAN [33]: In ESRGAN, a residual in residual block using dense connections has been proposed. However, different from ESRGAN we proposed a CAFD block to highlight the

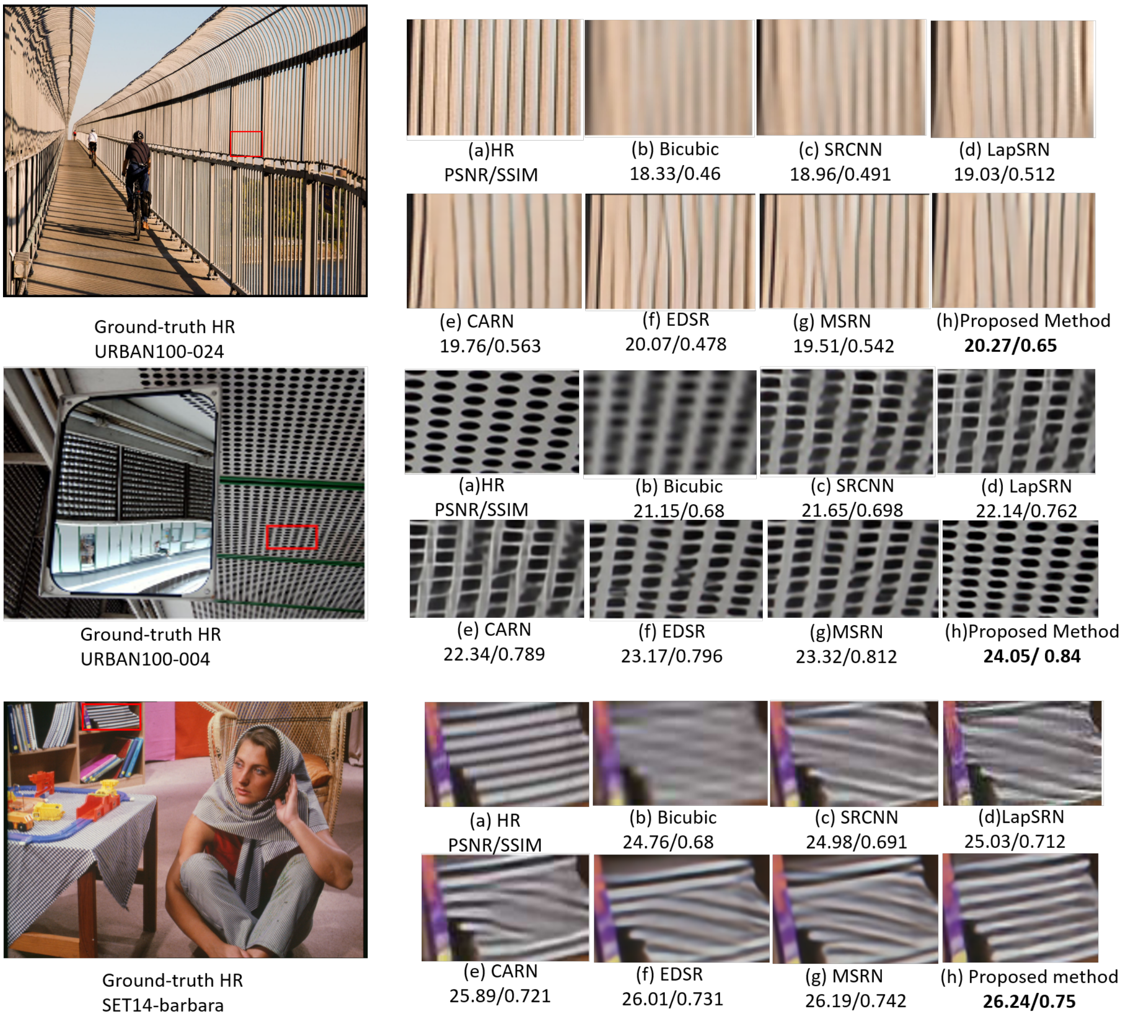


Figure 3.4: Visual Comparison on some challenging images from Set14 [31] and Urban100 datasets [21] where (a) denotes the cropped ground-truth HR, (b) Bi-cubic, (c) SRCNN [22], (d) LapSRN [23], (e) CARN [28], (f) EDSR [26], (g) MSRN [32], and (h) Proposed method.

important regions like texture and edges. Further, ESRGAN employs nearest neighbor interpolation as their up-sampling technique that is usually associated with blurring artefacts. On the other hand, our proposed up-sampling technique is effective in removing those unpleasant artefacts.

Computational Complexity

In Table 3.3, we compare our proposed MBUp-Net with the most popular SoTA super-resolution methods in terms of network parameters, FLOPs and execution time. Our MBUp-Net obtains the best results with lesser parameters. This certainly demonstrates that our method can very well balance the trade-off between the number of parameters and reconstruction performance. To further, reflect the efficiency of our method, we also compare the running time of MBUp-Net on Set5 [38] with other competitive methods. It is quite apparent from Table 3.3, that our proposed model is the fastest among all others.

Table 3.4: Ablation study on CAFD block, CA block and MLA block. The model is trained for 3 settings: With and Without CA block, MLA block, and CAFD block. Each setting is compared in performance for PSNR/SSIM on four benchmark datasets. Note: ✓ and ✗ indicate the model, with and without corresponding blocks, respectively.

Scale	CA	MLA	CAFD	Set5 [38]	Set14 [31]	BSD100 [21]	Urban100 [84]
×2	✗	✗	✗	35.31/0.912	30.55/0.892	28.96/0.857	28.73/0.883
	✗	✗	✓	37.64/0.943	32.41/0.901	30.16/0.878	30.91/0.908
	✗	✓	✓	38.04/0.958	32.68/0.821	31.43/0.891	30.58/0.913
	✓	✓	✓	**38.17/0.966**	**33.62/0.921**	**32.52/0.910**	**31.61/0.923**
×3	✗	✗	✗	31.78/0.879	27.88/0.814	26.49/0.764	25.14/0.824
	✗	✗	✓	33.09/0.907	29.72/0.837	28.76/0.799	27.98/0.855
	✗	✓	✓	33.56/0.915	30.01/0.841	28.99/0.800	28.17/0.862
	✓	✓	✓	**34.01/0.921**	**30.46/0.849**	**29.12/0.805**	**28.35/0.865**
×4	✗	✗	✗	29.31/0.843	25.58/0.746	25.96/0.692	24.73/0.713
	✗	✗	✓	31.54/0.859	27.88/0.771	27.62/0.710	25.18/0.719
	✗	✓	✓	31.84/0.865	27.92/0.767	27.69/0.727	25.89/0.722
	✓	✓	✓	**32.31/0.870**	**28.64/0.777**	**27.71/0.732**	**26.08/0.731**

Adversarial training

Here, we perform an additional experiment to demonstrate the effect of adversarial training [195] on our proposed MBUp-Net. We utilise our proposed MBUp-Net as the generator network and our discriminator network is inspired from SRGAN [34]. The discriminator is composed of six convolution layers, each of size 3×3 followed by LeakyReLU and a final Sigmoid layer for discriminating between real and generated SR images. We have trained the proposed network with adversarial and perceptual loss [196] and the visual result is shown in Figure 3.5 for ×4 SR. The model with adversarial training generates plausible details on regions with irregular structures like feathers. However, as visible in the zoomed portion, the reconstructed image is not able to produce faithful results in comparison to the ground-truth high-resolution image, thus lacking in accuracy when compared to the proposed MBUp-Net. Moreover, generative adversarial networks (GANs) are known to be susceptible to mode collapse, which may cause the generator to produce the same output over and over again. Though this problem can be resolved by several variants in the loss function [197], but the main problem we faced was the failure of GANs to converge which ultimately lead to unstable training.

3.1.3 Ablation Study

For examining the impact of individual architectural components of our proposed network, an ablation study is comprehensively conducted on DIV2K database. All the ablation

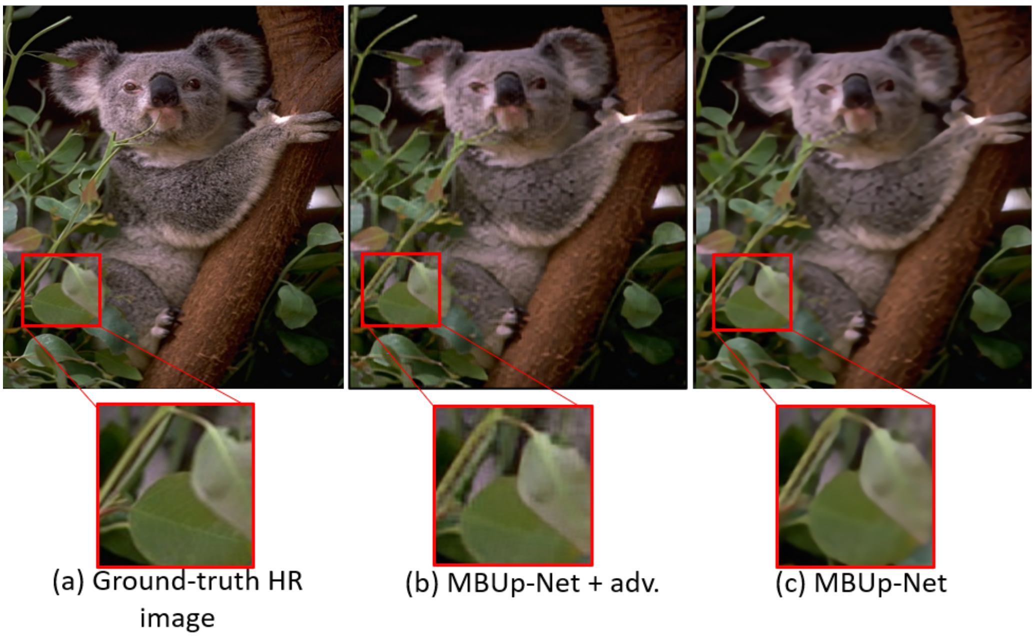


Figure 3.5: Visual analysis for adversarial training. We differentiated the results trained on our model with and without adversarial training, (b) Results with adversarial training, and (c) Original MBUp-Net results without any adversarial training.

models are trained for 10^5 iterations.

Impact of different components

Table 3.4 demonstrates the effectiveness of different modules in the proposed model. We study their impact by progressively introducing them in our model. The baseline is obtained without CA, MLA and CAFD block and performs very poorly (Here we employ just simple residual block [32] in place of CAFD block). It is quite evident that addition of CAFD block provides favorable performance gain of about (2.33 dB) over the baseline. Furthermore, addition of proposed MLA block causes 0.44 dB gain in the accuracy of the model. Table 3.4 also shows that incorporation of CA block inside our proposed MLA block further increases the PSNR score. It is evident that all the modules used together contribute positively towards the final image quality, indicating the importance of considering difference among features and further paying attention to channel-wise features.

Table 3.5: Ablation study on the effect of proposed up-sampling technique (MBU). Results are calculated for PSNR on Set5/Set14 for scale factor ×4.

Up-sampling Technique	Parameters	PSNR
Conv2DTranspose [198]	65.5K	32.21/28.45
Sub-Pixel Convolution [34]	55.3K	32.33/28.66
Bicubic Maximum	**30.5K**	**32.56/28.89**

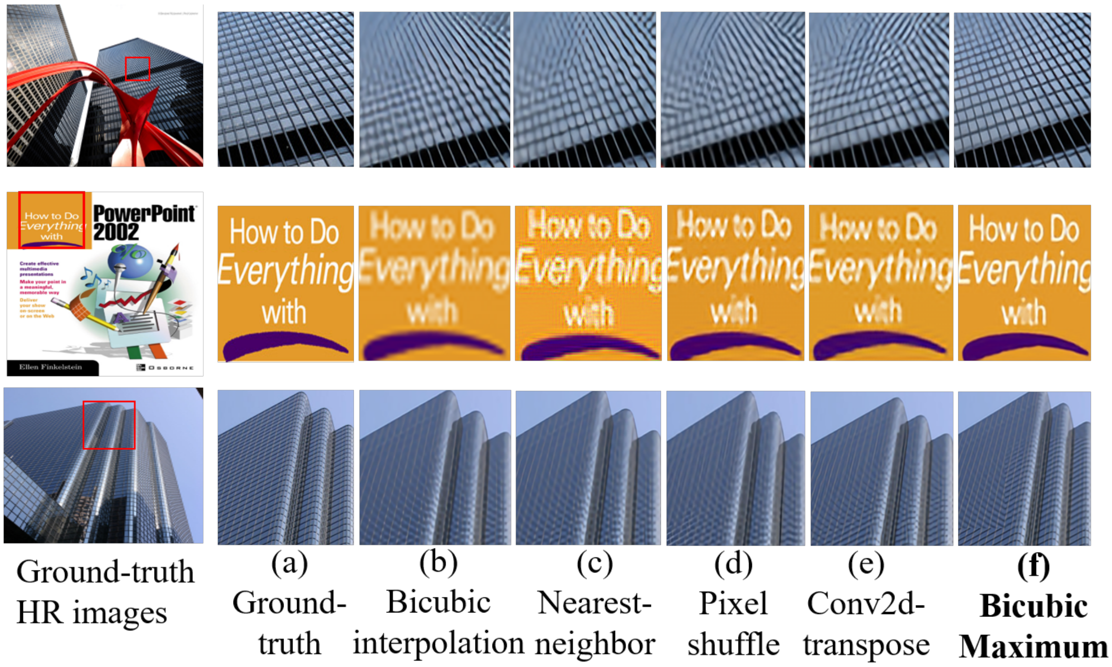


Figure 3.6: Visual analysis for different up-sampling layers. It is evident that (b) and (c) [33] introduce blurring, (d) [34] and (e) [35] introduce checker-board artefacts, and (f) bicubic maximum overcomes all the defects and obtains realistic images.

Effect of different up-sampling layers

The traditional non-learnable interpolation techniques like nearest-neighbor, and bilinear considered the spatial distance between the pixels for ushering the up-sampling process, hence it failed to add much new information to the LR image. As visible from Figure 3.6 (b, c) if our proposed network (under similar training settings) uses traditional up-sampling layers, the outputs of different images from Set5 [38], Set14 [31], and BSD100 [21] have certain artefacts and blurring. Similarly, subpixel layer [34] also leads to artefacts around the boundaries of different objects as shown in the Figure 3.6 (d). Deconvolution layer or transposed convolution layer generates checkerboard like-pattern and ultimately degrades the quality of the image which cannot be avoided even after proper learning as is visible in Figure 3.6 (e). Unlike other techniques, our proposed MBU block as shown in Figure 3.6 (f) demonstrates visually plausible results by generating sharper and crisp edges. Table 3.5 reveals that replacing the proposed up-sampling layer with popular up-sampling layers (Conv2DTranspose and pixel-shuffle) reduces the PSNR score with an increase in the number of parameters. Thus, we can conclude that the proposed MBUp-Net generates realistic-looking images closer to the ground truth and overcome the artefacts with a significant reduction in parameters.

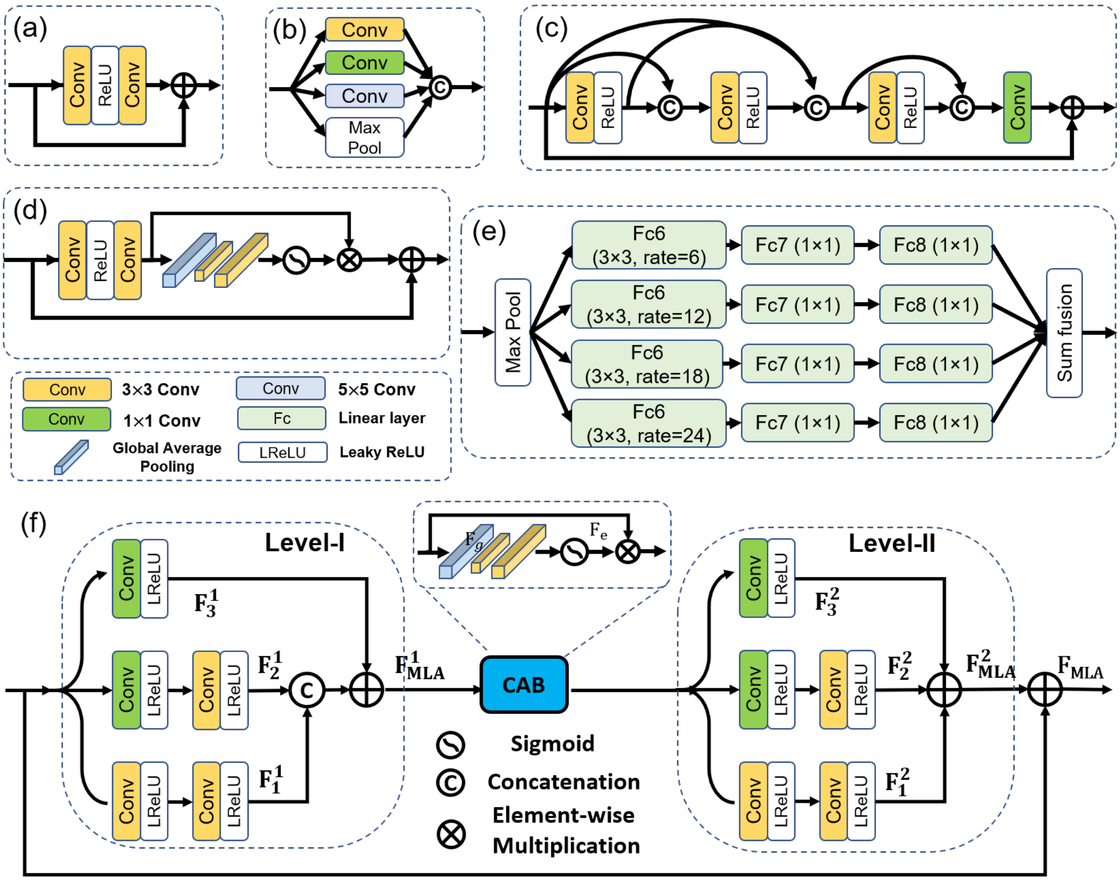


Figure 3.7: Comparison with prior network configurations. (a) Simple Residual block (Configuration1) [26], (b) Inception block (Configuration2) [32], (c) Dense block (Configuration3)[30], (d) Residual channel attention (RCAB) block (Configuration4) [36], (e) Atrous spatial pyramid pooling (ASPP) block (Configuration5) [37], and (f) our proposed Multi-level Attention Residual block.

Investigation on different network configurations

Next, we analyze the importance of our proposed MLA block by comparing it with other popular residual configurations as shown in Figure 3.7. All the configurations (Configuration1-6) are trained on the same dataset. As shown in Table 3.6, it is inevitable that for enlargement by factor ×4, the proposed multi-level attention block shows better performance on PSNR/SSIM for Set5 [38] and Set14 [31]. This improvement in the performance gain of the proposed configuration indicates that multi-scale feature extraction along with attention mechanism via MLA block is more adept for capturing contextual information as compared to simple residual block [199], or multi-scale residual [200, 36, 37] and dense blocks [30]. Further, we also demonstrate the convergence analysis of different residual configurations in Figure 3.8.

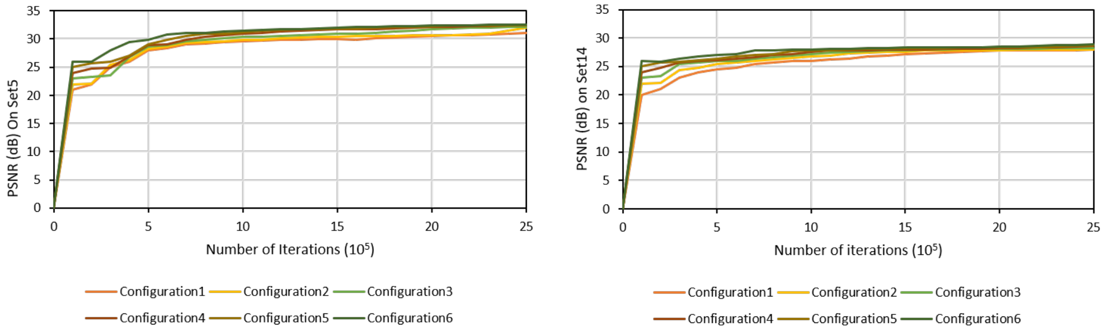


Figure 3.8: Convergence analysis for different residual configurations shown in Table VII. The curve is evaluated on Set5 [38] and Set14 [31] for ×4.

Table 3.6: Quantitative evaluation of PSNR/SSIM for different residual block configurations on Set5 [38] and Set14 datasets [31].

Residual Configurations	Set5 [38]	Set14 [31]
Configuration1 [26]	31.13/0.8656	27.92/0.7515
Configuration2 [32]	31.96/0.8674	28.01/0.7563
Configuration3 [30]	32.16/0.8866	28.25/0.7896
Configuration4 [36]	32.45/0.8897	28.67/0.7978
Configuration5 [37]	32.37/0.8784	28.69/0.7765
Configuration6	**32.54/0.8999**	**28.87/0.7988**

Table 3.7: Analysis of the number of CAFD blocks for PSNR/SSIM on Set5, Set14 and BSD100 datasets for × 4.

Approach	Set5 [38]	Set14 [31]	BSD100 [21]
with 4	32.36/0.8882	28.80/0.7871	27.69/0.7458
with 10	32.56/0.8990	28.85/0.7884	27.74/0.6462
with 16	32.62/0.8997	28.90/0.7889	27.81/0.7469
with 20	**32.63/0.8998**	**28.91/0.7889**	**27.81/0.7471**

Table 3.8: Performance evaluation of proposed MBUp-Net with and without edge information as input in terms of PSNR/SSIM.

Configuration	×2	×3	×4
Without edge prior	**38.31/0.9626**	**34.76/0.9296**	**32.67/0.9005**
With edge prior	38.00/0.9596	34.45/0.9276	32.45/0.8987

Table 3.9: Effectiveness of the proposed MLA block.

Method	×2	×4
MBUp-Net (With MLA)	**38.31/0.9626**	**32.67/0.9005**
MBUp-Net (Without MLA)	37.99/0.9463	32.33/0.8964

Investigation on using maximum operation in up-sampling

On account of the disadvantage of the learnable up-sampling methods, in our proposed up-sampling layer, bi-cubic interpolation followed by a convolution layer has been applied.

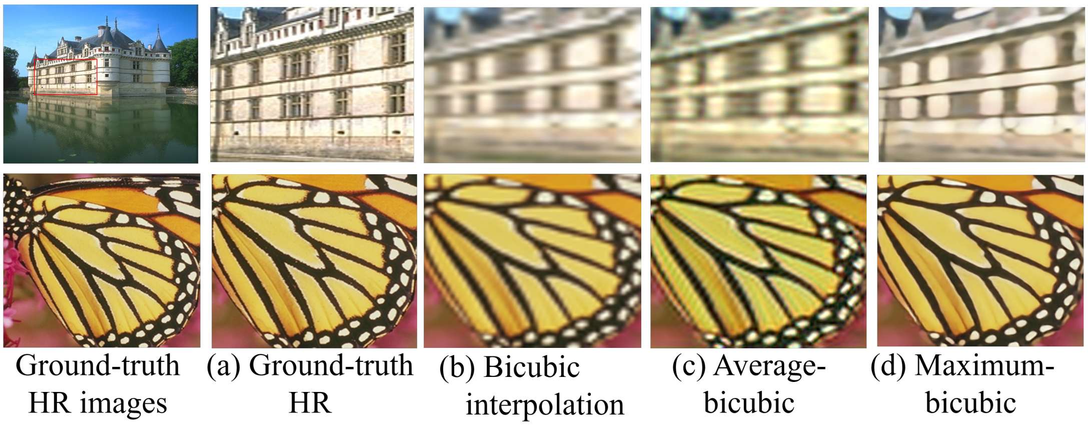


Figure 3.9: Visual analysis of our proposed maximum bi-cubic up-sampling versus average bi-cubic up-sampling and bi-cubic interpolation.

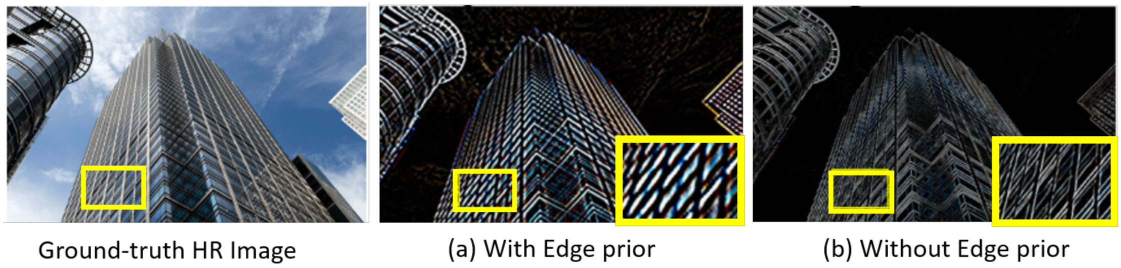


Figure 3.10: Visual comparison of the feature maps generated by our proposed CAFD block where, (a) represents the output with prior edge input, and (b) represents the output of our CAFD block. It is quite apparent that (b) generates high-fidelity output with better content preservation.

For better extraction of important features while reconstruction, maximum of all the channels is further considered to up-sample the features efficiently. For better analysis, we have compared the proposed maximum bi-cubic results with bi-cubic interpolation followed by the averaging operation and simple bi-cubic interpolation, as shown in Figure 3.9. The visual results using the maximum operation prove the efficiency of the maximum operation in generating high-frequency details.

Impact of number of Content-Aware Feature Difference block (CAFD) blocks

The capacity of the proposed network is determined mainly by the number of CAFD blocks. In this study, we test the effect of this parameter on image SR. Table 3.7 distinctly states that increasing the CAFD blocks beyond 20 causes very little refinement in the performance of our proposed network. Therefore, we opt for 20 CAFDs as a balanced choice.

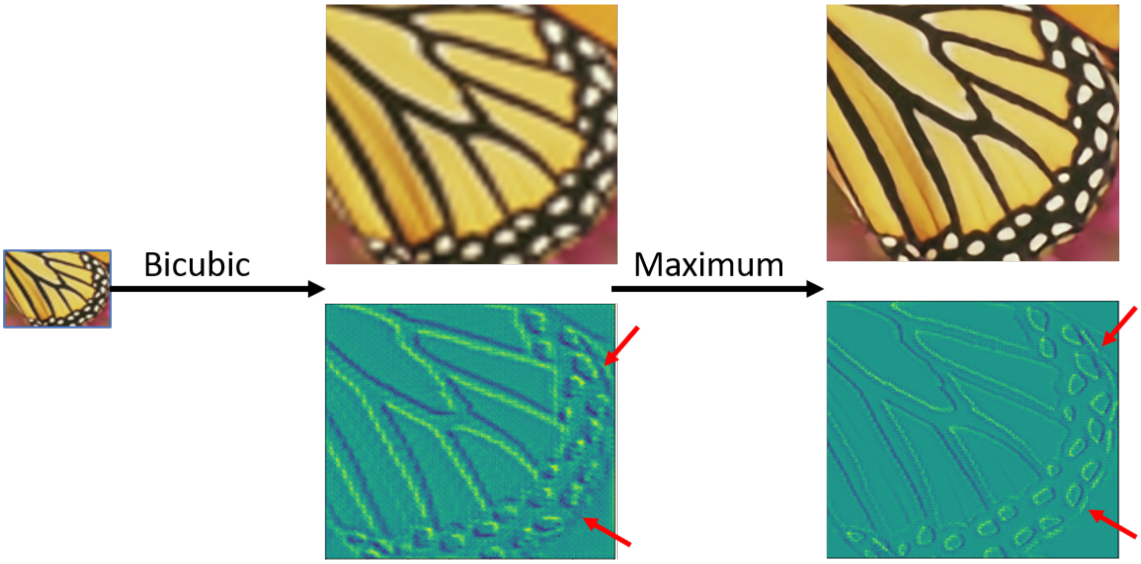


Figure 3.11: Intermediate feature maps after the application of bi-cubic maximum operation.

Effectiveness of CAFD and MLA block

To prove the efficacy of our proposed MLA block in preserving edges, we perform an experiment: where we first extract the edges of the input LR image by utilizing EdgeNet module (deep edge extraction) from SoftNet [178] and gave the extracted edges as a content input in MLA. As shown in Table 3.8 and Figures 3.10 and 3.11, giving edge prior information to MLA block reduces the overall accuracy of the model. The main reason for the results could be attributed to the fact that main goal of CAFD block (composed of MLA blocks) is the preservation of the overall content (low and high frequency) in an image. But, giving edge information as the input focuses only on high-frequency content thereby affecting the overall quality of the output reconstructed image.

To further prove the potency of our proposed MLA block in capturing contextual information, we perform an additional experiment. Basically, the role of our proposed MLA block is to extract the global contextual information by increasing the receptive field and its employment inside CAFD blocks further facilitates the preservation of high and low-frequency content. *Since, CAFD relies heavily on MLA, what if MLA is just a simple convolution layer?* Generally, in a simple convolution layer, every filter operates with a local receptive field and its resulting output feature map is unable to exploit the contextual information outside the local region. Henceforth, if we replace the MLA block with a simple convolution layer it would not be able to properly highlight the extracted features thus resulting in poor preservation of the content information and lower accuracy of the model. Table 3.9 shows a comparative analysis of our approach MBUp-Net (with MLA block) and MBUp-Net (without MLA block), where we replace the MLA block with a simple convolution layer. It is inevitable that for both the compared scale factors, MLA block shows much better performance in terms of PSNR and SSIM.

In summary, we discussed the impact of each component (up-sampling layers, popular network architectures, CAFD block, MLA block and the number of CAFD blocks) of our proposed MBUp-Net. All these components together account for better feature learning by generating plausible high-resolution images.

3.2 (MLE^{2}A^{2}U)-Net: Image Super-Resolution via Multi-Level Edge Embedding and Aggregated Attentive Upsampler Network

Albeit the rapid advancement of high profile convolutional neural networks (CNNs), the CNN-based SR methods have achieved remarkable progress, being quite successful in utilising the image statics inherent in the training datasets. *But what should be the next progress for SISR?* Nevertheless, the popular SR methods [26, 45, 201, 39, 42] showed remarkable results. But, still there is a great urge to revisit the various components and reconsider the previous methods to search for more efficient SR model and address some issues like: (1) How is it possible to improve the capability of the overall SR network and to utilize the low-frequency information for preserving the high-frequency details, **without involving much deeper architectures**. (2) Most of the recent SR approaches focus on using pixel-shuffle or Conv2D-Transpose as the up-sampling layers, **without exploring much the other possibilities for up-sampling**. (3) **Does enlarging the receptive field, really improves the final performance of the SR network?** and (4) Though, the local and the non-local attention blocks lead to good performance in SR results, but what is **the exact position in the SR architecture to place these blocks**.

For addressing the above issues and to improve the SR performance, we propose a purposeful method to improve the performance of super-resolution without using much deeper architecture. The overall architecture of the proposed method is shown in Figure 3.12. In summary, the main contributions of our work are four-fold:

1. An edge embedding and attentive up-sampler network (MLE^{2}A^{2}U-Net) is proposed to improve the desired high-frequency details and preserve the low-frequency information for SISR task, by incorporating both local and non-local attention mechanisms.

2. We design a novel multi-level edge embedding module, with stacked novel multi-receptive field extractor blocks, to exploit the features at various scales and reuse them for the next scale to faithfully preserve the precise spatial details at each resolution.

3. Also, we propose a novel aggregated attentive up-sampler block, aggregating information from the popular SoTA up-sampling layers, in an attentive manner to adaptively rescale the required important features.

4. We also propose a lightweight version of the proposed network (MLE^{2}A^{2}U-Net), MLE^{2}A^{2}U-Net$_L$ by adjusting the hyper-parameter settings of the main network.

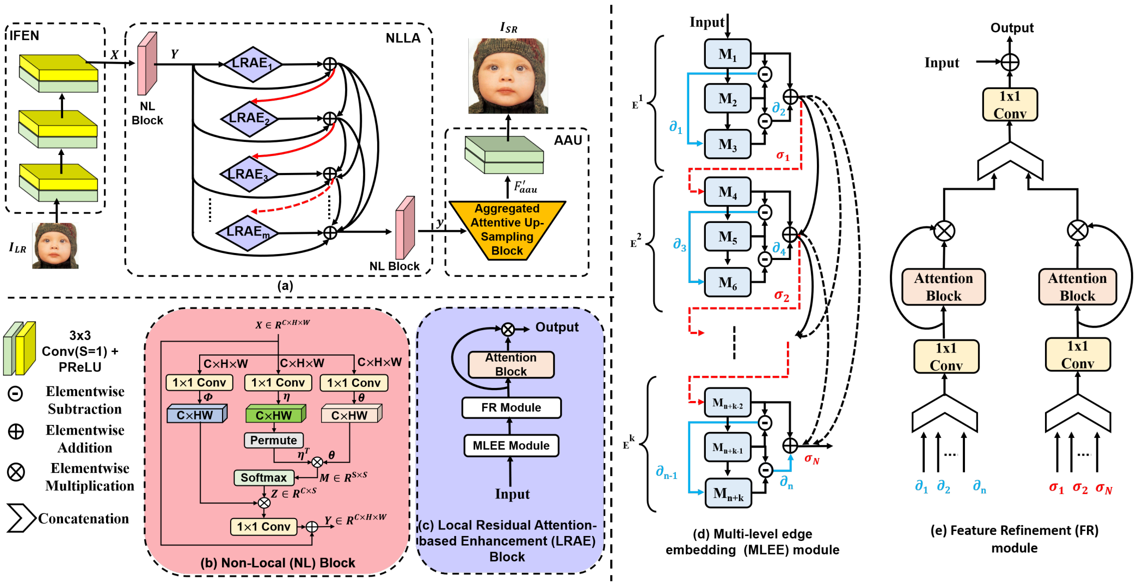


Figure 3.12: (a) The proposed MLE^{2}A^{2}U-Net architecture for SISR, (b) Non-local (NL) block, (c) Local residual attention-based enhancement (LRAE) block, (d) Multi-level edge embedding (MLEE) module, and (e) Feature refinement (FR) module.

3.2.1 Proposed Method

For SISR application, stacking profoundly deep networks with variations in the architectural designs, loss functions and attention blocks is an efficient way of representing the non-linear mapping between input LR and HR image. However, there is a requirement to analyse the various components that goes in the designing of SR architecture and finalize a generic model capable of inheriting the strengths of the existing approaches.

Inspired by this, we propose an architecture, outlined in Figure 3.12, comprising of an initial feature extraction network (IFEN), a non-local and local attention-based (NLLA) block and an aggregated attentive up-sampler (AAU) block. Specifically, IFEN network is proposed to represent the input LR image as a set of feature maps via convolution layers, with PReLU as the activation function. Taking the extracted LR features as input, the proposed NLLA block focuses on extracting more informative features, by paying more attention to the detail fidelity. Next, embedding of the acquired deep features to detailed up-scaled features, is obtained through an AAU block. Now, we individually illustrate the design details of the two basic components of the proposed SR model, including the NLLA block and AAU block.

Non-Local and Local Attention Block

To make a trade-off between the local and non-local properties of features, NLLA block is built comprising of non-local block and local residual attention-based enhancement block. Such a combination scheme, not only gathers the contextual information within the local receptive field but also exploits the information outside the local region. Non-local operations are more useful for images with repetitive details, whereas local operations are a suitable choice for images with complex textures [202] and when used together, they can complement each other and improve the reconstruction performance. Further, to bypass redundant low-frequency information in the input LR images and to facilitate the training of our network, we have utilised several share-source skip connections [52] and local source skip connections in our proposed NLLA architecture.

Non-local Block

The non-local (NL) block [126] aims to strengthen the features of the query spatial position, through aggregation of information from all other spatial positions as shown in Figure 3.12 (b). NL block can be considered as global context modelling block accumulating query-specific global context features to each query position.

For an input feature $X \in R^{C\times H\times W}$ where, C, H, W denote the number of channels, spatial height and spatial width, respectively. Initially three 1×1 convolutions, W_θ, W_ϕ and W_η are used to transform the input feature X, into different embeddings, θ, ϕ and η.

$$\theta = W_\theta(X), \phi = W_\phi(X), \eta = W_\eta(X) \tag{3.12}$$

Following it, the three embeddings are flattened to size $C \times S$, where S indicates the total number of spatial locations and $S = HW$. The correlation matrix, C in embedding space is defined in our work by Embedded Gaussian [126] and is calculated by matrix calculation as,

$$M = \eta^T \theta \tag{3.13}$$

where $M \in R^{S\times S}$. Then, we apply a normalization operation to M, in the form of Softmax, and get $Z \in R^{C\times S}$ as the output, which is defined as,

$$Z = Softmax(M) \times \phi \tag{3.14}$$

The final output, $Y \in R^{C\times H\times W}$ is obtained by adding the original input X with the weighting parameter W_c, implemented through 1×1 convolution.

$$Y = W_c(Z) + X \tag{3.15}$$

Thus, the non-local block helps the network to access long-range information via its flexible residual architecture by making use of less number of layers and parameters. Moreover, since the non-local block helps in capturing better interactions between any two positions make it a suitable candidate for SR applications.

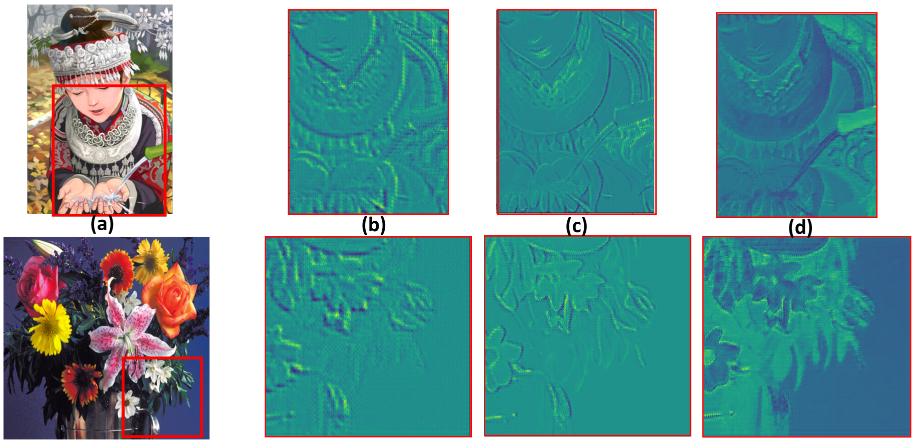


Figure 3.13: Visualization of the attention maps of the three different outputs of an nth MLEE block. (a) Represents sample input images from Set14 [2] dataset, (b), (c) Attention maps of $\partial_n{}'s$ (d) Attention maps of $\sigma_n{}'s$.

Local Residual Attention-based Enhancement Block

After non-local operations, to enhance the features via extracting their spatial correlation, a local residual attention-based enhancement block (LRAE) comprising of multi-level edge embedding (MLEE) module with multi-receptive field extractor (MRFE) module stacked inside it is proposed. Since, LRAE blocks are itself composed of several MLEE modules, incorporating skip connections between these modules help in better propagation of the contextual information towards end of the network by avoiding the problem of vanishing gradients. The MLEE module has been further followed by a feature refinement module to exploit better correlation among the features as shown in Figure 3.12 (c). Various components of the LRAE block are discussed in detail as follows:

Multi-Level Edge Embedding Module In the proposed module, we target at two main problems in SR. Firstly, for SR task, the feature distribution is varying across different frequency bands, with low-frequency information consisting of simpler textures and high-frequency information consisting of complex textures [203]. Hence, a dedicated architecture for the preservation of low and high-frequency information is required. Secondly, since the receptive field of CNNs tends to grow slowly with increase in the depth of network [204], thereby adversely effecting the extraction of long-range relationships among pixels, hence we require a remedial solution for the same. Inspired by the work in [204], we propose a novel multi-level edge embedding (MLEE) module. *The proposed MLEE module, is designed to boost the enhanced features at the current level and to pass the learnt high-frequency details onto the next level using share-source skip connections.* To further enhance the representational ability of the different resolution features and to effectively remedy the missing information from multi-levels, the proposed MLEE module consists of several MRFE blocks stacked together in a multi-level way. Specifically, as

shown in Figure 3.12 (d), k-level MLEE stacks a chain of n multi-receptive field extractor blocks, MRFE $\{M_m\}_{m=1}^{n}$. Considering the k^{th} level MLEE module be denoted as E^k, the input of which consists of features from the previous embedding module, E^{k-1}. Further, for every k^{th} embedding module we have used three MRFE (M_n) blocks, outputting two subtraction (∂_n) and one addition operation (σ_n). As clear from the Figure 3.12 (c), the last embedding module E^k, receives input from all the preceding levels of MLEE modules alleviating the issue of vanishing gradient [205].
The overall operation of the k^{th} embedding block with $M_{n_{th}}$ MRFE block is defined as,

- Compute the difference ∂_n between the adjacent MRFE features M_n at the same level. For every level of edge embedding module, we define the difference output as,

$$\partial_n = M_{n+k} - M_{n+k-1}, \ \forall n = [1, 10]; \ k = \left\lfloor \frac{n+1}{2} \right\rfloor \tag{3.16}$$

- The enhanced feature σ_n is obtained after considering the boosted features from all the previous levels and detailed features from the same level and is given by

$$\sigma_n = \sigma_{n-1} + M_{3n-2} + (M_{3n} - M_{3n-1}), \ n \in [1, 5] \tag{3.17}$$

The feature maps obtained after the subtraction operation, ∂_n encodes the details like edges and textures, whereas the feature maps obtained after addition operation, σ_n encodes the high level semantic and detailed information as shown in Figure 3.13. Since, both the extracted features consists of redundant and complimentary information and focus on different frequency information **(with $\partial_n^{\prime}s$ focusing on high-frequency and $\sigma_n^{\prime}s$ focusing on the low-frequency information)** in the image, directly combining them using simple concatenation or addition operations could ignore the information between different layers. Hence, we introduce a feature refinement module for exploitation of the complementary information.

Multi-Receptive Field Extractor Block Since SR requires prediction for every single pixel in the input LR image, it is important for every output pixel to have large receptive field in order to avoid missing important information while making predictions. Moreover, limited spatial size of the input image is insufficient to learn the diverse LR-to-HR mappings. To relieve this situation by rethinking the influence of multi-scale learning and skip connections in the field of SR, we design a novel multi-receptive field extractor (MRFE) block as shown in Figure 3.14 (a). It is capable of extracting features from various receptive fields to cover different shapes and textures of the objects. In MRFE block, different levels of the convolution layers corresponds to different degrees of feature information extraction. The cross-scale concatenation between the multi-scale features is

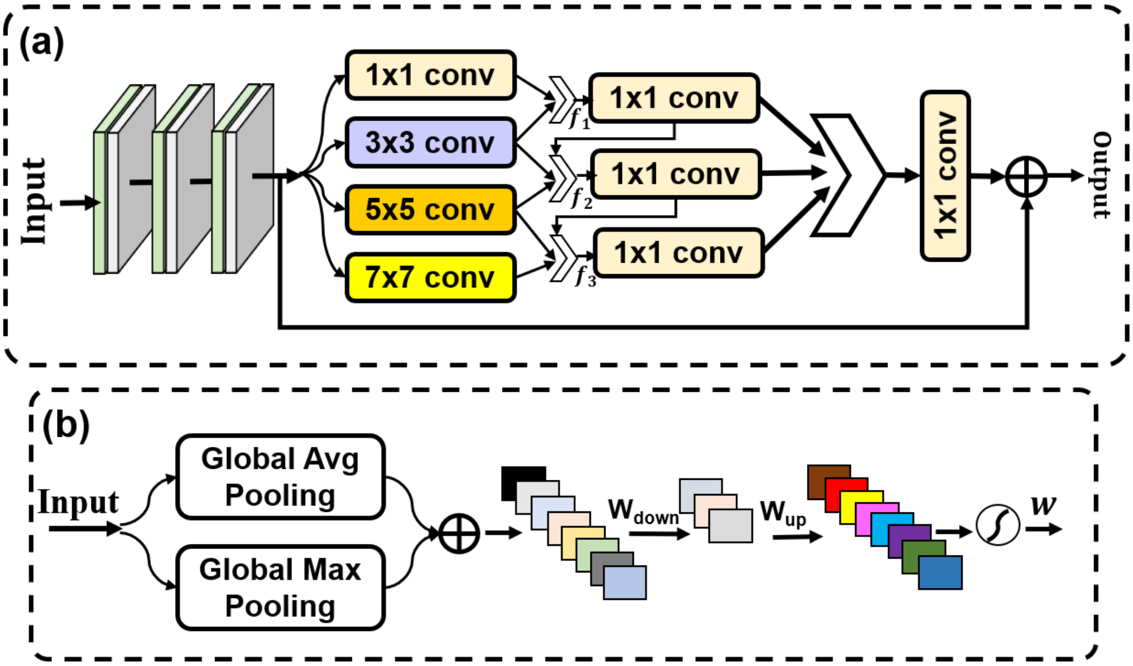


Figure 3.14: (a) The proposed Multi-receptive Field Extractor (MRFE) block. Represented as M in Figure 3.12 (d), (b) Attention block used in Figure 3.12 (c) and Figure 3.12 (e).

shown in Figure 3.14 (a). The obtained feature subset, f_i at each level i, where $i \in \{1, 2, 3\}$ are further concatenated and passed through the 1×1 convolution layer. Notice, that at each level, f_i could potentially receive information from all the previous levels $f_j; j \leq i$. This way of feature representation not only allows the features to be explored sequentially, but also explores their distinctive complimentary characteristics. Further, passing the fused features of one level to other level helps in better sharing the information at different scales.

Contrary to the existing MDCN [115] which uses dual path network by densely connecting the features at different scales, our MRFE is capable of exploiting more contextual information with increased receptive field by utilizing less number of parameters. The receptive field, *(RF)* is generally calculated as,

$$RF_i = RF_{i-1} + (k-1)\prod_{i=1}^{L} S_i \tag{3.18}$$

where, i denote the layer under consideration. L denotes the total number of layers in the network, S denotes the cumulative stride of all the previous $i-1$ layers and k indicates the filter size. Specifically, after passing through k level MLEE block with n MRFEs, the receptive field RF increases progressively helping the network for gathering more contextual information.

Feature Refinement Module The addition σ_n and subtraction ∂_n outputs of the MLEE module contain complimentary and redundant information and directly fusing these different features may limit the expressive power of the overall network. Hence, we need to design a novel way of fusing these contradictory feature response from different layers. Most deep SR networks integrating multi-level features, directly by concatenation or addition operations (without any post-processing), generally ignores the gap between different features and may lead to some undesirable artifacts [206]. Working on this goal, in our proposed feature refinement (FR) module, we propose a non-linear mechanism for combining the complimentary features using attention mechanism. Our FR module separately receives the subtraction and addition outputs of all the MLEE modules as inputs as shown in Figure 3.12 (e). Firstly, all the subtraction outputs ∂_n's are concatenated, and then these discriminative concatenated feature maps are passed through a 1×1 convolution layer to generate fused subtraction features (FSF). The FSF is defined as,

$$FSF = W_S * ([\partial_1, \partial_2, \partial_3 ..., \partial_n]) + b_S \tag{3.19}$$

where, W_S, b_S denote the weight and biases of the convolution layer for FSF features, learned during training. ∂_i denotes the i^{th} feature map obtained after subtraction operation, where $\partial_i \in R^{C\times H\times W}$.

On a similar note, the fused addition features (FAF) aggregating all the $\sigma_n{}'s$ is defined as,

$$FAF = W_A * ([\sigma_1, \sigma_2, \sigma_3 ..., \sigma_n]) + b_A \tag{3.20}$$

where, W_A, b_A denote the weight and biases of the convolution layer for FAF features learned during training.

To further remove any redundant information, the feature maps of FSF and FAF are passed through an attention block to adaptively re-calibrate the feature maps. Then, the refined FSF and FAF are aggregated through concatenation. For selection of useful multi-level information with respect to the features of every layer and for reducing the number of channels to the original channels, a final convolution layer with weight W_R and bias b_R has been added after the aggregated features. The final refined feature (FRF) maps can be formulated as,

$$FRF = W_R * ([FSF, FAF]) + b_R \tag{3.21}$$

Attention Block To strengthen the role of useful feature channels (excite) and weaken the role of useless channels (squeeze), we ought to assign different weights to separate channels [201]. Contrary to the previous SR works [44], [52] which uses only global average pooling to capture the global statistics of feature maps, we leveraged both global average pooling (GAP) and global max pooling (GMP) to aggregate the global information. Since, spatial details are more important for the discriminability of SR network, and max-pooling being a good option for preserving the most activated pixels, hence it becomes a suitable choice in our network design. For excitation operation, we re-calibrate the feature descriptor (*obtained after merging the two squeezed feature vectors by element-wise addition*) by using two convolution layers with weights W_{down} and weights W_{up}. After receiving the excited feature maps, we pass it through Sigmoid function to capture the channel-wise dependencies from the aggregated information. The overall operation of the attention block is summarised as:

$$\begin{aligned}
&w = sig(W_{up}(\mathrm{Re}LU(W_{down} * ({z_c}^{avg} + {z_c}^{\max})))) \\
&{z_c}^{avg} = GAP(X_c) = \tfrac{1}{h \times w} \sum_i \sum_j X_c(i,j) \\
&{z_c}^{\max} = GMP(X_c) = \max(X_c(i,j))
\end{aligned} \tag{3.22}$$

where, *sig* represents the Sigmoid function, inserted at the end of attention block.

The attention block has been inserted in every module of our network and serves several purposes as mentioned below:

- In the proposed upsampler block, it serves the benefit of sharing information within a tensor along the channel dimension.
- In the proposed feature refinement module which receives several ∂_n and σ_n as inputs, attention block serves the purpose of discriminatively focusing on the important information and bypassing the redundant information.
- Since different scales of the receptive fields generate features with different levels of discrimination, hence an attention module is added in LRAE module after every FR module for modelling rich contextual dependencies over the local features.

SR being a pixel-to-pixel correspondence task, proper care must be taken, while preserving the fine spatial details and on collecting long-distance spatial contextual information. Towards this goal, as discussed our designed NLLA block with MLEE module is suitable for capturing the fine details (edges and textures), and the stacked MRFE's serves the purpose of increasing the receptive field for capturing large context of the input image.

Aggregated Attentive Up-sampler Block

Upsampling, being the key step for reconstructing the HR images from LR images, any improper up-sampling technique may lead to certain redundant information in the reconstructed image. Majority of the SoTA SR methods adapt similar type of up-sampling techniques by either using deconvolution, pixel-shuffle or learn-able convolution layers. Bilinear and Bicubic interpolation are implemented in accordance with the spatial

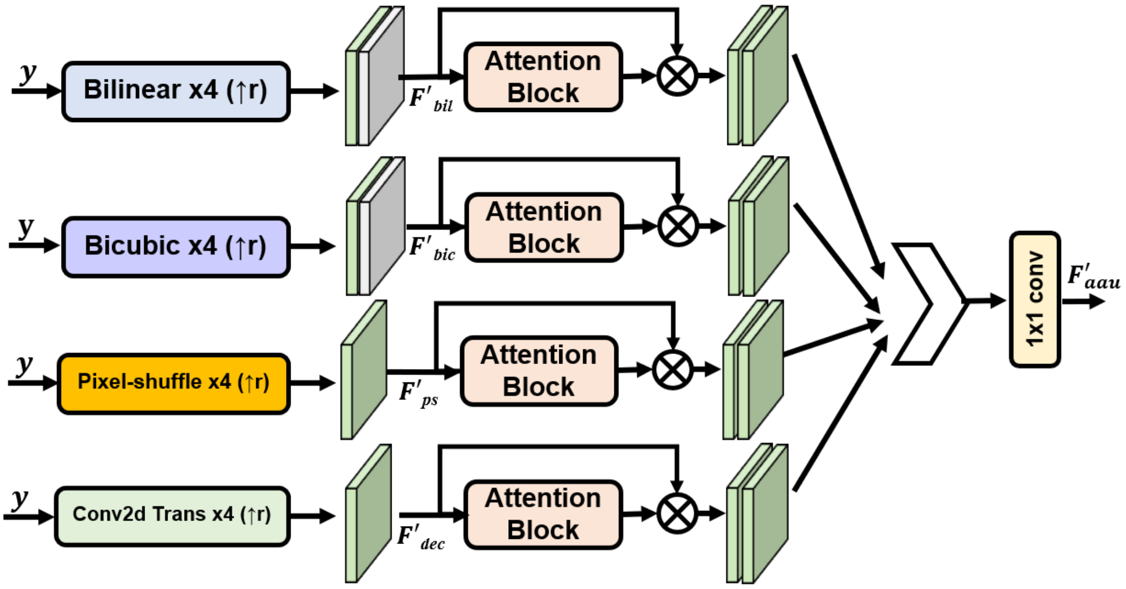


Figure 3.15: The proposed Aggregated Attentive Upsampler (AAU) block for ×4 upsampling.

distances, the basic difference between the two is the number of positions being considered (4 pixels for bilinear and 16 for bicubic). Images reassembled with bicubic interpolation tend to produce smoother outputs with less interpolation distortion. Bilinear interpolation also tends to remove some visual distortions. The operation of transposed convolutions, commonly known as deconvolution layers, is equivalent of interleaving the input features with 0's and then applying a standard convolution layer. Pixel-shuffle, also known as sub-pixel convolution, shifts the feature channels into spatial dimensions and preserves all floats inside the high-dimensional representation of the image, since it only changes the placement of pixels. However, these techniques when incorporated individually in the network may not be that much effective, as they introduce certain redundant information in the reconstructed image. Moreover, these methods map the LR image to SR image with content-irrelevant up-sampling weights, which implies that weights for up-sampling are same for different pixels. Hence, it may result in over-smooth SR results. To combine the advantages of all the up-sampling layers and to overcome the shortcomings imposed, when each of them used individually, we propose a novel Aggregated Attentive Up-sampler (AAU) block, which aggregates all the up-sampling techniques in an attentive way. The structure of our proposed AAU block is shown in Figure 3.15. *It adaptively learns the up-sampling weights for different pixels to produce SR results.*

Given a LR feature map $F(y) \in R^{C\times H\times W}$ to be up-sampled, our goal is to generate an up-sampled feature map $F'(y) \in R^{C\times rH\times rW}$, where r denotes the up-sampling factor. For position (*u'*,*v'*) in *F'(y)*, the corresponding source position *(u,v)* is solved by equating $u' = \lceil \frac{u}{r} \rceil$ and $v' = \lceil \frac{v}{r} \rceil$. We aim to learn an up-sampling weight w for each position in *F'(y)*. Applying these learned weights to each channel of the up-sampled feature map denoted

Table 3.10: Hyper-parameters setting for the proposed model and its lightweight version.

Method	MLEE (E)	MRFE (M)	LRAE (L)
MLE^{2}A^{2}U-Net	5	15	5
MLE^{2}A^{2}U-Net$_L$	2	6	5

by $F'(y) \in R^{1 \times rH \times rW}$, helps to dynamically characterize the useful set of kernels from each branch. We further ensemble the features through concatenation to adaptively learn more accurate representation, after passing the attentive response from every up-sampling branch, through two convolution layers to get the aggregated response as $F^{'}_{aau}$.

$$(F^{'}_{aau})_c = [(w_{bic} \times F^{'}_{bic})_c, (w_{bil} \times F^{'}_{bil})_c, (w_{ps} {\times} F^{'}_{ps})_c, (w_{dec} {\times} F^{'}_{dec})_c]) \tag{3.23}$$

Here, $F^{'}_{aau}$ represents the final up-sampled feature map, obtained after passing through the AAU block. $[\cdot]$ denotes the concatenation operation performed after passing the attentive weight response, through two convolution layers with filter size 3×3. $(w_{up} \times F^{'}_{up})_c$ denotes the element-wise multiplication of the corresponding learned up-sampling weights (given by Eq. 3.22) of the c^{th} channel and the up-sampled feature map *(Here, up represents the different up-sampling techniques (Bicubic (bic), Bilnear (bil), Pixel-shuffle (ps) and deconvolution (dec)))*. It is worth noting that to make bilinear and bicubic up-sampling comparable to the other learning methods, we have used an additional convolutional layer after the up-sampling operation.

3.2.2 Experimental Analysis

In this section, initially, we provide the experimental setup followed by the comparison between the proposed MLE^{2}A^{2}U-Net and other state-of-the-art methods on several benchmark datasets for SISR. Further, we discuss the contributions of the different components in the proposed model via detailed ablation study.

Experimental Details

Dataset

The performance of the proposed MLE^{2}A^{2}U-Net has been validated on Set5 [1], Set14 [2], BSD100 [3], Urban100 [4] and Manga109 [49] datasets for SISR.

In this work, we have trained the proposed network on DIV2K dataset [5] that is composed of high-quality 800 training images, 100 validation and 100 testing images. Data augmentation in form of horizontal flipping and rotation by 90° is performed on 800 DIV2K images. Degraded data for training has been further obtained by bicubic interpolation.

Training Details

Training hyper-parameters have been implemented in PyTorch, using NVIDIA DGX station with processor 2.2 GHz, Intel Xeon E5-2698, NVIDIA Tesla V100 1×16 GB GPU. Following the settings of [26], we have pre-processed the images, by subtraction of the mean RGB value of the DIV2K dataset. Sixteen low-resolution patches of size 48×48 has been randomly sampled from the training batch. The patch size for ground-truth HR images depends on the corresponding scale factor. The model had been optimized using ADAM optimizer [207] with $\beta1$, $\beta2$ and ϵ set to 0.9, 0.999 and 10^{-8}, respectively. During training, learning rate is set to 10^{-4} and decreases after every 2×10^{5} mini-batch updates. The network has been trained with $L1$ loss for 1000 epochs. Following, the trend in SISR, evaluation of the measured PSNR and SSIM metrics has been done on the luminance channel of the image, with boundary pixel cropping.

Implementation Details

As shown in Figure 3.12, our proposed MLE^{2}A^{2}U-Net depends on various hyper-parameters, including the number of LRAE blocks (L), MLEE (E) modules and the number of MRFE (M) modules in every MLEE and the specific settings of the attention block and aggregated attentive up-sampler block. For lightweight applications, we propose a lightweight version of our MLE^{2}A^{2}U-Net model as MLE^{2}A^{2}U-Net$_L$ with different configurations of hyper-parameters as shown in Table 3.10. The shown Table compare the settings of the main modules of our originally proposed model and its lightweight version. The non-local block is embedded at the beginning and the end of proposed NLLA module. For the attention block, we have set the kernel size of convolution layer as 1×1 and reduction ratio is set to 16. For convolution filters, outside the attention block, we have set the filter size and the number of filters as, 3×3 and 64, respectively. For the up-sampler block, we have set the kernel size of convolution layer as 3×3 and the number of filters $= 64$ for all the layers, except the last 1×1 layer and the deconvolution layer, for which the filter size, stride and padding are set to 6, 2, and 2 respectively for $\times$ 4 factor.

Table 3.11: Quantitative comparisons of the state-of-the-art super-resolution models on the benchmark datasets for ×2 and ×4 scale factor.

Scale	Method	Publication	Parameters	Set5 [1]		Set14 [2]		BSD100 [3]		Urban100 [4]		Manga109 [49]	
			(M)	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
×2	Bicubic			33.66	0.9299	30.24	0.8688	29.56	0.8431	26.88	0.8403	30.80	0.9399
	EDSR [26]	ECCV-2018	40.7	38.11	0.9602	33.92	0.9195	32.32	0.9013	32.93	0.9351	39.10	0.9773
	RDN [45]	CVPR-2018	5.6	38.16	0.9603	33.88	0.9199	32.31	0.9009	32.89	0.9353	39.09	0.9771
	DBPN [40]	CVPR-2018	5.9	38.09	0.9600	33.85	0.9190	32.27	0.9000	32.55	0.9324	38.89	0.9775
	RCAN [44]	ECCV-2018	15.1	38.27	0.9614	34.12	0.9216	32.41	0.9027	33.34	0.9384	39.44	0.9786
	SRFBN [46]	CVPR-2019	3.9	38.18	0.9611	33.90	0.9203	32.34	0.9015	32.80	0.9341	39.28	0.9784
	OISR-RK3 [208]	CVPR-2019	44.2	38.13	0.9602	33.81	0.9194	32.34	0.9011	33.00	0.9357	39.13	0.9773
	CSFM [125]	TCSVT-2019	12.0	38.17	0.9605	33.94	0.9200	32.34	0.9013	33.08	0.9358	39.30	0.9775
	SAN [52]	CVPR-2019	15.7	38.31	**0.9620**	34.07	0.9213	32.42	0.9028	33.10	0.9370	39.32	0.9792
	RFA-Net [43]	CVPR-2020	11.1	38.26	0.9615	34.16	0.9220	32.41	0.9026	33.33	0.9389	39.44	0.9783
	USRNet [47]	CVPR-2020	17.1	37.77	0.9592	33.49	0.9156	32.10	0.8981	31.79	0.9255	38.37	0.9760
	HAN [42]	ECCV-2020	15.2	38.27	0.9614	34.16	0.9217	32.41	0.9027	33.35	0.9385	39.46	0.9785
	TSAN [209]	TCSVT-2021	14.1	38.30	0.9619	34.17	0.9218	32.40	0.9026	**33.45**	0.9387	-	-
	DeFiAN$_L$ [210]	TIP-2021	15.1	**38.33**	0.9618	**34.28**	**0.9231**	**32.43**	**0.9029**	33.39	**0.9390**	**39.78**	**0.9797**
	MLE^{2}A^{2}U-Net		8.1	**38.35**	**0.9625**	**34.24**	**0.9233**	**32.44**	**0.9030**	**33.40**	**0.9392**	**39.49**	**0.9794**
×4	Bicubic			28.42	0.8101	25.99	0.7023	25.96	0.6672	23.14	0.6573	24.89	0.7866
	EDSR [26]	ECCV-2018	43.0	32.46	0.8976	28.71	0.7857	27.72	0.7414	26.64	0.8029	31.02	0.9148
	RDN [45]	CVPR-2018	5.7	32.45	0.8979	28.82	0.7860	27.72	0.7400	26.61	0.8020	30.98	0.9141
	DBPN [40]	CVPR-2018	10.4	32.47	0.8980	28.66	0.7839	27.67	0.7385	26.38	0.7938	30.89	0.9127
	RCAN [44]	ECCV-2018	15.4	32.62	0.8992	28.75	0.7862	27.74	0.7424	26.74	0.8058	31.18	0.9160
	SRFBN [46]	CVPR-2019	4.1	32.47	0.8983	28.81	0.7868	27.72	0.7409	26.60	0.8015	31.15	0.9160
	OISR-RK3 [208]	CVPR-2019	44.2	32.51	0.8984	28.76	0.7863	27.75	0.7423	26.78	0.8065	31.24	0.916
	CSFM [125]	TCVST-2019	12.2	32.57	0.8988	28.77	0.7864	27.75	0.7424	26.77	0.8057	31.30	0.9172
	SAN [52]	CVPR-2019	15.6	32.64	0.9003	28.92	0.7888	27.78	0.7436	26.79	0.8068	31.18	0.9169
	USRNet [47]	CVPR-2020	17.2	32.42	0.8978	28.83	0.7871	27.69	0.7404	26.44	0.7976	31.11	0.9154
	RFA-Net [43]	CVPR-2020	11.2	32.66	0.9004	28.88	0.7894	27.79	0.7442	26.92	0.8112	31.41	0.9187
	HAN [42]	ECCV-2020	15.4	32.64	0.9002	28.90	0.7890	27.80	0.7442	26.85	0.8094	31.42	0.9177
	DRLN [41]	PAMI-2020	32.5	32.63	0.9002	28.94	0.7900	27.83	0.7444	26.98	0.8119	31.54	0.9196
	TSAN [209]	TCVST-2021	14.4	32.65	0.9004	28.91	0.7888	27.81	0.7443	26.95	0.8110	-	-
	DeFiAN$_L$ [210]	TIP-2021	15.3	**32.67**	**0.9009**	**28.99**	**0.7906**	**27.84**	**0.7448**	**27.05**	**0.8134**	**31.67**	**0.9211**
	MLE^{2}A^{2}U-Net		8.2	**32.80**	**0.9100**	**28.97**	**0.7910**	**27.85**	**0.7446**	**27.05**	**0.8165**	**31.48**	**0.9198**

Table 3.12: Quantitative comparisons of the state-of-the-art light-weight super-resolution models on the benchmark datasets for ×2 and ×4 scale factor.

Scale	Method	Publication	Parameters	Set5		Set14		BSD100		Urban100		Manga109	
			(M)	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
×2	LapSRN [211]	CVPR-2017	0.43	37.52	0.9591	33.25	0.9137	32.05	0.8973	31.23	0.9188	37.88	0.9750
	CARN [50]	ECCV-2018	1.61	37.80	0.9589	33.44	0.9161	32.10	0.8978	31.93	0.9256	38.31	0.9754
	SRMNDF [212]	CVPR-2018	1.51	37.79	0.9601	33.32	0.9159	32.05	0.8985	31.33	0.9204	-	-
	OISR-RK2 [208]	CVPR-2019	1.37	37.98	0.9604	33.58	0.9172	32.18	0.8996	32.09	0.9281	-	-
	Latticenet [213]	ECCV-2020	1.74	**38.15**	**0.9610**	**33.78**	**0.9193**	**32.25**	**0.9005**	**32.43**	**0.9302**	-	-
	MADNet [214]	TSMC-2020	0.87	37.85	0.9600	33.38	0.9161	32.04/	0.8979	31.62	0.9233	-	-
	DeFIAN$_S$ [210]	TIP-2021	1.02	38.03	0.9605	33.63	0.9181	32.20	0.8999	32.20	0.9286	**38.91**	**0.9775**
	SMSR [56]	CVPR-2021	0.98	38.00	0.9601	33.64	0.9179	32.17	0.8990	32.19	0.9284	38.76	0.9771
	MLE^{2}A^{2}U-Net$_L$		2.01	**38.23**	**0.9614**	**33.84**	**0.9198**	**32.31**	**0.9011**	**32.59**	**0.9336**	**38.98**	**0.9782**
×4	LapSRN [211]	CVPR-2017	0.43	31.54	0.8850	28.09	0.7687	27.31	0.7255	25.21	0.7545	29.08	0.8853
	CARN [50]	ECCV-2018	1.65	32.12	0.8936	28.50	0.7791	27.58	0.7349	26.06	0.7837	30.45	0.9043
	SRMNDF [212]	CVPR-2018	1.55	31.96	0.8925	28.35	0.7787	27.49	0.7337	25.68	0.7731	-	-
	OISR-RK2 [208]	CVPR-2019	1.52	32.21	0.8950	28.63	0.7822	27.58	0.7364	26.14	0.7874	-	-
	Latticenet [213]	ECCV-2020	1.79	**32.30**	**0.8962**	**28.68**	**0.7830**	**27.62**	0.7367	26.25	0.7873	-	-
	LWSR [215]	TIP-2020	2.27	32.28	0.8960	28.34	0.7800	27.61	**0.7385**	**26.38**	**0.7938**	-	-
	MADNet [214]	TSMC-2020	1.00	32.11	0.8939	28.52	0.7799	27.52	0.7340	25.89	0.7782	-	-
	DeFIAN$_S$ [210]	TIP-2021	1.06	32.16	0.8942	28.63	0.7810	27.58	0.7363	26.10	0.7862	**30.59**	0.9084
	SMSR [56]	CVPR-2021	1.00	32.12	0.8932	28.55	0.7808	27.55	0.7351	26.11	0.7868	30.54	**0.9085**
	MLE^{2}A^{2}U-Net$_L$		2.03	**32.41**	**0.8983**	**28.84**	**0.7881**	**27.72**	**0.7404**	**26.41**	**0.7961**	**30.84**	**0.9099**

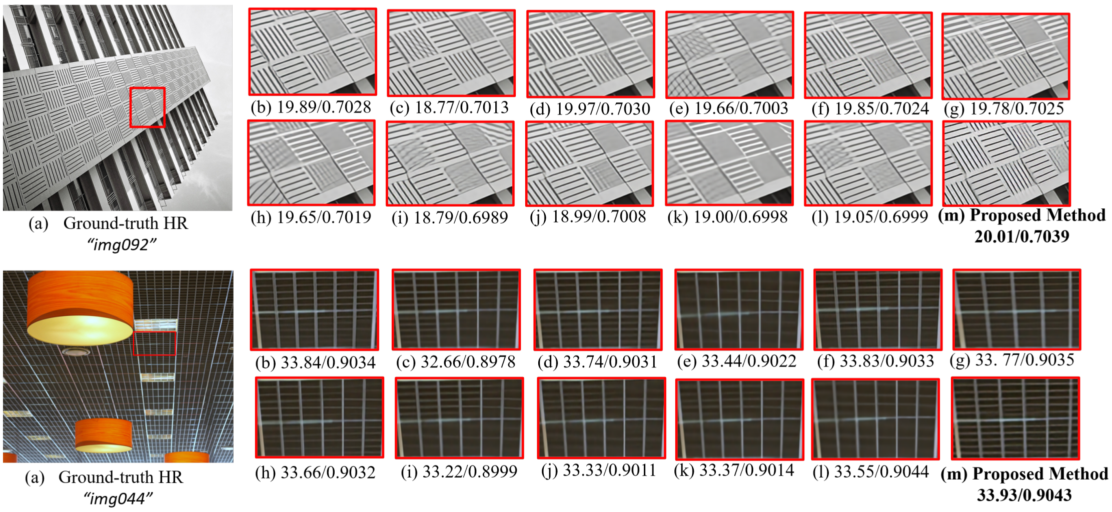


Figure 3.16: Subjective evaluation for ×4 upscaling on images from Urban100 [4] dataset. (a) Ground truth HR image, (b) CSNLN [39], (c) DBPN [40], (d) DRLN [41], (e) EDSR [26], (f) HAN [42], (g) RFA-Net [43], (h) RCAN [44], (i) RDN [45], (j) SRFBN [46], (k) USRNet [47], (l) SRGAT [48] and (m) **MLE^{2}A^{2}U-Net (Proposed Method)**.

Comparison with State-of-the-art Methods

For evaluating the effectiveness of the proposed MLE^{2}A^{2}U-Net model, several state-of-the-art methods are compared in terms of quantitative and qualitative evaluation and local attribution maps [51].

Quantitative Evaluation

We compare the proposed MLE^{2}A^{2}U-Net with fourteen state-of-the-art methods for SISR. For quantitative purpose, we have compared the PSNR and SSIM values of different methods for scales ×2 and ×4. From Table 3.11, it is clear that our model, **with 8.2 Million parameters,** yields the best performance, with highest PSNR and SSIM on almost all the datasets. It is worth noting, that the proposed model clearly outperforms heavy models (with about 40 Million parameters) EDSR [26], OISR-RK3 [208], by a large margin of about 0.3 dB and 0.2 dB for Set5 and Set14, respectively for × 4 up-scaling. As reported in Table 3.11, the improvement margins of PSNR when compared with SAN [52], for ×2, ×4 are beyond 0.2 dB, 0.3 dB, respectively for challenging dataset of Manga109[49] and Urban100 [4]. Both these datasets, consists of highly complicated and repetitive bands, making it quite challenging for task of super-resolution. Furthermore, the proposed MLE^{2}A^{2}U-Net shows an improvement of 0.2 dB when compared with recently proposed TSAN [209] and DeFiAN [210] on Set5 [1] dataset for ×4. This proves the accuracy of our model, in better learning the relationship among the training patches in feature space.

To further prove the effectiveness of the proposed model, we have compared our lightweight MLE^{2}A^{2}U-Net$_L$ with 8 state-of-the-art lightweight SR methods as shown in Table 3.12.

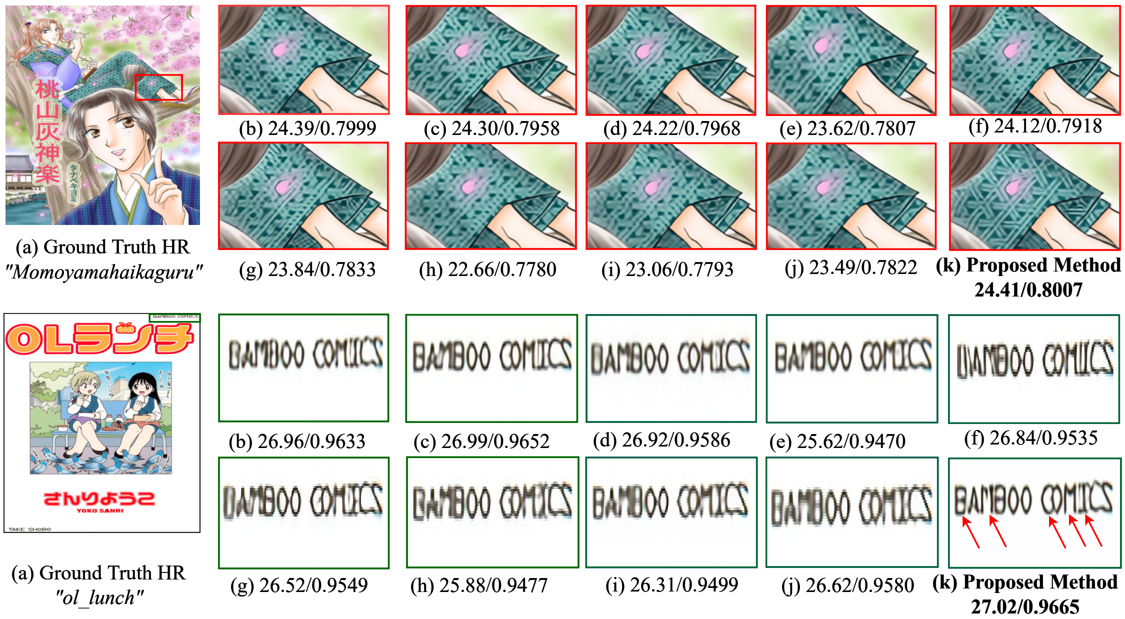


Figure 3.17: Subjective evaluation for ×4 upscaling on *"Momoyamahaikagura"* and *"Ollunch"* from Manga109 [49] dataset. (a) Ground truth HR image, (b) CSNLN [39], (c) DRLN [41], (d) HAN [42], (e) DBPN [40], (f) RFA-Net [43], (g) RCAN [44], (h) RDN [45], (i) SRFBN [46], (j) SRGAT [48] and (k) **Proposed Method**.

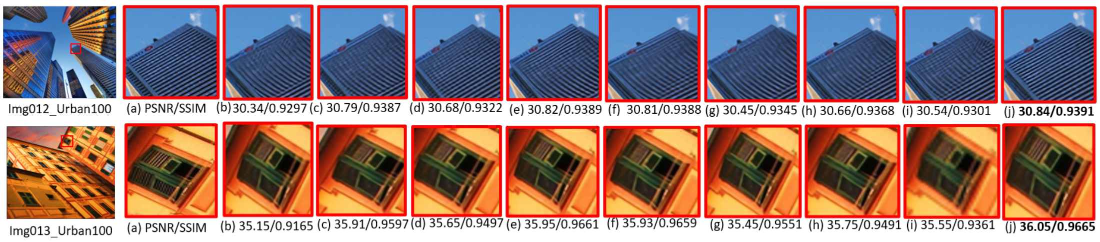


Figure 3.18: Subjective evaluation for ×2 upscaling on *"img012"* and *"img013"* from Urban100 [4] dataset. (a) Ground truth HR image, (b) CARN [50], (c) CSNLN [39], (d) DBPN [40], (e) DRLN [41], (f) HAN [42], (g) MSRN [34], (h) RCAN [44], (i) RDN [45] and (j) **MLE^{2}A^{2}U-Net (Ours)**. (Better view in zoom)

Our lightweight model **with 2.03 Million parameters**, clearly outperforms the recently proposed DeFian [210] by 0.16 dB and 0.3 dB on Set5 for × 2 and × 4 scaling, respectively.

Qualitative Evaluation

We further assess qualitatively the performance of the proposed MLE^{2}A^{2}U-Net on some challenging images from the benchmark datasets for × 4 and × 2 as shown in Figure 3.16, 3.17 and 3.18. For better understanding, we have zoomed in the details of the considered example and labelled the PSNR/SSIM under each image patch. Our proposed method is more accurate at reconstructing most of the parallel straight lines and grid patterns, such as the highlighted stripes of the wall and the rectangular grid as shown in Figure 3.16. The compared methods are hardly able to reconstruct the right patterns of these lines and suffer from unpleasant blurring artifacts. On the other hand, the

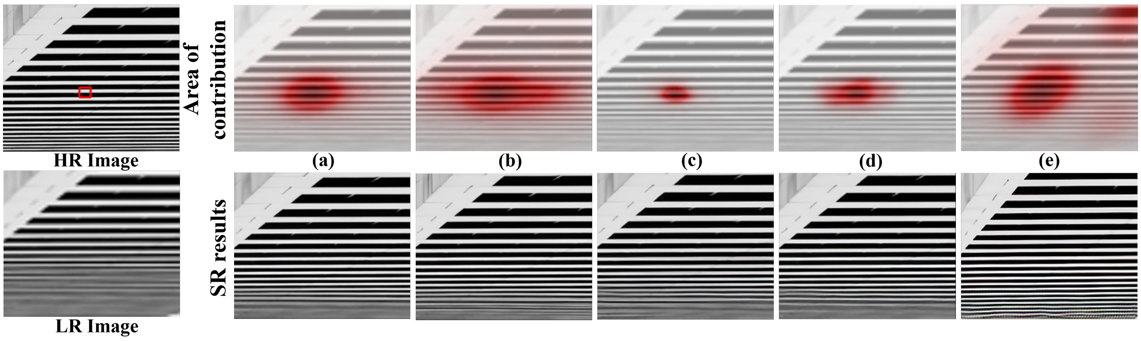


Figure 3.19: Demonstration of the Local Attribution Maps (LAM) for SR network interpretation [51]. Here, (a) RCAN [44], (b) SAN [52], (c) CARN [50], (d) EDSR [26] and (e) Proposed method.

super-resolved results of the proposed MLE^{2}A^{2}U-Net contain visually pleasant patterns, with more high-frequency details, such as textures and edges. This phenomenon of better reconstruction is further highlighted for images with complicated and repetitive bands, like for examples in Figure 3.17, our proposed method is better at reconstructing the texts and characters (*the characters marked with red color arrows are better generated by our proposed model*). Besides, as shown in Figure 3.18, the visual comparisons for scale ×2 demonstrate that our approach is best at reconstructing the high-frequency image details with high fidelity to the ground-truth HR image. This is mainly, because we attempt to exploit information among patches by utilising feature correlations, and thus parallely recover the structured information and semantic details.

Results for Local Attribution Maps

Gu *et al.* [51], recently proposed the concept of local attribution maps (LAM) for SR, to find the contribution of input pixels, that strongly influence the SR results. LAM, aims to highlight the most important pixels, and for the same input LR patch, if local attribution maps for the SR image covers a wide range of pixels, it indicates that more information has been extracted and used for reconstruction. It can be concluded from Figure 3.19 that network like CARN [50], covers only limited range of pixels, on account of its limited receptive field. EDSR [26] and RCAN [44] covers a wide range of pixels and exhibits better results than [50]. Further, it is worth noting that some of the networks like SAN [52] may cover a broader range of pixels, but still lead to the generation of wrong textures in the final reconstructed SR image. While our model, besides covering large range of pixels is also capable of generating crisp edges. This proves the effectiveness of the modules incorporated in our network for utilising global information to assist SR.

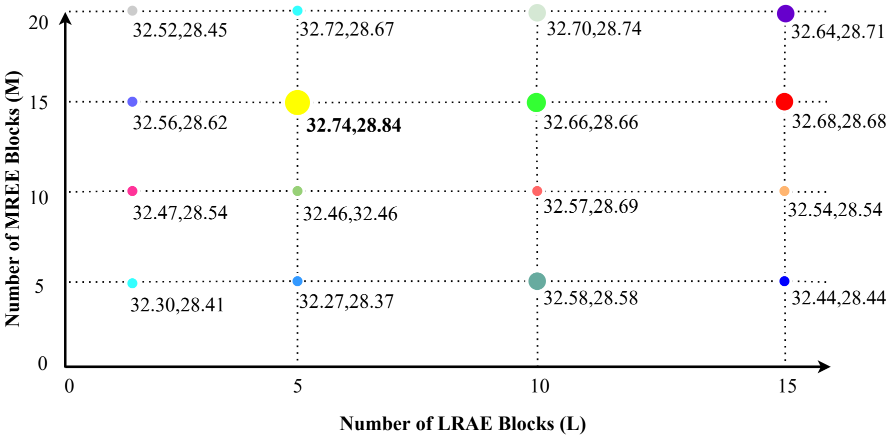


Figure 3.20: Comparison of MLE^{2}A^{2}U-Net on different model sizes. Average PSNR on Set5 and Set14 datasets for × 4 scale factor. Here (x,y) denotes the corresponding PSNR values for Set5 [1] and Set14 [2].

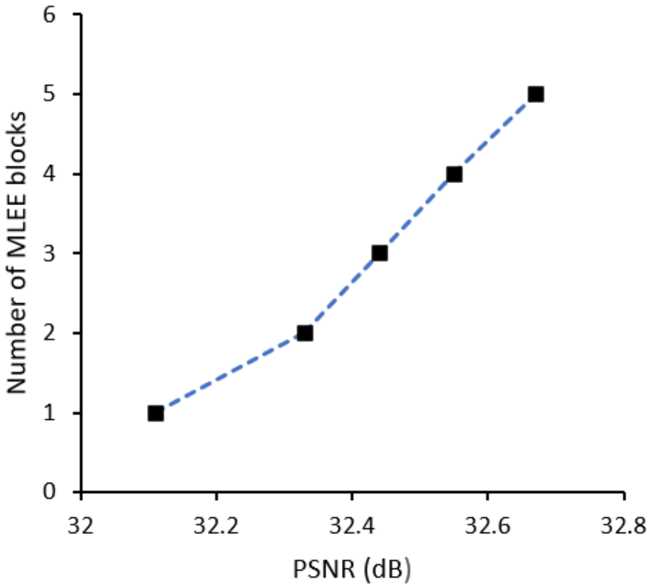


Figure 3.21: Comparison of PSNR vs. the number of MLEE blocks for Set5 dataset.

3.2.3 Ablation Study

In this subsection, the effects and the contributions of different components in the proposed model are analysed by conducting a series of experiments, including the effect of model size, effect of incorporating MRFE, MLEE, LRAE, effect of AAU, feature refinement module and the effect of pooling operations. For all experiments, the hyper-parameters of the models in this subsection are set as $L = 5$, $E = 5$ and $M = 15$ and have been trained on DIV2K datasets.

Analysis of Model Size

In this work, several hyper-parameters play an important role in designing our full model; including the number of LRAE blocks, MLEE blocks, the number of MRFE blocks, etc. We, then conduct the ablation analysis for different settings of hyper-parameters as shown in Figure 3.20 and Figure 3.21. From Figure 3.20, showing the results of PSNR performance (`represented using different colors on Set5 and Set14 for` ×4) versus the number of LRAE and MRFE blocks, we find that as the model size increases (increase in number of LRAE, MRFE blocks), the performance of our network improves, but we find increasing it beyond a particular range, limits the performance of our SR network. It is evident from Figure 3.20 that network with 5 LRAE and 15 MRFE blocks exhibits PSNR gain of about 0.1 dB when compared to deeper model 15 LRAE and 20 MRFE blocks.

Effect of Different Modules

Next, we conduct an ablation study for each component of our proposed MLE^{2}A^{2}U-Net, including the multi-level edge embedding module (MLEE), attention block (AB) and aggregated attentive up-sampler block (AAU). All these variants have been tested on Set5 dataset and the detailed performance has been shown in Table 3.13. Further, to focus on the high-frequency details of the features, we visualize the intermediate feature maps as shown in Figure 3.22. The 8 different feature maps corresponds to the 8 different settings of Table 3.13, by removing or adding different modules. It is clear from the Figure 3.22 that undesirable discrepancies in detailed regions have been reduced greatly from (a) → (h).

We further construct a baseline model by removing all the main modules in the network. We can check from Table 3.13 that the baseline model reaches 31.56 dB PSNR on Set5 dataset for ×4. Meanwhile, results in the first four columns of Table 3.13 demonstrates the efficiency of each module, exhibiting significant improvement over the Baseline. As an example, the effect of adding AAU block is visible from Figure 3.22 (c) (*showing the feature maps of each combination*), that shows more details when compared to (a). When

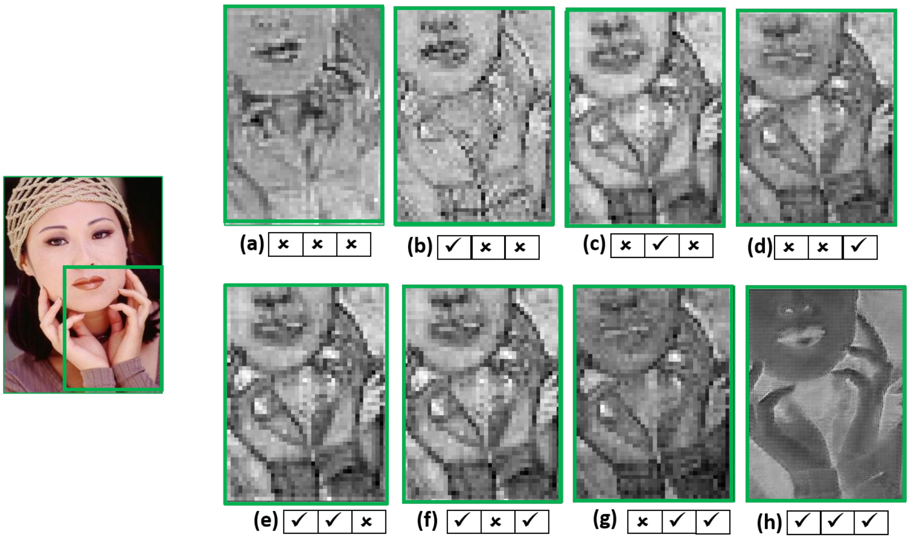


Figure 3.22: Visualization of feature maps for different combinations given in Table 3.13.

we integrate the two modules (MLEE and AAU), the ability of the network for detail preservation is improved by about 1 dB from the baseline model. The major reason lies in the removal of excessive redundant information and collection of large contextual information via increased receptive field. Furthermore, when all the three modules are used, the performance is significantly improved from 31.56 dB to 32.68 dB; showing the effectiveness of the proposed architecture.

Table 3.13: Ablation study on Set5 dataset for × 4 scale factor with different modules.

MLEE AAU AB	× × ×	✓ × ×	× ✓ ×	✓ ✓ ×	✓ × ✓	× ✓ ✓	✓ ✓ ✓
PSNR (dB)	31.56	32.10	31.97	32.56	32.29	32.44	**32.68**

Effect of AAU

After confirming the validity of the proposed components, we next compare the performance of our network for various up-sampling techniques. Our baseline network is the proposed model employing AAU for up-sampling. Next, we train our network separately, adopting bicubic, deconvolution, pixel-shuffle and bilinear for up-sampling in place of our proposed AAU block. It is clear from the Table 3.14 that the baseline method using our proposed AAU module is better than the Pixel Shuffle, Bilinear and Deconvolution operation in terms of PSNR by approximately 0.2, 0.3 and 0.4 dB, respectively for × 4 up-scaling on Set5 dataset.

Table 3.14: Effect of different upsampling techniques. Average PSNR on different datasets for × 4 scaling.

Methods	Set5	Set14	BSD100	Urban100
Bicubic	32.46	28.77	27.66	26.79
Deconvolution	32.44	28.78	27.63	26.75
Pixel-Shuffle	32.69	28.90	27.79	26.88
Bilinear	32.52	28.86	27.77	26.66
Proposed Method	**32.80**	**28.97**	**27.85**	**27.05**

Table 3.15: Effect of feature fusion strategy for ×4 on Set5.

Method	PSNR/SSIM
Concat	32.15/0.8993
Addition	32.23/0.8996
Feature Refinement Module	**32.66/0.9004**

Table 3.16: Effect of pooling operation in the Attention Block (AB). GAP denotes Global Average Pooling and GMP denotes Global Max Pooling.

Model	Set5	Set14	B100
AB + GAP	32.46/0.8958	28.69/0.7884	27.58/0.7316
AB + GMP	32.52/0.8974	28.85/0.7895	27.71/0.7327
AB + GMP + GAP	**32.69/0.9000**	**28.93/0.7900**	**27.82/0.7338**

Table 3.17: **Different Locations:** Non-local Blocks are inserted at different positions in our proposed model for Set5 dataset.

Model	PSNR /SSIM
Without NLB	31.95/0.8984
After every LRAE	32.21/0.9004
Parallel with every LRAE	32.05/0.8999
Proposed Method	**32.57/0.9007**

Effect of Feature Refinement Module

We show the importance of feature refinement module by comparing it with other commonly used fusion strategies in literature i.e. feature addition and feature concatenation. From the Table 3.15, it is found that our feature refinement module obtains the best results on all the datasets, by showing an improvement of nearly 0.4 dB for Set5 dataset when compared with other fusion strategies, and thus demonstrates the effectiveness of incorporating it in our network. This comparison proves the fact that since, both the addition and subtraction operations of MLEE module contains complementary information; and fusion of these features using trivial concatenation or element wise feature addition may overlook the redundant information. Hence, we need to introduce some sort of attentive mechanism in our combination framework, justifying the importance of feature refinement module.

Effect of Pooling operation in Attention Block

In order to highlight the effect of different pooling operations in the attention block, we compared three different settings as shown in Table 3.16. It should be noticed, that while employing both GAP and GMP in the attention block, our accuracy results surpasses the other settings by a huge margin, proving the efficacy of the proposed attention block.

Which Stage do We Need to add Non-Local Block?

As discussed in [126], the position of non-local block in the network plays an important role in capturing the precise spatial information. We also conducted series of experiments, as tabulated in Table 3.17, for placing the non-local block at different positions in the proposed architecture. The comparisons have been performed for Set5, Set14 and BSD100 at scale ×4. `Without NLB` refers to the proposed architecture without any non-local block and exhibits the lowest performance in terms of PSNR/SSIM. Further, it is clearly visible that placing the non-local blocks either in parallel or in series, after every LRAE module, does not contribute much in improving the performance. One possible explanation is that, non-local blocks are not that much effective in linking the long-range dependency information when incorporated in such a way. It is evident that plugging one non-local block at the beginning of the LRAE module and one before the up-sampling module (or after the last LRAE module) gives the best performance.

3.3 Summary

This chapter addresses the problem of frequency extraction in super-resolution without any requirement of explicit prior information. In the first solution (Section 3.1), we propose an effective framework, novel multi-level bi-cubic up-sampler network (MBUp-Net) for modelling the process of super-resolution. Specifically, a novel content aware feature difference (CAFD) block is designed to effectively encode multi-scale contextual information and to extract high-frequency details. In each CAFD, multi-level attention (MLA) blocks enable full utilization of the multi-scale features by allowing only more informative ones to pass further. Our proposed up-sampling strategy ensures superior results by removing the artefacts and aliasing introduced by other layers to a great extent. In our second solution 2 (Section 3.2), we propose an edge embedding with attentive up-sampler (MLE^{2}A^{2}U-Net) network for single image super-resolution aiming to preserve the precise spatial details by capturing long-range information without any prior information. Particularly, to extract the multi-scale high-frequency information for generating images with high visual quality, a novel multi-level edge embedding module with stacked multi-receptive field extractor block has been proposed. Furthermore, a novel aggregated attentive up-sampling module is proposed to effectively merge the attentive features from different layers.

However, there is more requirement of computationally intensive architectures for saving the memory consumption. A detailed discussion is given in the next chapter.

Chapter 4

A Novel Lightweight Approach for Image Super-resolution

Though the existing super-resolution networks yield photo-realistic outputs, but the models are quite heavy with millions of parameters [68, 216] and the inference through such a complex network demands billions of floating point operations (FLOPs). Consequently, it results in larger computational cost, more power consumption and limited flexibility of the overall network while deploying on hardware. The best possible solution for this is to come up with some lightweight networks that offer a trade-off between accuracy and speed. In order to balance these issues of quality and complexity, in this chapter we proposed two different lightweight solutions with relatively less computational complexity and efficient super-resolution results. The proposed two solutions are:

1. MSAR-Net: Multi-scale Attention based Light-Weight Image Super-Resolution.

2. Con-Net: A Consolidated Light-Weight Image Restoration Network.

These solutions are explained in detail in the following sections.

4.1 MSAR-Net: Multi-scale Attention based Light-Weight Image Super-Resolution

Despite considerable improvement brought up by the existing SR techniques [36, 68, 216, 217, 186], a computationally intensive method for consolidating the feature representation and edge-enhancing capability in a single network with less computational burden needs to be exploited. Attention mechanism that focuses on the correlation of the features either spatially or channel-wise has shown promising results in the field of image super-resolution, but at the cost of large number of parameters [36] and [176]. Owing to the effectiveness of attention blocks, our proposed method embedded a novel lightweight attention module into the proposed multi-scale residual block. Additionally, to effectively super-resolve an image it is necessary to focus on the edges in an image, thus a novel up and down sampling projection block has been used after each multi-scale attention residual block for collecting the high-frequency information. *But still a natural question arises, is it possible to have an effective consolidated framework capable of promoting the understanding of image contents, with less number of parameters?* It is obvious that increasing the

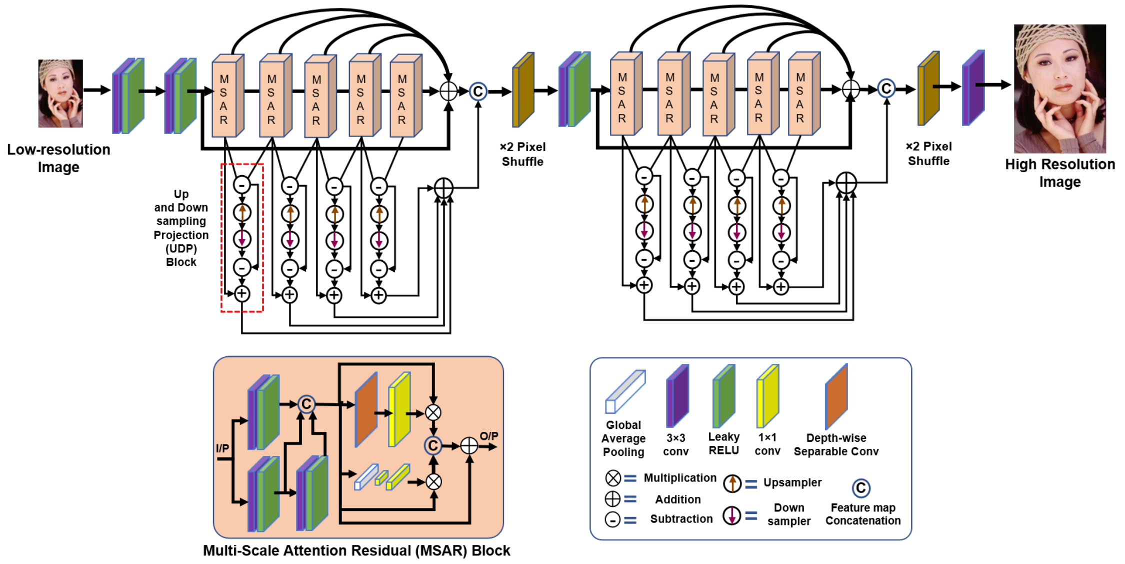


Figure 4.1: The proposed architecture of the network for image super-resolution.

layer depth will cause vanishing gradient and computational burden problems. Hence, to promote better reconstruction performance with fewer number of parameters, we propose a Lightweight MSAR-Net for exploiting both the feature and edge information in single network. The main contributions of our work are listed as below:

1. We propose a progressive multi-scale network to sequentially explore the hierarchical information with fewer parameters. This lightweight architecture makes it possible to handle the image features efficiently for high quality image restoration.

2. We propose Multi-scale attention residual (MSAR) blocks for adaptively capturing the multi-scale correlations among the features and an up and down sampling projection (UDP) block for edge refinement of the extracted multi-scale features.

4.1.1 Proposed Method

The overall pipeline of the proposed network shown in Figure 4.1 is composed of three modules: (1) Feature Extraction Block (FEB) that processes a LR input image to collect the robust features, (2) stack of multi-scale attention residual (MSAR) blocks for performing the non-linear mapping after exploring the relationship between features, (3) up and down projection (UDP) block for performing the edge refinement of the extracted features. The feature extraction module consists of two 3×3 convolution layers which are used to extract the features by collecting the activations of the inputs and generate LR feature maps. Eq. (4.1) defines the basic function of FEB.

$$H_0 = \psi_{FEB}(I_{LR}), \tag{4.1}$$

where, ψ_{FEB} denotes the function of feature extraction block, H_0 represents the extracted features, and I_{LR} denotes the input LR image. After extracting the features, a stack of

MSAR blocks are used to further explore the features non-linearly and finally map the LR features to SR using pixel-shuffle layer. Each MSAR block consists of multi-scale attention block and the features further extracted are fused by global learning. This process for k blocks can be formulated as H_k, where H_k denotes the output of k^{th} MSAR block. After passing through all MSAR's and UDP's, the concatenated features will pass through the up-sampling block which will up-sample the feature maps through pixel-shuffling. The reconstructed feature maps consisting of refined upsampled features is defined as,

$$H_{rec} = \psi_{UDP}(\psi_{MSAR}(H_0)), \tag{4.2}$$

where, ψ_{UDP} and ψ_{MSAR}, denote the function of the UDP block and MSAR block, respectively. The final high-quality SR image we seek is defined as,

$$I_{SR} = f_{3\times3}H_{rec}, \tag{4.3}$$

where, $f_{3\times3}$ denotes the function of convolution layer with kernel size 3. We have used pixel-shuffle [181] as our upsampling layer. It is worth mentioning, that processing the information at different scales and subsequently aggregating helps in the abstraction of features for the next stage, thus making the model capable of extracting a variety of information. Further, the residual connections are been used in the model that helps in eliminating the problem of vanishing gradients, thus ultimately stabilising the training procedure. In the next subsection, we discuss the two main components of the proposed MSAR-Net, Multi-scale Attention Residual Block and Up and Down projection block in detail.

Multi-Scale Attention Residual Block

An overall progressive multi-scale model for better feature-correlation, while moving deeper in the network has been proposed. Different from the other residual and inception blocks proposed in existing literature [30, 32], we made an attempt to increase the receptive field for better feature extraction. For allocation of the available resources toward the more informative contents in the image, we have used the concept of attention inspired from [36]. To further increase the network capability on learning more important features, a spatial attention unit and channel attention unit has been designed. The multi-scale features obtained by passing the information from parallel convolution layers of size 1, 3, 5[1] are concatenated as:

$$\psi_{res} = [f_{1\times1}, f_{3\times3}, f_{5\times5}], \tag{4.4}$$

where, $[\cdot]$ and ψ_{res} denote the concatenation and multi-scale features, respectively. Each convolution layer from the proposed MSAR block consists of 32 filters.

[1] $f_{5\times5}$ in Figure 4.1 has been implemented using two $f_{3\times3}$ convolutions [218].

For further contextual information, not gathered by the local receptive field, global average pooling [36] has been used. We have opted to extract the channel characteristics using a sigmoid function for enhancing the non-linear interactions between the channels. The channel attention block is defined below:

$$x_c = V_{GP}(\psi_{resc}) = \frac{1}{h \times w} \sum_{i=1}^{h} \sum_{j=1}^{w} \psi_{resc}(i, j), \tag{4.5}$$

$$\psi_{CA} = [\delta(f_{1\times1}(\lambda(f_{1\times1}(x_c))))] \times \psi_{resc}, \tag{4.6}$$

Here, $V_{GP}(\cdot)$ represents the average global pooling operation to take into account the channel-wise spatial information. $\delta(\cdot)$ represents the sigmoid function and $\lambda(\cdot)$ represents the LeakyReLU activation function, respectively. ψ_{CA} represents the output of channel attention block. ψ_{resc} represents the c^{th} channel feature map and x_c represents the statistics obtained by shrinking ψ_{res} spatially. To emphasize on the non-linear activations between the channels, the obtained features are first downsampled channel-wise via convolution layer of 32 filters and kernel size of 1 × 1, followed by channel upsampling performed via a convolution layer of 1 × 1 with 64 filters. For rescaling the input, the obtained channel statistic is multiplied by the feature map in the c^{th} channel as shown in Eq. (4.6) to scale the important channel features.

To modulate the features locally, spatial attention unit has been used, which is defined as,

$$\psi_{SA} = (f_{1\times1}(\phi(\psi_{res}))), \tag{4.7}$$

$$\psi_{cat} = [\psi_{SA} \times \psi_{res}, \psi_{CA} \times \psi_{res}]\psi_{MSAR} = \psi_{res} + \psi_{cat}, \tag{4.8}$$

Here, ψ_{SA} represents the output of spatial attention block. ψ_{MSAR} represents the final output of the MSAR block. $\phi(\cdot)$ represents the depth wise convolution with filter size 3×3. ψ_{cat} represents the concatenated attention features. The proposed attention blocks are capable of exploiting the inter and intra channel information, where the use of Depth-wise convolution further helps in generation of different 2D spatial attention maps for each channel. The obtained maps are then passed through a convolution layer with 64 filters for better refinement. To utilise the benefits of both the blocks simultaneously, we have combined them through concatenation.

Up and Down projection Block

After obtaining the refined features from MSAR block, we ought to increase the content of high-frequency information in the image by using the proposed UDP block as shown in Figure 4.1. The overall operation of UDP block has been summarized in Eqs. (4.9) and (4.10). Firstly, the difference of the consecutive multi-scale feature maps of MRFE is evaluated. It focuses on the high-frequency information, then the subtracted features are passed through an upsampling layer `Conv2DTranspose with stride 2`. The upsampled features are converted back to the LR space by using a `convolution layer of stride 2`. The final subtraction operation, outputting ζ_n in Eq. (4.9) helps in removing the redundant information. The addition operation shown in Eq. (4.10), with ψ_{UDP} output, extracts the relevant features required for the reconstruction of sharp image and consequently boosts the multi-scale features. Further, the features from all the UDP blocks are added together for better gradient propagation. Finally, the concatenation of all the residual features and edge features have been performed for full exploitation of the multi-scale edge features.

$$\begin{aligned} &\zeta_n = \Delta_n - i_n \\ &\Delta_n = \psi_{res_n} - \psi_{res_{n-1}} \\ &i_{n=}\downarrow_2(\uparrow_2(\Delta_n)), \end{aligned} \tag{4.9}$$

$$\psi_{UDP} = \zeta_n + \psi_{res_n} \tag{4.10}$$

Here, the $\uparrow_2$ and $\downarrow_2$ represent the upsampling and downsampling operations by ×2, respectively.

4.1.2 Experimental Analysis

In experiments, we have trained our model using DIV2K dataset [36]. The training LR images are generated by downsampling the HR image through bicubic interpolation. Our model has been further evaluated on Set5 [1], Set14 [2], BSD100 [3], Urban100 [4] and Manga109 [49]. All these datasets further consist of a variety of scenarios, thus completely validating the performance of the proposed method. The proposed network has been trained for 500 epochs and implemented in Tensorflow 2.0 deep learning framework. The model has been optimized using ADAM with $\beta 1$, $\beta 2$ and ϵ set to 0.9, 0.999 and 10^{-8}, respectively. During training, learning rate is set to 10^{-4} and halved every 10^4 iterations. The model has been trained on NVIDIA RTX 2080Ti GPU using mean absolute error function between the ground truth I_{HR} and reconstructed image I_{SR}.

Table 4.1: Quantitative results of the proposed and state-of-the-art algorithms. PSNR/SSIM for scale factor ×2, ×3, ×4 on datasets Set5 [1], Set14 [2], BSD100 [3], Urban100 [4] and Manga109 [49].

Dataset	**Scale**	**Set5**	**Set14**	**BSD100**	**Urban100**	**Manga109**
RFDN [43]	×2	38.05/0.9606	33.68/0.9184	32.16/0.8994	32.12/0.9278	38.88/0.9773
MSICF [219]	×2	37.89/0.9605	33.41/0.9153	32.15/0.8992	31.47/0.9220	-/-
SMSR [56]	×2	38.00/0.9601	33.64/0.9179	32.17/0.8990	32.19/0.9284	38.76/0.9771
HDRN [54]	×2	37.75/0.9590	33.49/0.9150	32.03/0.8980	31.87/0.9250	38.07/0.9770
MIPN [55]	×2	38.12/0.9609	33.73/0.9188	32.25/0.9006	32.42/0.9312	38.88/0.9773
ACNet [220]	×2	38.12/0.9609	33.73/0.9188	32.25/0.9006	32.42/0.9312	38.88/0.9773
CFSRCNN [221]	×2	38.12/0.9609	33.73/0.9188	32.25/0.9006	32.42/0.9312	38.88/0.9773
LESRCNN [222]	×2	38.12/0.9609	33.73/0.9188	32.25/0.9006	32.42/0.9312	38.88/0.9773
MSAR-Net (Ours)	×2	**38.22/0.9616**	**33.79/0.9189**	**32.27/0.9108**	**33.46/0.9322**	**33.46/0.9322**
RFDN [43]	×3	34.41/0.9273	30.34/0.8420	29.09/0.8050	28.21/0.8525	33.67/0.9449
MSICF [219]	×3	34.24/0.9266	30.09/0.8371	29.01/0.8024	27.69/0.8411	-/-
SMSR [56]	×3	34.40/0.9270	30.33/0.8412	29.10/0.8050	28.25/0.8536	33.68/0.9445
HDRN [54]	×3	34.24/0.924	30.23/0.840	28.96/0.804	27.93/0.849	33.17/0.942
MIPN [55]	×3	34.53/0.9280	30.43/0.8440	29.15/0.8061	28.38/0.8573	33.86/0.9460
ACNet [220]	×3	38.12/0.9609	33.73/0.9188	32.25/0.9006	32.42/0.9312	38.88/0.9773
CFSRCNN [221]	×3	38.12/0.9609	33.73/0.9188	32.25/0.9006	32.42/0.9312	38.88/0.9773
LESRCNN [222]	×3	38.12/0.9609	33.73/0.9188	32.25/0.9006	32.42/0.9312	38.88/0.9773
MSAR-Net (Ours)	×3	**34.59/0.9285**	**30.53/0.8446**	**29.35/0.8044**	**28.44/0.8586**	**33.98/0.9470**
RFDN [43]	×4	32.24/0.9266	28.61/0.7819	27.57/0.7360	26.11/0.7858	30.58/0.9089
MSICF [219]	×4	31.91/0.8923	28.35/0.7751	27.46/0.7308	25.64/0.7692	-/-
SMSR [56]	×4	32.12/0.8932	28.55/0.7808	27.55/0.7351	26.11/0.7868	30.54/0.9085
HDRN [54]	×4	32.23/0.8960	28.58/0.7810	27.53/0.7330	26.09/0.7870	30.43/0.9080
MIPN [55]	×4	**32.31/0.8971**	28.65/0.7832	27.61/0.7375	26.23/0.7906	**30.67/0.9107**
ACNet [220]	×4	38.12/0.9609	33.73/0.9188	32.25/0.9006	32.42/0.9312	38.88/0.9773
CFSRCNN [221]	×4	38.12/0.9609	33.73/0.9188	32.25/0.9006	32.42/0.9312	38.88/0.9773
LESRCNN [222]	×4	38.12/0.9609	33.73/0.9188	32.25/0.9006	32.42/0.9312	38.88/0.9773
MSAR-Net (Ours)	×4	32.29/0.8989	**28.67/0.7841**	**27.95/0.7410**	**26.25/0.7907**	30.66/0.9100

Results and Discussion

We have compared the performance of the proposed network with popular CNN-based SR methods including, SRCNN [22], VDSR [191], DRCN [111], MemNet [24], SRFBN [185], IDN [25], CARN [28], MSRN [32], IMDN [53], RFDN [43], MSICF [219], MIPN [55] and SMSR [56]. We have quantitatively evaluated the performance of the proposed MSAR-Net for PSNR and SSIM metrics and is given in Table 4.1. The proposed method clearly outperforms the recent MIPN [55], SMSR [56] on PSNR and SSIM for ×2, ×3, ×4 scaling factor. It is clear from the Table 4.1 that the proposed method performs stably well on all the datasets when compared with the existing methods.

Furthermore, referring to Images, Urban100img_092, Urban100img_093, Urban100img_011, Urban100img_046, Urban100img_033, Urban100img_076, Urban100img_072 shown in Figure 4.2, 4.3 and 4.4, from Urban100 dataset, it is clearly visible that the compared existing SR methods are unable to capture clean, crisp edges and subsequently, add undesirable textures. Moreover, we find that our proposed MSAR-Net has clear advantage over the recently proposed MIPN [55], HDRN [54] and SMSR [56] methods in preserving the image edges and thus consequently, generating realistic textures in the reconstructed image. From the above-detailed analysis, the capability of the proposed network for synthesizing better textures and structures is proved for image super-resolution.

4.1.3 Ablation Study

We have further investigated the performance of the proposed network with different numbers of MSAR blocks. As clear from Table 4.2, both the performance metrics are incremental with an increase in the number of MSAR blocks. But, the increase in the number of MSARs beyond 5, gives a negligible improvement in the performance of the proposed network. Hence, we have used 5 MSAR blocks to ensure better feature learning and obtain efficient results. Further, to know the importance of the proposed UDP block, we have analysed the performance of the proposed network with and without UDP block. Table 4.2 witnessed the improvement in the performance of the proposed network with UDP block.

Demonstration of the effect of channel attention (CA) block and spatial attention (SA) block in the proposed MSAR-Net has been set-up with four scenarios. Table 4.3 presents a detailed analysis and indicates the PSNR/SSIM values for ×4 scaling factors on Set5 and Set14 testing datasets. It is clearly visible that the proposed channel and spatial attention blocks leads to the significant improvement for the considered scale, indicating the importance of paying attention to the channel-wise and spatial features.

Table 4.2: Analysis of the number of MSAR blocks and the importance of UDPB in the proposed network for × 4 on Set5.

# MSAR	**PSNR**	**SSIM**
with 4	31.96	0.8904
with 5	**32.11**	**0.8985**
with 6	32.12	0.8985

UDPB	**PSNR**
without	28.31
with	**28.51**

Table 4.3: Analysis of the effect of Attention Block on PSNR/SSIM for ×4.

CA block	SA block	Set5 [1]	Set14 [2]
×	×	31.04/0.8432	26.41/0.6724
×	✓	31.44/0.8785	27.68/0.7123
✓	×	31.75/0.8664	27.86/0.7707
✓	✓	**32.11/0.8997**	**28.51/0.7808**

Table 4.4: Average time consumption on different datasets for scale ×4

Datasets	**Set5**		**Set14**	
	Resolution	**time(s)**	**Resolution**	**time(s)**
VDSR [191]	512 × 512	2.2	720 × 516	3.46
LapSRN [23]	512 × 512	4.78	720 × 516	6.29
HDRN [54]	512 × 512	1.66	720 × 516	2.14
Proposed MSAR-Net	512 × 512	**0.78**	720 × 516	**1.86**

Comparisons on Time Complexity

Further, we provide a comparison on model's efficiency in terms of parameters and time complexity on different datasets. Comparably, our method ensures better performance with less number of parameters. As obvious from Table 4.4, that our MSAR-Net shows clear advantages, when compared with other methods in term of speed, thus promoting the efficiency of our proposed SR method in real-world applications. This further ensures the suitability of MSAR-Net for image SR tasks in large resolution.

Besides run-time, for the proposed MSAR-Net method, computational complexity has been evaluated for SR images of size 154 × 154 in terms of parameters and flops. We have compared several lightweight methods VDSR [191], LapSRN [23], DRCN [111], MemNet [24], CARN [28], DRRN [192], IDN [25], SRFBN [185] and CSFRCNN [221] as tabulated in Table 5. It is evident from the table that the proposed MSAR-Net with least number of flops and less number of parameters is quite suitable for real world applications. This proves the efficiency of proposed MSAR-Net in terms of visual quality, computational efficiency and complexity.

Table 4.5: Comparison of the proposed MSAR-net with state-of-the-art light-weight models in terms of parameters and flops on Set5 for ×2 upscaling.

Methods	Parameters	FLOPs
VDSR [191]	0.66M	10.90G
LapSRN [23]	0.43M	18.90G
DRCN [111]	1.77M	29.07G
CARN [28]	1.59M	42.07G
MemNet [24]	0.66M	11.09G
DRRN [192]	**0.29M**	1275.5G
IDN [25]	0.71M	31G
SRFBN [185]	3.63M	22.24G
CSFRCNN [221]	1.20M	11.08G
MSAR-Net	1.48M	**8.82G**

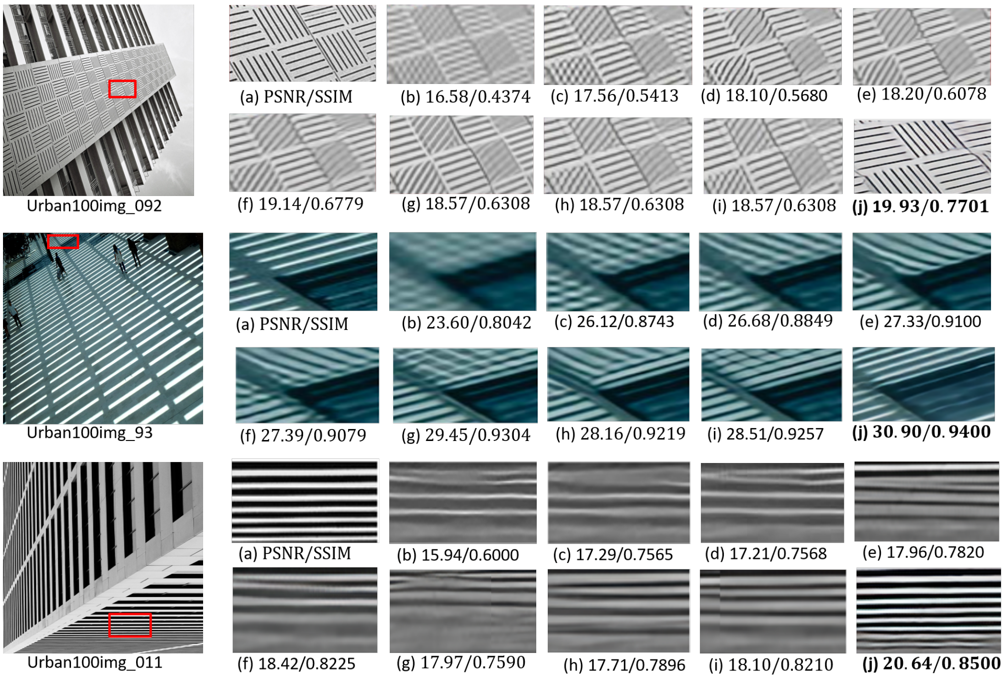


Figure 4.2: Visual comparison on challenging images of Urban100 dataset for ×4 where (a) Represents the cropped ground-truth HR, (b) Bicubic, (c) SRCNN [22], (d) MemNet [24], (e) IDN [25], (f) IMDN [53], (g) HDRN [54], (h) MIPN [55], (i) SMSR [56], (j) Proposed MSAR-Net.

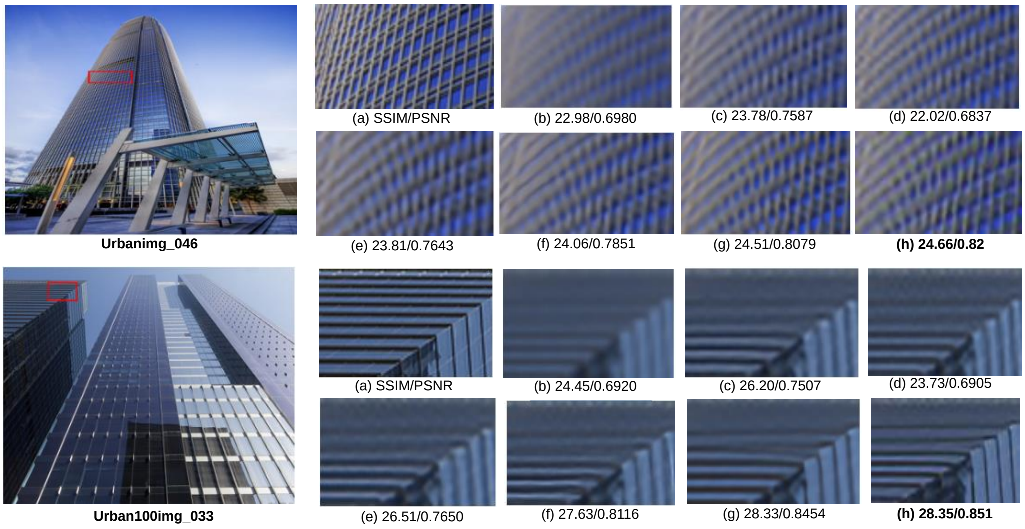


Figure 4.3: Visual comparison for ×3 of Urban100 dataset where, (a) Represents the cropped ground-truth HR, (b) Bicubic, (c) MemNet [24], (d) IDN [25], (e) IMDN [53], (f) HDRN [54] (g) SMSR [56], (h) Proposed MSAR-Net.

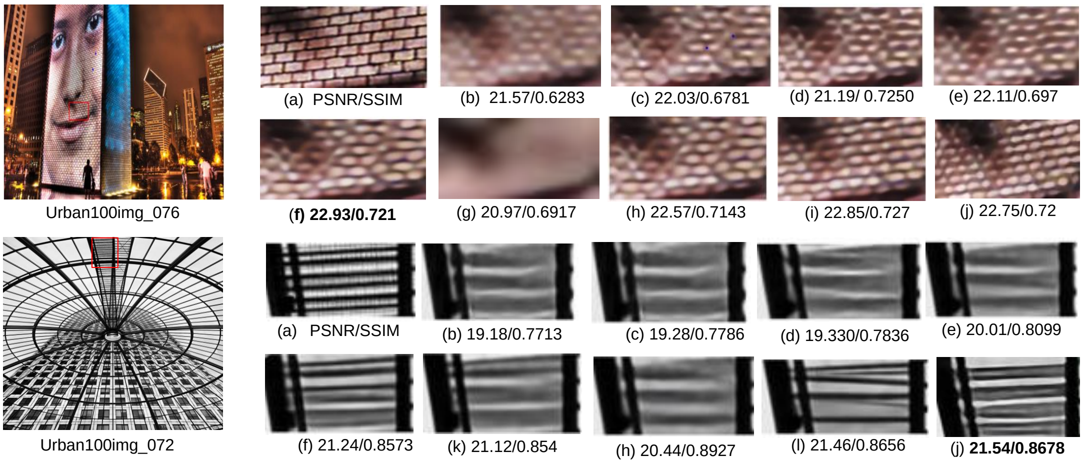


Figure 4.4: Visual comparison for ×2 of Urban100 dataset where, (a) Represents the cropped ground-truth HR, (b) Bicubic, (c) SRCNN [22], (d) MemNet [24], (e) IDN [25], (f) IMDN [53], (g) HDRN [54], (h) MIPN [55], (i) SMSR [56], (j) Proposed MSAR-Net.

4.2 Con-Net: A Consolidated Light-Weight Image Restoration Network

This work is focused on designing efficient modules that offer a better accuracy-speed trade-off in comparison to cascaded convolution layers. Motivated from the understanding, that a restoration network can benefit from the uneven distribution of different types of information in the image and especially, for real-world cases, where the degradation model is highly non-uniform across the spatial position of image, we adopt a consolidated network (Con-Net) for the recovery of the degradation in a fully non-linear manner.

We present the first attempt to design a lightweight restoration network by exploiting the edge and textural details for restoring the image content in areas with complex textures and highly repetitive details. Our proposed Con-Net is suitable for different types of degradations that selectively affect parts of an image. It is composed of two main components - a spatial-degradation aware network (Net$_{SDA}$) and a holistic attention refinement network (Net$_{HAR}$). Net$_{SDA}$ collects information from the entire image for localizing the degradations and extracting the diverse information in the image, which further steers the processing in Net$_{HAR}$ for selectivity and improving the degraded regions by exploiting the relationships in channel and spatial dimensions.

The proposed Net$_{HAR}$ comprises of two degradation-aware blocks - coupled attention module (Co-Attn) and a selective dual-branch merging module (SDBM). Co-Attn utilizes the extracted features from Net$_{SDA}$ for modulating the feature statistics of the intermediate features of Net$_{HAR}$. It basically aims at improving the features by gathering global context from all the clean regions. And, SDBM deploys two parallel branches for fusing the features globally and locally.

On every task, we achieve significant reduction in parameters without any compromise in the visual quality. The exhaustive experiments and detailed ablation study manifest the generalizability of our components on a variety of low-level restoration tasks and investigate its potential breadth. The key contributions of our work are four-fold:

- A lightweight approach capable of generating spatially accurate and contextually enriched features by using Spatial-Degradation Aware and Holistic Attention Refinement networks. These components ease spatially-varying degradations, besides controlling receptive field within an image in a location-adaptive manner.

- A CNN with regularly repeated structure, where the multi-scale edge and detail information are fused in an attentive manner, to improve the network representation ability and finally obtain high-quality image restoration results.

- A new mechanism for merging the features using a selective dual-branch merging module that dynamically combines different types of information and also preserves the original feature information.

- A novel module named channel attentive upsampling (CAU) is designed to efficiently exploit the high-resolution clues during upsampling of the extracted features.

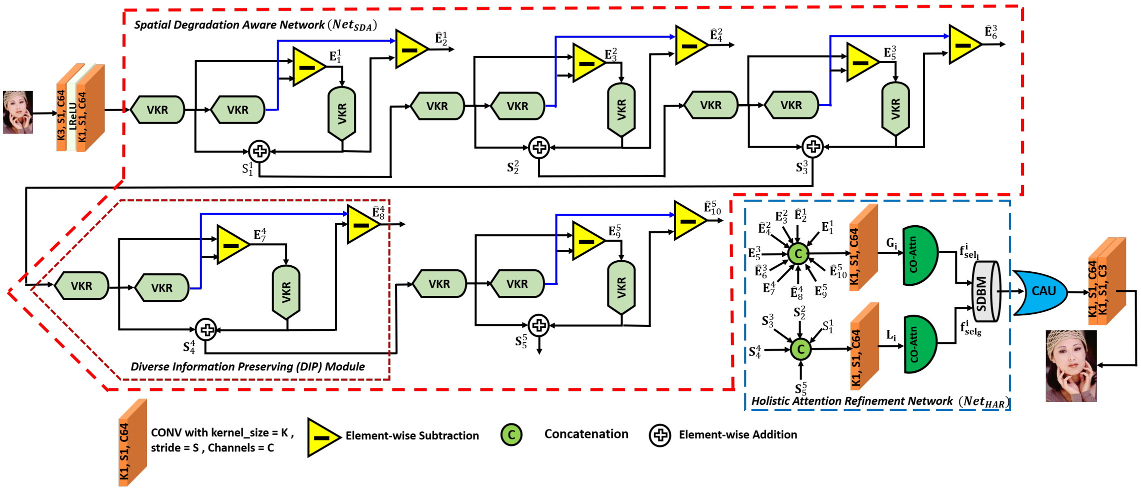


Figure 4.5: Illustration of our proposed network **(Con-Net)**. Con-Net has two main components. (1) Spatial-Degradation Aware Network (2) Holistic Attention Refinement Network.

4.2.1 Proposed Method

An image restoration model ought to solve a few important tasks: (1) Capturing of relevant content from the corresponding degraded regions and the simultaneous preservation of the diverse information inherent in any degraded image. (2) Exploiting semantically-richer and spatially-accurate feature representations in the channel and spatial dimensions. While, spatial-degradation aware network (Net_{SDA}) addresses the former task, we realise the latter through holistic attention refinement network (Net_{HAR}). A schematic layout of the proposed Con-Net network is shown in Figure 4.5. To realize the twofold goals of restoration and enhancement, the refinement of the extracted features in Net_{HAR} is enabled through coupled attention and selective dual-branch merging modules.
In this section, we provide the details of the *Spatial-Degradation Aware network* and *Holistic Attention Refinement network*, the fundamental building blocks of our network.

Spatial-Degradation Aware Network

To maximize the generalizability of our proposed network, it is first necessary to understand the need for preservation of the diverse information (low and high-frequency details) in any restoration network. Generally, the low-frequency information represents the global structure in a given image but with less perception of the minute details. In contrast, the high-frequency information represents the local details of an image, but being more robust to noise [223]. It is further worth emphasizing that there is also a huge difference regarding the distribution of this diverse information in both the low-quality (LQ) and high-quality (HQ) images. For example, in case of denoising, most of the homogeneous regions in both LQ and HQ images share almost the same low-frequency information, while the high-frequency regions are relatively different in both type of images. Moreover, in comparison to the low-frequency, the high-frequency information is more prone to be corruptible, resulting in image degradations [224]. Hence, it is conjectured that restoration of the local details (high-frequency information) is an important step for improving the quality of images. However, individual use of this local information, cannot guarantee efficient recovery of degraded information, hence it is highly expected to use the diverse information together and apply right filter operations so that we can fully exploit their combined merits. Further, we intuit that heavily contaminated regions can benefit a lot from the ability to gather relevant features from the whole image [225].

As discussed earlier, contrary to the existing restoration methods that lack discriminative power and generally consider all types of information equally, we propose a spatial-degradation aware network (Net$_{SDA}$). For rectifying the aforementioned shortcomings, the proposed Net$_{SDA}$ employ vari-kernel residual modules and diverse information processing modules at multiple levels for effective awareness of the degradation in the feature map and thus becoming a suitable candidate for restoring the degraded pixels.

Vari-kernel Residual Module

Most often, propagation of high-frequency information struggles in most of the existing restoration architectures as these networks tend to saturate with low-frequency thereby hindering the effective learning of degraded regions in the image [225]. One possible reason for this could be the use of same kernel across the entire spatial extent of the input features that results in ineffective capture of the high-frequency details. Therefore, in our proposed network to overcome this shortcoming we incorporated multiple Vari-kernel residual (VKRs) modules stacked inside the Diverse Information Processing (DIP) module.

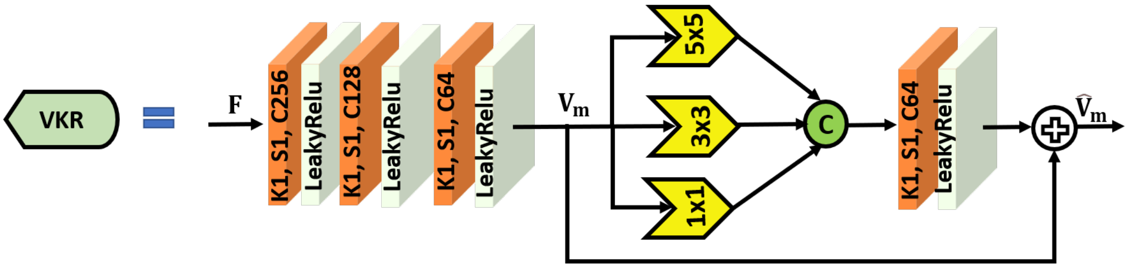


Figure 4.6: Vari-kernel Residual (VKR) module is added in the Diverse Information Processing (DIP) module to facilitate the easy flow of diverse information.

Moreover, larger receptive field helps to generate fine-level feature maps and facilitates the reconstruction of the corrupted pixels in the contaminated images. In large contrast to the high-level computer vision problems such as classification, and object detection which obtain large receptive field by successfully downsampling the feature maps with max-pooling, image restoration tasks needs finer pixel details [216] that are bit hard to achieve from highly downsampled features. And, especially for tasks like super-resolution and denoising that require high-frequency operations these flaws are altogether more exacerbated. Some of the other solutions proposed in this direction either consider the use of non-local [216] information across the whole inputs or used the concept of dilated convolution filtering. However, where the non-local operation demands huge computational complexity, the dilated filtering often suffer from gridding effect [226], an undesirable effect in SR. Thus, one should be careful enough to enlarge the receptive field, while avoiding the computational complexity and the sacrifice of performance improvement. Working towards this realm to cover the whole features at a time with controlled number of parameters, we design a VKR module as shown in Figure 4.6. The vari-kernel residual module can be formulated as follows,

$$\begin{aligned}\hat{\mathbf{V}}_{\mathrm{m}} &= \mathbf{V}_{\mathrm{m}} + \ell(\Re^{\mathbf{1\times 1}}(\mathbf{\Psi})),\\ \Psi &=< \Re^{1}(V_m), \Re^{3}(V_m), \Re^{5}(V_m) >\end{aligned} \tag{4.11}$$

Here, $\Re^k$ represents the k × k convolution filter, ℓ represents LeakyReLU, $< \, . \, >$ is the symbolic representation for the concatenation operation, $\mathbf{V_m}$ denotes the extracted features as shown in Figure 4.6 and $\hat{\mathbf{V}}_\mathbf{m}$ represents the output of a VKR module. As illustrated from the Figure 4.6, the three convolution kernels are stacked in parallel to extract the multi-scale features at different scales. It is to be specified that gathering of multi-scale large context information helps in the utilization of the full feature map effectively which further helps in identifying the degraded regions in the image. Note that except for VKR module we refrain from using larger filter sizes (greater than 1) of convolution layers in the entire network to keep a check on the computational complexity.

Diverse Information Processing Module

The next main concern of the proposed Net$_{SDA}$ is to acquire the knowledge of the prominent degraded regions in the image. And, we believe that in order to acquire this knowledge, the network needs to be aware of the diverse information in the extracted features. And, this requires capturing enough contextual information so as to easily segregate the contaminated pixels in the image. Thus, we propose diverse processing module (DIP) as shown in Figure 4.5 (with deep red dotted line) for accomplishing this objective. Every DIP module consists of several stacked VKR modules capable of extracting features at multiple scales, which helps in providing local and global

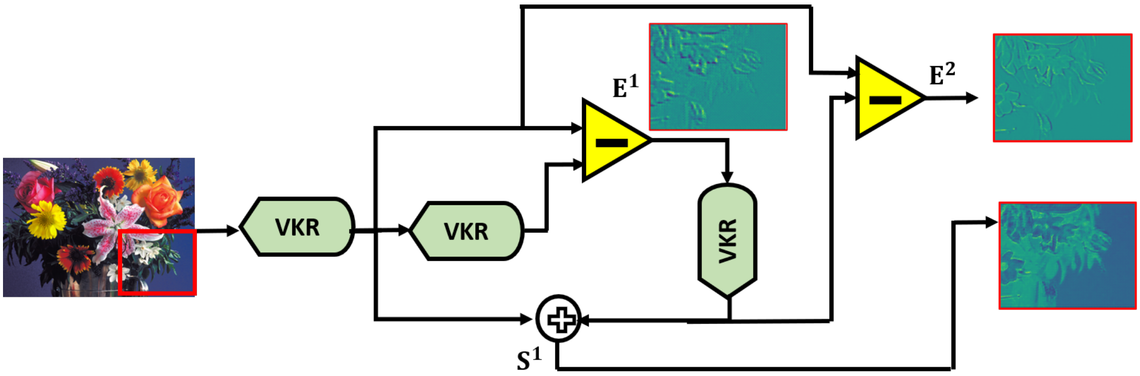


Figure 4.7: Visual illustration of the feature maps obtained for every DIP module, where E^1 and E^2 are concerned with collecting local details such as edges, textures and S^1 is more focused on collecting global details.

context [227] via increased receptive field. Furthermore, every DIP module generates three outputs, where *E's* are concerned with the high-frequency information extraction and *S's* are concerned with low-frequency information extraction. Since, low-frequency information represents the global structure (minute details) in an image and the high-frequency information represents the local details (edges, texture) in an image. Hence we can say that via this diverse information extraction, the architecture is capable of capturing the desired local and global details. Although this information can be learnt to some extent by simple convolution layers (the basic building block for all the prior works), but they struggle for heavy degradations. Additionally, in more detail every d^{th} DIP module ($d \in [1, 5]$) as shown in Eq. (4.12) generates three outputs, $\mathbf{E}^d_m$, $\hat{\mathbf{E}}^d_{2m}$ and $\mathbf{S}^d_m, (\forall m = d)$. Initially, we obtain the multi-scale high-frequency output by subtracting the outputs of two consecutive VKRs. But, since both the corruptions and structured regions are high-frequency signals in most cases, directly just learning the residual information is not sufficient. Thus, we further resort to subtract the obtained output and the difference between $(3m-2)^{th}$ and the $(3m)^{th}$ VKR for effectively learning the formatted residual information, to get a more enhanced output, $\hat{\mathbf{E}}^d_{2m}$. And, the low-frequency information, $\mathbf{S}^d_m$ for the corresponding dth DIP module in Net$_{SDA}$ is obtained as below:

$$\begin{aligned}
\mathbf{S}^d_m &= \mathbf{E}^d_{2m-1} + \hat{\mathbf{E}}^d_{2m} + 2\hat{\mathbf{V}}^d_{3m-1} \\
\mathbf{E}^d_{2m-1} &= \hat{\mathbf{V}}^d_{3m-2} - \hat{\mathbf{V}}^d_{3m-1} \\
\hat{\mathbf{E}}^d_{2m} &= \mathbf{E}^d_{2m-1} + \hat{\mathbf{V}}^d_{3m} - \hat{\mathbf{V}}^d_{3m-2}
\end{aligned} \tag{4.12}$$

Holistic Attention Refinement Network

Our proposed Holistic Attention Refinement (Net$_{HAR}$) Network, comprises of two parts: Coupled Attention (Co-Attn) module and Selective Dual Branch Merging (SDBM) module. Local attention plays a vital role for image restoration tasks, as the

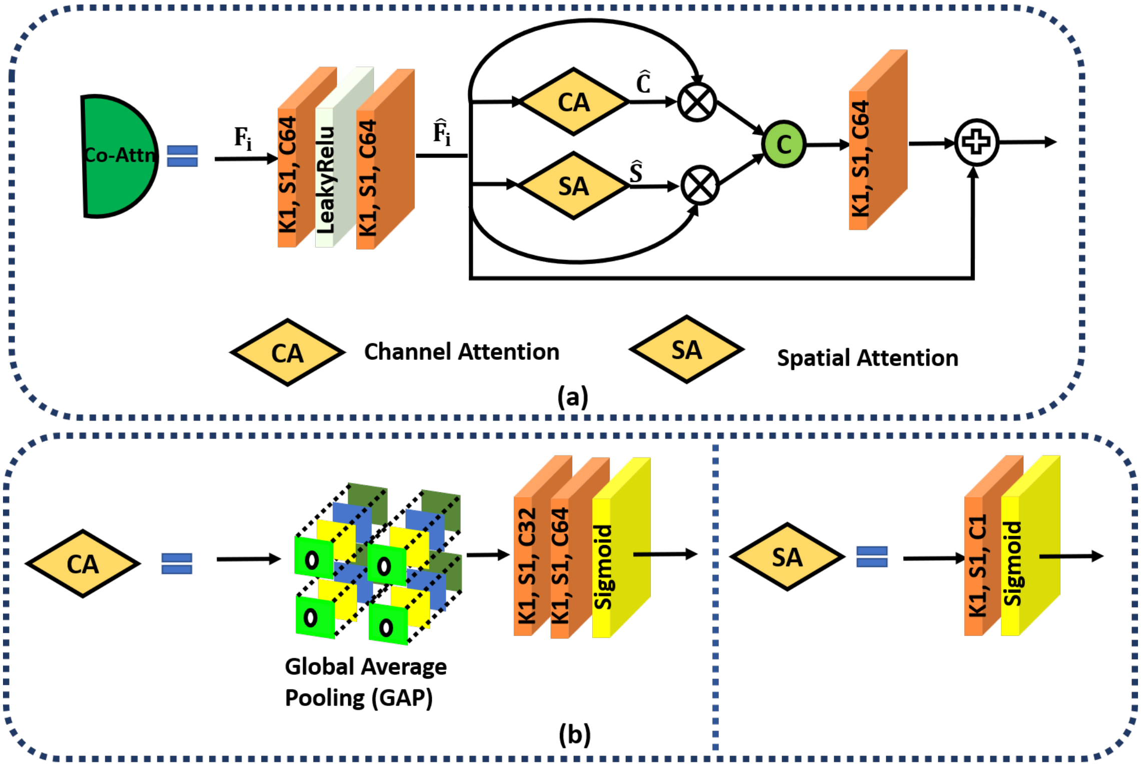


Figure 4.8: (a) Coupled attention module generates complementary features. Note: To avoid confusion, we used $\mathbf{F}$ to represent L, G of Eq. 4.13 and $\hat{\mathbf{F}}$ represents $\hat{\mathbf{L}}$ and $\hat{\mathbf{G}}$ in Eq. 4.14. (b) channel attention (CA) and spatial attention (SA) employed in Coupled attention (Co-Attn) module.

neighbourhood for a degraded pixel could be leveraged for restoring the clean image. And, the role of our proposed Net_{HAR} is to restore and formulate the interdependencies among the global and local features in channel and spatial domains extracted from Net_{SDA} and thus endow the network a capability to exploit semantically richer outputs. The architecture of the two main modules (Co-Attn and SDBM) of Net_{HAR} has been elaborated in the following subsections.

Coupled-Attention Module

To selectively process the vast amount of discriminative information from Net_{SDA}, we propose a coupled-attention (Co-attn) module. The attention mechanism trying to mimic the human visual cortex system has been widely employed in image restoration problems [228]. Based upon the same realm, our proposed Co-attn module, as shown in Figure 4.8 works collaboratively with its two attention mechanisms (channel and spatial attention) to enhance the originality of the respective extracted features while reducing their negative correlation. In which, the spatial attention focuses on the different areas in the feature maps and channel attention emphasizes the relativity of different channels, that redistributes the critical channel information and over-goes the unrelated information. As shown in Figure 4.5, we denote $\mathbf{G_i}$ and $\mathbf{L_i}$, each consisting of C feature maps with size

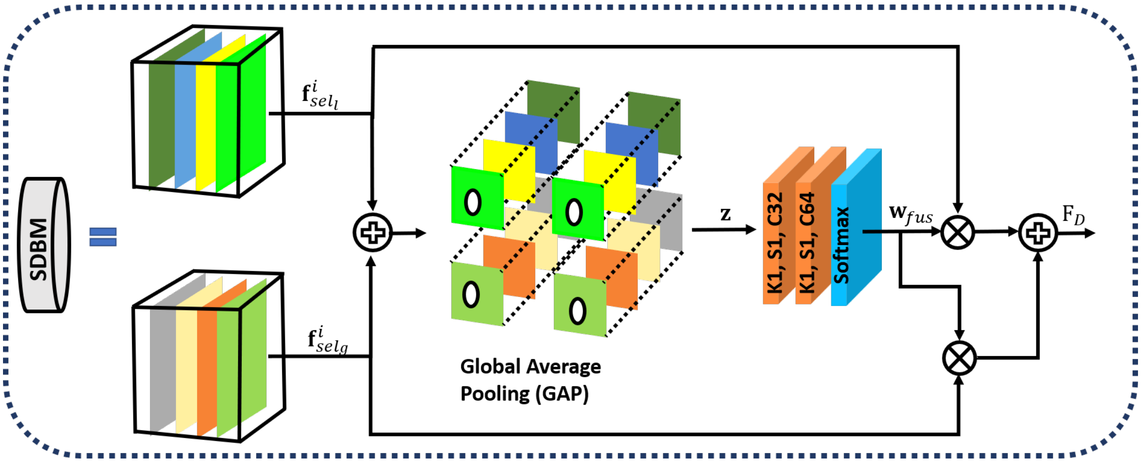


Figure 4.9: SDBM module handles the combination of the complimentary properties of both types of information extracted from Co-Attn module.

$C \times H \times W$, as the input of our i^{th} Co-Attn module.

$$\begin{aligned}\mathbf{G_i} &= \Re^1 \langle \mathbf{E}^1_1, \hat{\mathbf{E}}^1_2......\mathbf{E}^d_{2m-1}, \hat{\mathbf{E}}^d_{2m} \rangle \\ \mathbf{L}_i &= \Re^1 \langle \mathbf{S}^1_1, \mathbf{S}^2_2....\mathbf{S}^d_m \rangle\end{aligned} \tag{4.13}$$

where, $\mathbf{E}^d_m$, $\hat{\mathbf{E}}^d_{2m}$ and $\mathbf{S}^d_m$ denote the outputs of the Net$_{SDA}$ and d,m are equal to five. $\Re^1$ denotes the convolution operation with kernel size 1 and $\langle . \rangle$ denotes the concatenation operation. Now, we elaborate on the details of the attention mechanisms in our proposed Co-Attn module.

Channel attention: Considering $X \in R^{C\times H\times W}$, the image tensor in the network (where, C, H and W denotes the number of channels, height and width of the feature map, respectively). The information expressed in the feature map of each channel is different and channel attention (CA) aims to use the relationships between each channel of the feature map to learn a 1-D scalar weight $\hat{W}_c \in R^{C\times 1\times 1}$. This scalar usually represents and evaluates the importance of each channel. After obtaining the attention vector of all channels, each channel of the input $\hat{\mathbf{F}}$ is scaled by the corresponding attention value (as shown in Figure 4.8 (a). It should be noted that since, the degraded regions are usually not distributed uniformly hence spatial attention comes into picture with major goal of emphasizing on important areas.

Spatial attention: In order to maximize the effectiveness of our coupled attention module, it is best to use the spatial attention (SA) mechanism in conjunction with channel attention. It is a kind of comprehensive information module that tends to utilize the global spatial information. One of our main consideration while designing the spatial attention mechanism is it should be lightweight, besides capable of covering the spatial contents of key importance. The used SA submodule is shown in Figure 4.8 (b) and is capable of getting representative features. Further, we embed the spatial and channel attention in a parallel manner in the proposed Co-Attn module. In brief outline, the overall function of both the Co-Attn modules (in Net$_{HAR}$) shown in Figure 4.5, generating two outputs, $\mathbf{f^i}_{\mathbf{sel}_g}$ and $\mathbf{f^i}_{\mathbf{sel}_l}$ (both being the inputs to the subsequent block) could be formulated as,

$$\begin{aligned}\mathbf{f^i}_{\mathbf{sel}_g} &= \Re^1\langle \hat{\mathbf{G}}_i \bullet \hat{\mathbf{C}}_i, \hat{\mathbf{G}}_i \bullet \hat{\mathbf{S}}_i\rangle + \hat{\mathbf{G}}_i, \;\; \hat{\mathbf{G}}_i = \Re^1(\ell(\Re^1(\mathbf{G}_i))) \\ \mathbf{f^i}_{\mathbf{sel}_l} &= \Re^1\langle \hat{\mathbf{L}}_i \bullet \hat{\mathbf{C}}_i, \hat{\mathbf{L}}_i \bullet \hat{\mathbf{S}}_i\rangle + \hat{\mathbf{L}}_i, \;\; \hat{\mathbf{L}}_i = \Re^1(\ell(\Re^1(\mathbf{L}_i)))\end{aligned} \tag{4.14}$$

where, $< \; . \; >$ represents the concatenation operation, $\bullet$ is used for representing element-wise multiplication. $\hat{\mathbf{C}}_\mathbf{i}$ and $\hat{\mathbf{S}}_\mathbf{i}$ are the outputs of channel and spatial attention modules, respectively and $\mathbf{L_i}$ and $\mathbf{G_i}$ are given by Eq. (4.13).

Selective Dual-Branch Merging Module

We tend to utilize both spatial attention as well as its dual form, channel-wise feature attention. The input to the proposed Selective Dual-Branch Merging (SDBM) module (as shown in Figure 4.9) has two parallel branches, where the first branch B1 carries a reflection of the image contour $\mathbf{f^i}_{\mathbf{sel}_l}$, and the second branch B2 carries a reflection of the image details $\mathbf{f^i}_{\mathbf{sel}_g}$, hence the main role of this module aims at highlighting the respective representation ability of the extracted features. The proposed SDBM block acts as a gate for controlling the information flow from two branches, carrying different types of information into the neurons in the next layer. It first merges the resulting feature maps from two parallel ($\mathbf{f^i}_{\mathbf{sel}_l}, \mathbf{f^i}_{\mathbf{sel}_g}, \in R^{C\times H\times W}$) attentive feature maps, followed by global pooling to produce a global vector, $\mathbf{z} \in R^C$. Soft attention across the channels is further employed to adaptively select different spatial scales of information, being guided by the feature descriptor $\mathbf{z}$. While restoring the degraded pixels, it assigns non-zero weights to the more useful features, delivering strong performance as it suppresses the influence of less important features. Contrary to the existing methods that employ concatenation or addition of the extracted features, our merging block is adept for selecting the useful set of attentive features from each branch representations by utilising an attention branch and thus fusing the dual-branch results in an adaptive way. To summarize the overall operation of SDBM module, after getting $\mathbf{f^i}_{\mathbf{sel}_g}$ and $\mathbf{f^i}_{\mathbf{sel}_l}$, the final output F$_D$ is obtained by the SDBM module in the following three steps:

$$\begin{aligned}\mathbf{z} &= GAP(\mathbf{f}^i_{sel_l} + \mathbf{f}^i_{sel_g}) \\ \mathbf{w}_{fus} &= \Omega(\Re^1(\Re^1(\mathbf{z})) \\ \mathbf{F}_D &= \mathbf{f}^i_{sel_l} \bullet \mathbf{w}_{fus} + \mathbf{f}^i_{sel_g} \bullet \mathbf{w}_{fus}\end{aligned} \quad . \tag{4.15}$$

where, Ω is used for Softmax operation.

Channel Attentive Upsampling Block

We focus on another fundamental operation that has received relatively less attention for the task of super-resolution: upsampling block. Since, the main aim of super-resolution is to improve the resolution of the image, hence the upsampling method plays an indispensable role for any SR algorithm and the way it is performed affects the overall final result. The proposed channel attentive upsampling (CAU) block has been designed specifically to integrate the low-level details into the final feature maps for recovering the required lost information. Figure 4.10 shows the schematic layout of the proposed upsampling block with scale factor × 4.

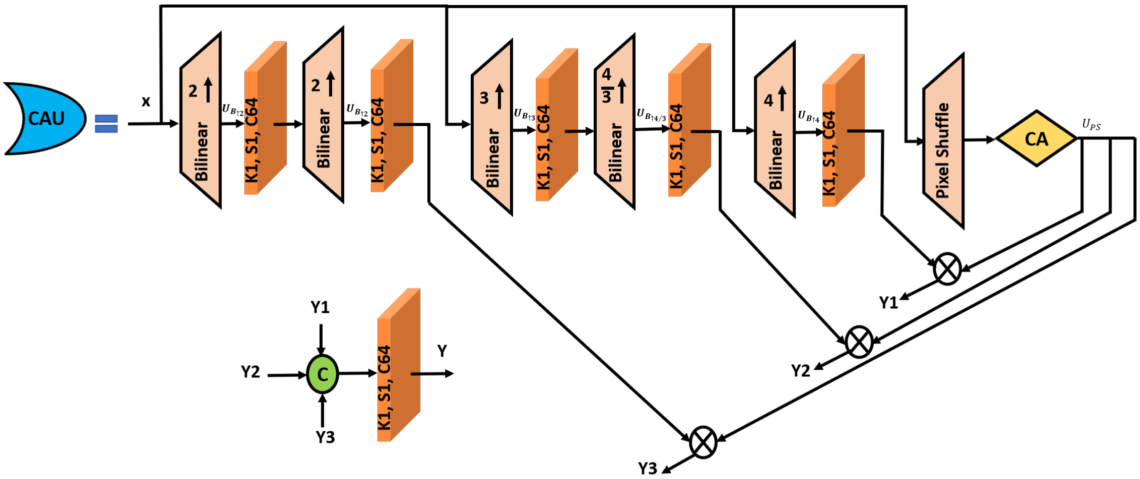


Figure 4.10: Schematic layout of Channel Attentive Upsampling Block upscaling by factor ×4.

Initially, we perform element-wise multiplication of the feature maps obtained through pixel shuffle layer with the following three feature maps: (1) obtained directly through bilinear interpolation by scale factor of 4, ($\mathbf{Y}_1$), (2) obtained progressively by factor of 3, ($\mathbf{Y}_2$) and (3) progressively by a factor of 2, ($\mathbf{Y}_3$) followed by their concatenation. The proposed module serves the binary purpose of exploiting the advantages of 1) both direct and progressive upsampling, 2) both traditional and recent upsampling techniques. The overall concatenation of the individual feature maps (Y_1,Y_2, Y_3) helps in fusing different level features becoming a suitable candidate for providing rich information in SR reconstruction. The overall formulation of the proposed CAU block is given below:

$$\begin{aligned}
\mathbf{Y_1} &= \hat{\mathbf{C}}(U_{PS}(\mathbf{x})) \bullet (\Re^1(U_{B_{\uparrow 4}}(\mathbf{x}))) \\
\mathbf{Y_2} &= \hat{\mathbf{C}}(U_{PS}(\mathbf{x})) \bullet (\Re^1(U_{B_{\uparrow 4/3}}(\Re^1(U_{B_{\uparrow 3}}(\mathbf{x}))))) \\
\mathbf{Y_3} &= \hat{\mathbf{C}}(U_{PS}(\mathbf{x})) \bullet (\Re^1(U_{B_{\uparrow 2}}(\Re^1(U_{B_{\uparrow 2}}(\mathbf{x}))))) \\
\mathbf{Y} &= \Re^1\langle \mathbf{Y_1}, \mathbf{Y_2}, \mathbf{Y_3} \rangle
\end{aligned} \tag{4.16}$$

where, $\hat{\mathbf{C}}$ denotes output of channel attention module. $U_{PS}(\mathbf{x})$ denotes the pixel-shuffling operation and $U_{B_{\uparrow s}}(\mathbf{x})$ shows interpolation by bilinear operation with factor s.

Table 4.6: Performance evaluation on four datasets for ×2 and ×4 super-resolution.

Scale	Size Scope	Methods	Params	FLOPs	Set5 [1]	Set14 [2]	B100 [3]	Urban100 [4]
×2	<5 × 10^2 K	Bicubic	-	-	33.66/0.9299	30.24/0.8688	29.56/0.8431	26.88/0.8403
		DRCN [229]	1774K	17,974.3G	37.63/0.9588	33.04/0.9118	31.85/0.8942	30.75/0.9133
		VDSR [230]	666K	612.6G	37.53/0.9587	33.03/0.9124	31.90/0.8960	30.76/0.9140
		MemNet [231]	678K	2662.4G	37.78/0.9597	33.28/0.9142	32.08/0.8978	31.31/0.9195
		IDN [60]	590K	174.6G	37.83/0.9600	33.30/0.9148	32.08/0.8985	31.27/0.9196
		CARN [58]	1592K	222.8G	37.76/0.9590	33.52/0.9166	32.09/0.8978	31.92/0.9256
		BSRN [57]	594K	1666.7G	37.78/0.9591	33.43/0.9155	32.11/0.8983	31.92/0.9261
		IMDN [59]	694K	158.8G	38.00/0.9605	33.63/0.9177	32.19/0.8996	32.17/0.9283
		MADNet [232]	878K	187.1G	37.94/0.9604	33.46/0.9167	32.10/0.8988	31.74/0.9246
		LatticeNet [61]	756K	169.5G	**38.15/0.9610**	**33.78/0.9193**	**32.24/0.9005**	32.43/0.9302
		OverNet [233]	900K	200G	38.11/**0.9610**	33.71/0.9179	**32.25/0.9007**	**32.44/0.9311**
		Ours	**640K**	**86G**	**38.19/0.9616**	**33.80/0.9194**	32.19/0.8992	**32.32**/**0.9306**
×4	<5 × 10^2 K	Bicubic	-	-	28.42/0.8101	25.99/0.7023	25.96/0.6672	23.14/0.6573
		DRCN [229]	1774K	17,794.3G	31.53/0.8854	28.02/0.7670	27.23/0.7233	25.14/0.7510
		VDSR [230]	666K	612.6G	31.35/0.8830	28.02/0.7680	27.29/0.7260	25.18/0.7540
		MemNet [231]	678K	2662.4G	31.74/0.8893	28.26/0.7723	27.40/0.7281	25.50/0.7630
		IDN [60]	590K	32.3G	31.99/0.8928	28.52/0.7794	27.52/0.7339	25.92/0.7801
		CARN [58]	1592K	90.9G	32.13/0.8937	28.60/0.7806	27.58/0.7349	26.07/0.7837
		BSRN [57]	594K	451.8G	32.14/0.8937	28.56/0.7803	27.57/0.7353	26.03/0.7835
		IMDN [59]	694K	40.9G	32.21/0.8948	28.58/0.7811	27.56/0.7353	26.04/0.7838
		MADNet [232]	878K	54.1G	31.95/0.8917	28.44/0.7780	27.47/0.7327	25.76/0.7746
		LatticeNet [61]	756K	43.6G	32.30/**0.8962**	28.68/**0.7830**	27.62/0.7367	26.25/0.7873
		OverNet [233]	940K	200G	**32.32**/0.8956	**28.71**/0.7826	**27.67/0.7373**	**26.31/0.7923**
		Ours	**670K**	**35G**	**32.57/0.9001**	**28.73/0.7857**	**27.67/0.7395**	**26.44/0.7972**

4.2.2 Experimental Analysis

In this section, initially we describe the datasets and the implementation details and next we report the results on three tasks a) image super-resolution, b) image enhancement, and c) image denoising. Following which we describe the contributions of the several components in the proposed model through detailed ablation study.

Datasets

Image Super-resolution: In our work, for the task of super-resolution, we have trained our model on **DIV2K** dataset [5], that is composed of high-quality 800 training images, 100 validation and 100 testing images. The trained model has been evaluated on Set5 [1], Set14 [2], B100 [3] and Urban100 [4] datasets.

Image Enhancement: The performance of the proposed Con-Net has been further validated on two real-world enhancement datasets: **MIT-Adobe FiveK** [14] and **LoL** [13] datasets. MIT-Adobe FiveK consists of 5000 images from various indoor and outdoor scenes captured using DSLR cameras in different lighting conditions. The tonal attributes of all the images have been manually adjusted by five different trained photographers (labelled as experts A to E). Inspired by [234], we have considered the enhanced images of expert C as ground-truth. We have considered the first 4500 images for training and the last 500 has been used for testing. LoL dataset, with 485 images for training and 15 for testing consists of low-light input image and its carefully-exposed reference images.

Image Denoising: Since real noises are usually more signal dependent and are spatially varying depending on different in-camera pipelines, real image denoising is a highly challenging task. For real image denoising, aiming at restoring the high-quality images given noisy inputs, we used **Smartphone Image Denoising Dataset** (SIDD) [10] consisting of 320 noisy image pairs for training and 1280 images for validation. It is collected using five smartphone cameras in 10 static scenes. On account of small sensor and high-resolution, the noise in smartphone images is quite higher as compared to DSLRs. Besides, we also use **Darmstadt Noise Dataset** (DND) [9] consisting of 50 image pairs from four consumer cameras. As the images are of high resolution, dataset holders have extracted 20 crops from each image, yielding 1000 patches in total. Since, for DND the ground truth images are not released publicly, hence we obtained the image quality scores through online server.

Training settings

The proposed model is end-to-end trainable without any requirement of pre-training of sub-modules. We have trained three different networks for restoration and enhancement tasks. Except for the task of super-resolution, where we use an extra upsampling module, the following trainable parameters are common to all the experiments: We use five DIPs,

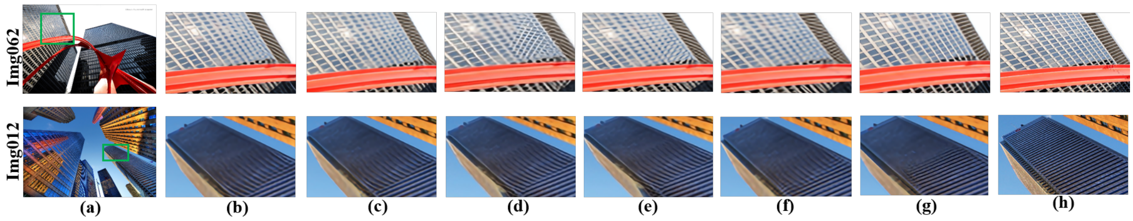


Figure 4.11: Comparisons for scale factor ×4 from Urban100 [4] dataset. (a) Input image, (b) BSRN [57], (c) CARN [58], (d) IMDN [59], (e) IDN [60], (f) Latticenet [61], (g) **Con-Net (Ours)**, (h) Ground-truth.

each of which contains three VKRs in the Net$_{SDA}$ and two Co-attn modules in Net$_{HAR}$. The training hyper-parameters have been implemented using NVIDIA DGX station with processor 2.2GHz, Intel Xenon E5-2698, NVIDIA Tesla V100 1×16 GB GPU. The model has been optimized using ADAM optimizer [207] with $\beta1$, $\beta2$ and ϵ set to 0.9, 0.999 and 10^{-8}, respectively. The initial learning rate has been set to 10^{-4} and we used cosine annealing strategy [235] to steadily decrease the learning rate from initial value to 10^{-6} during the training. Batch size for all the three tasks is set to 1 and, for data augmentation, we performed the horizontal and vertical flips. Following the trend in most of the image restoration applications, we adopted PSNR and SSIM as the evaluation metrics.
To show the effectiveness of the proposed Con-Net, we have chosen the *L1* loss as our objective function. Given a training set with a large number of pairs, of corrupted inputs $\widehat{x}_i$ and their clean labels x_i. The major objective of the training is defined as:

$$L1 = |f_\phi(\widehat{x}_i) - x_i| \tag{4.17}$$

where, f_ϕ is a parametric family of mappings of the proposed Con-Net under *L1* loss function. And, the notation $\widehat{x}_i$, underlines the fact that the corrupted input $\widehat{x} \sim p(\widehat{x}|x_i)$ is a random variable, that is distributed according to the clean target.

Comparison with State-of-the-Art methods

In this section, we compare our proposed Con-Net with some recent SOTA methods in synthetic image super-resolution, real-world de-noising and real-world low-light enhancement tasks.

Results on Super-resolution

We compare our proposed Con-Net against fourteen SoTA SR approaches on four benchmark datasets for scale factor ×2 and ×4. Following the common practise in SR literature, we computed the PSNR and SSIM metrics on Y channel. Table 4.6 shows quantitative evaluation results, including the number of parameters and FLOPs. The results clearly show that our method significantly advances SoTA lightweight super-resolution methods and consistently outputs better image quality scores for both

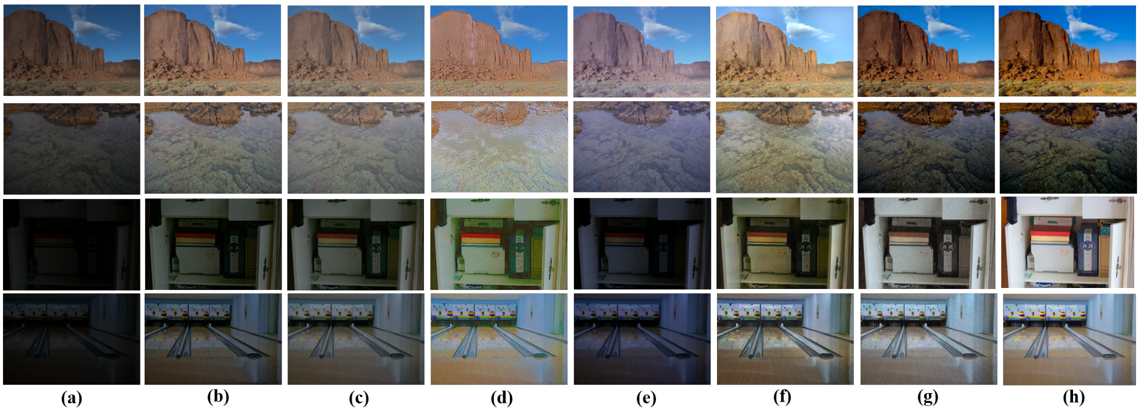


Figure 4.12: Visual Comparisons for low-light enhancement tasks on LoL [13] and MIT-Adobe FiveK [14] datsets. (a) Input image, (b) SRIE [62], (c) CRM [63], (d) RetinexNet [64], (e) Zero-DCE [65], (f) EnGAN [66], (g) **Con-Net (Proposed Method)**, (h) Ground-truth.

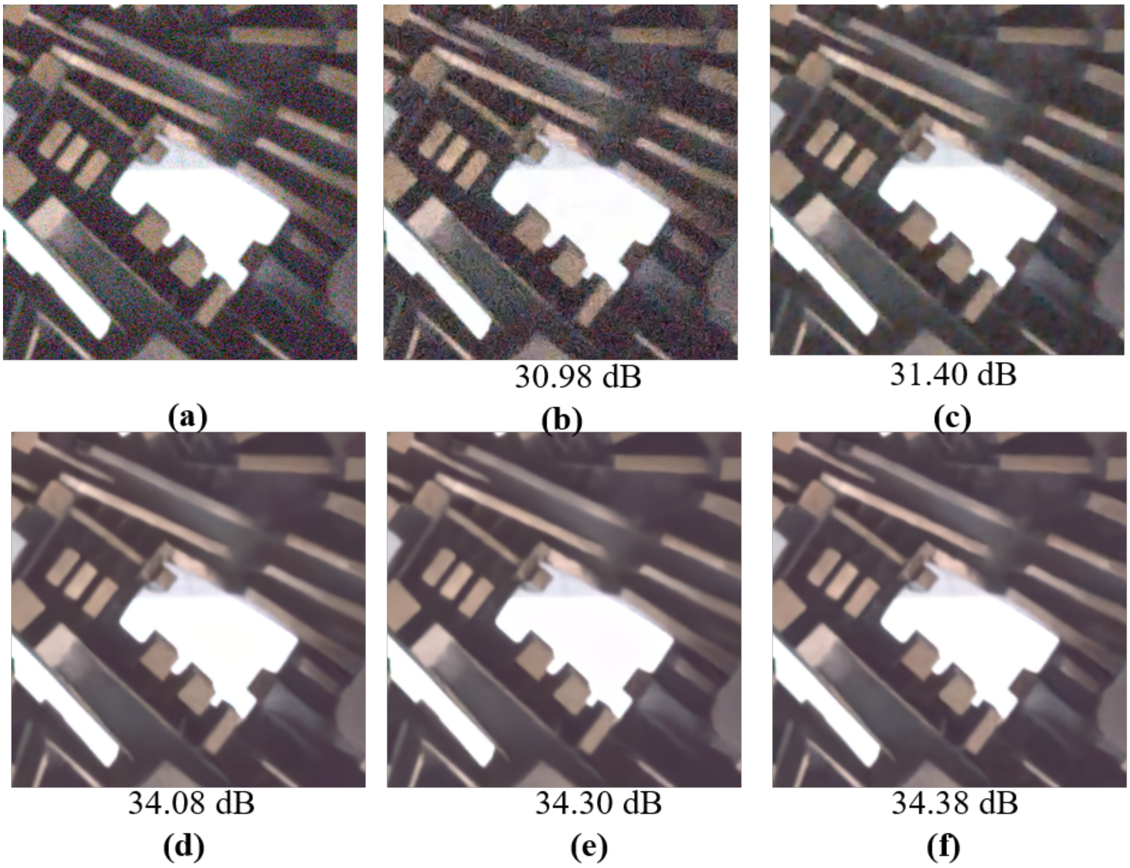


Figure 4.13: Denoising example from DND [9] dataset. (a) Input Image, (b) DnCNN [67], (c) CBDNet [68], (d) VDN [69], (e) DANet [70], and (f) **Con-Net (Proposed Method)**.

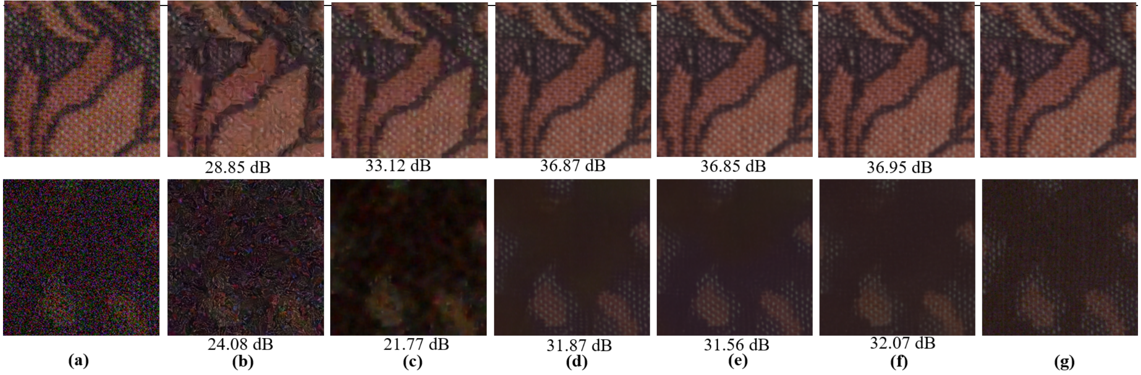


Figure 4.14: Denoising examples from SIDD dataset [10]. Our proposed Con-Net, besides effectively removing noise from the challenging images, has higher PSNR gain. (a) Input image, (b) DnCNN [67], (c) CBDNet [68], (d) VDN [69], (e) DANet [70], (f) **Con-Net (Proposed Method)**, and (g) Ground-truth.

scale factors. Particularly, when compared to the recent lightweight SR methods, Overnet [233] and Latticenet [61], our proposed method shows a performance gain of about 0.27 dB and 0.25 dB, respectively on Set5 [62] dataset for scale factor 4.

Figure 4.11 presents the visual comparisons on B100 and Urban100 for scale factor ×4. The figure clearly shows the supremacy of our method in comparison to others by reconstructing images much closer to the HR images. For instance, in image Img062, we observe that unlike our proposed Con-Net all the compared methods fail to recover the orientation of lines. And, for image Img012, the texture in case of all the predicted SR images for the compared SR methods contains blur or aliasing. Whereas, our method is more efficient at reconstructing the structural patterns and edges, thus generating images that are more natural looking alike ground-truth images.

Table 4.7: Quantitative Results (PSNR, SSIM) of SoTA methods and ours on MIT-Adobe 5K and LOL datasets for low-light enhancement tasks.

Methods	Parameters	MIT [14]		LoL [13]	
	(M)	PSNR	SSIM	PSNR	SSIM
MBLLEN [236]	0.45	15.59	0.71	13.93	0.49
GLADNet [237]	1.13	16.73	0.76	16.19	0.61
RetinexNet [64]	2.11	12.69	0.77	13.09	0.43
EnGAN [66]	8.64	15.01	0.77	15.64	0.58
KinD [238]	8.04	17.17	0.69	14.62	0.64
DRBN [239]	0.58	15.95	0.70	15.32	0.70
FIDE [240]	8.62	17.17	0.69	16.71	0.67
DeepUPE [241]	2.99	18.78	0.82	13.04	0.48
CSDNet [242]	17.29	18.48	0.85	**21.63**	**0.85**
ZeroDCE [65]	7.94	16.46	0.76	15.51	0.55
BLNet [243]	-	-	-	20.14	0.72
RUAS [244]	0.41	**20.83**	**0.85**	18.23	0.35
Ours	**0.57**	**23.36**	**0.91**	**21.98**	**0.88**

Results on Low-Light Image Enhancement

To prove the effectiveness of our proposed work on real-world enhancement tasks, we compared it with twelve SoTA low-light image enhancement algorithms. We reported the quantitative scores on the MIT-Adobe-5K [14] dataset in Table 4.7, where our method achieves the best numerical scores. Contrary to other methods, in Figure 4.12 (first two rows) featuring outdoor scenes, Con-Net enhances both the dark regions and preserves the color of the input images. The results generated by our method are visually more pleasing, without obvious noise and color casts. Though, some methods [66], [65] were successful at enhancing the brightness successfully, but they failed to restore the clear image textures. In contrast, our Con-Net restores both the details and brightness perfectly, at the same time. To further prove the efficacy of our approach in real-world scenarios, we evaluate Con-Net on LoL [13] dataset, which contains sensible noise to hinder the image enhancement. Obviously, as mentioned in the Table 4.7, our Con-Net outperforms several recently proposed methods [243], [244], [241] while maintaining an attractive computational complexity. Notably, when compared with the recent best methods, proposed Con-Net achieves 2 dB performance gain over BLNet [243] on the LoL dataset and 3 dB improvement over RUAS [244] for the Adobe-Fivek dataset. A similar trend follows for the SSIM scores, as well. As clear from the Figure 4.12 (last two rows) featuring indoor scenes, of LoL [13] dataset where, most of the compared methods fail at generating vivid and true colors, our proposed Con-Net is efficient in preserving colors and contrast.

Table 4.8: Comparison of Con-Net against other SoTA methods in real-world image denoising On DND [9] and SIDD [10].

Methods	Parameters (M)	DND [9]		SIDD [10]	
		PSNR	SSIM	PSNR	SSIM
DnCNN [245]	0.56	32.43	0.7900	23.66	0.5830
EPLL [246]	-	33.51	0.8244	27.11	0.8700
TNRD [247]	-	33.65	0.8306	24.73	0.6430
FFDNet [248]	0.48	34.40	0.84740	-	-
BM3D[249]	-	34.51	0.8507	25.65	0.6850
NC [250]	-	35.43	0.8841	-	-
KSVD [251]	-	36.49	0.8978	26.88	0.8420
CBDNet [68]	4.34	38.06	0.9421	33.28	0.8680
RIDNet [252]	1.49	39.26	0.9528	-	-
PRIDNet [253]	-	39.42	0.9528	-	-
DRDN [254]	-	39.43	0.9531	-	-
GradNet [255]	1.60	39.44	0.9543	38.34	0.9460
AINDNet [256]	13.76	39.53	0.9561	39.08	0.9530
VDN [69]	7.81	39.38	0.9518	39.26	0.9550
DANet [70]	63.01	**39.58**	**0.9545**	39.25	0.9550
InvDN [257]	2.64	**39.57**	**0.9522**	**39.28**	**0.9550**
Ours	**0.57**	39.32	0.9514	**39.31**	**0.9610**

Results on Real-world Denoising

In this section, we demonstrate the effectiveness of the proposed Con-Net for the task of real-world denoising, where we compare it with sixteen SOTA denoising approaches. As already mentioned, we have trained our model on the SIDD [10] medium training set. Since, DND [9] does not provide any training set, therefore for it we have used the model

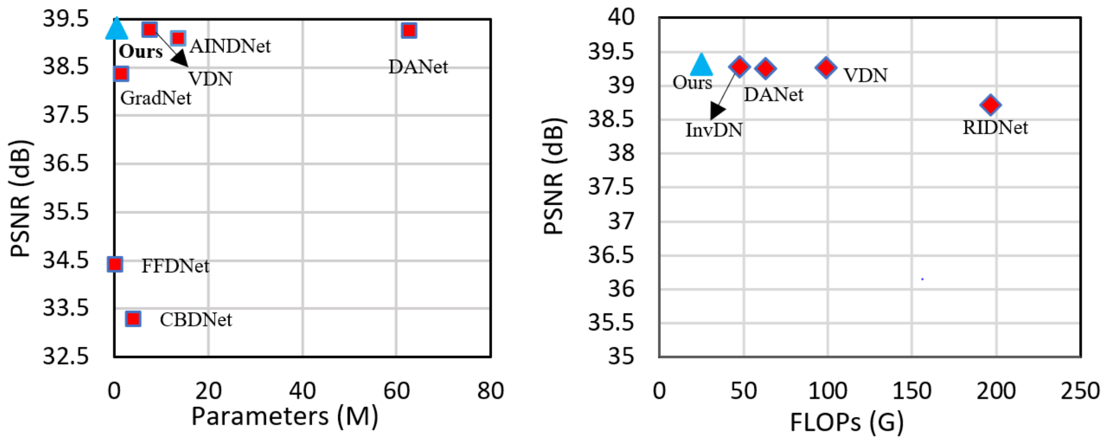


Figure 4.15: Trade-off between performance *vs.* model size on SIDD [10].

trained on SIDD. For a fair comparison, PSNR/SSIM of other compared models on the test set is directly taken from the official leaderboards of DND and SIDD and further verified from their respective articles. Table 4.8 represents the test results of various data-driven, as well as conventional, SOTA denoising modules on SIDD and DND dataset clearly proving the attractive performance of our proposed Con-Net. To further prove the superiority of our proposed Con-Net, we also compare it with the recently proposed InvDN [257], and from Table 4.8 we can clearly see the superiority of Con-Net on SIDD dataset and its comparable performance on DND dataset. Furthermore, it is worth emphasising that CBDNet [68] and RIDNet [252] used additional training data, yet our method yield significantly better results. For instance, our method achieves 6.03 dB performance gain on SIDD dataset and 1.26 on DND dataset, when compared to CBDNet [68].

In Figure 4.13 and Figure 4.14, we present the visual comparisons of our results with other methods for image denoising. It is evident that our proposed Con-Net is more capable at maintaining the smoothness of the homogeneous regions without any sacrifice of the structural content and fine textural details. Moreover, our proposed model is capable of recovering the subtle edges while other models results in introduction of blockiness, fuzziness, and random dots particularly along the edges.

Computational Complexity

Keeping in mind, that the number of parameters are not the complete reflection of the complexity of model, hence we employ both the number of trainable parameters and floating point operations (FLOPs) that a model takes for the processing of 256×256 image. The complexity analysis of the compared representative methods has been reported in Figure 4.15. We have provided the denoising performance on the SIDD dataset of various methods. As clear from the above figure compared with the recent SoTA denoising algorithms, our proposed Con-Net have an attractive complexity and achieve good promising performance with 26.72 G FLOPs and 0.57 Million parameters.

Table 4.9: Importance of Con-Net modules evaluated on LoL dataset for low-light enhancement tasks. DIP, VKR, SDBM, Co-attn, NOD denotes Diverse Information Processing, Vari-kernel Residual, Selective Dual-Branch, Coupled Attention modules and Number of DIP modules, respectively.

Methods	DIP	VKR	SDBM	Co-Attn	NOD	PSNR
Net1	×	✓	✓	✓	5	17.07
Net2	✓	×	✓	✓	5	18.01
Net3	✓	✓	✓	×	5	20.14
Net4	✓	✓	×	✓	5	18.18
Net5	✓	✓	×	✓	5	18.93
Net6	✓	✓	✓	✓	4	18.67
Net7	✓	✓	✓	✓	3	18.42
Net8	✓	✓	✓	✓	2	18.11
Con-Net	✓	✓	✓	✓	5	**21.98**

4.2.3 Ablation Study

We study the influence of each of our architectural components and design choices on the final performance of our proposed model. Table 4.9 has been presented to quantify the effect of the performance of Con-Net for LoL [13] dataset.

For validating our design considerations, we implemented the following baselines (reported in Table 4.9). Net1: We modified all the Diverse Information Processing (DIP) modules from our proposed Con-Net by employing VKR module in series, generating 1 output (instead of 3) per DIP module. Net2: We replaced all the vari-kernel residual modules with simple residual block as proposed in [217]. Net3: We removed the coupled attention modules from our proposed network. Net4: We replaced the SDBM module of our proposed Con-Net with simple addition operation. Net5: We replaced the SDBM module with concatenation operation. In Net6, Net7 and Net8, keeping all other components same we reduced the number of DIP modules (NOD) employed.

Effectiveness of DIP, VKR and Co-Attn modules: It is clear that our proposed Con-Net exhibits more superiority over its incomplete versions, including Net1, Net2 and Net3 surpassing them by 4.91 dB and 3.97 dB and 1.84 dB (PSNR), respectively. The substantial drop in PSNR while removing these modules from our proposed Con-Net clearly demonstrates the advantages of employing DIP, VKR and Co-Attn modules in our network.

Effectiveness of SDBM module: To verify the necessity of Selective Dual Branch Merging (SDBM) module, we compare our proposed Con-Net with simple addition and concatenation operations in place of SDBM module. The 3.8 dB gain improvement by our proposed Con-Net over Net4 and 3.05 dB gain over Net5 is attributed to SDBM design which is capable of effectively merging the information in both the degraded and clean regions. This comparison proves the fact that since both the addition and subtraction operations of DIP module contains complementary information; and fusion of these features using trivial concatenation or element wise feature addition may overlook the redundant information, hence signifying the advantage of using SDBM module.

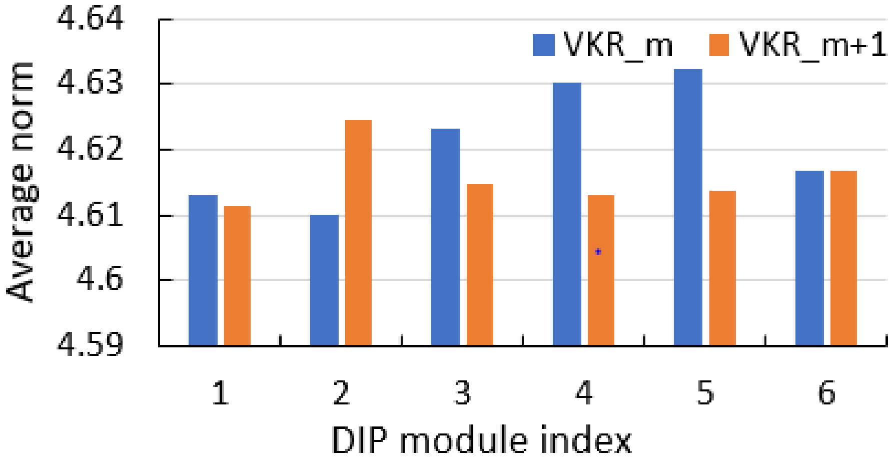


Figure 4.16: The norm of filter weights *vs.* the DIP module index. Every set of histogram corresponds to one DIP module. Out of the three VKR modules, present in every DIP module, we have shown the response of first and last VKR.

Table 4.10: Quantitative comparison of the popular upsampling techniques for scale factor ×4 On Set5 [1] for super-resolution.

Methods	Upsampling	Set5 [1]		Set14 [2]	
		PSNR	SSIM	PSNR	SSIM
M1	Pixel Shuffle (PS)	32.46	0.8993	28.66	0.7848
M2	Deconvolution (DEC)	32.38	0.8982	28.53	0.7837
M3	BiL + Convolution	32.40	0.8986	28.58	0.7840
M4	**CAU**	**32.52**	**0.8999**	**28.71**	**0.7852**

Effectiveness of the number of DIP modules: We now assess the influence of the number of DIP modules on the overall image restoration performance. As clearly shown in the Table 4.9 (Net6-Net8), the performance decreases with decline in number of DIP modules. This indicates the important role of DIP in exploiting the multi-scale edge information. By making a tradeoff between performance and comparison, the default number of DIP for our proposed network is set to 5.

We further illustrate the effect our proposed DIP module in the different stages of the network. Inspired by [231], we adopt weight-norm as an approximation for the dependency of current layer on its preceding layers. We have calculated the weight norm using the corresponding weights from all the filters w.r.t each feature map in the DIP module. Figure 4.16 represents the norm of the filter weights *vs.* DIP module index. Generally, more is the norm, then stronger is the dependency on the particular feature map. From the plot, we further draw the following conclusions: (1) The weights of the feature maps have been spread over all the blocks, indicating all the residual features have been used to produce the output features of DIP module. (2) The variance of weight norms in latter modules are much larger than that of the initial modules proving the discriminability of the network to distinguish residual features. (3) While increasing the number of DIP > 5, the variance of weight norms seems to be almost same in both the Vari-kernel residual (VKR) modules of DIP indicating no benefit of increasing the DIP module beyond 5.

Table 4.11: Comparison with prior network configurations. Here we replace our VKR module with residual configurations of the popular SOTA SR methods for ×4.

Configuration	Parameters	Set5		B100	
		PSNR	SSIM	PSNR	SSIM
Con-Net+ [26]	458k	32.27	0.8946	27.47	0.7353
Con-Net+ [44]	616k	32.35	0.8983	27.59	0.7368
Con-Net+ [81]	892k	32.40	0.8995	27.62	0.7389
Con-Net	670k	32.57	0.9001	27.67	0.7395

Effectiveness of Channel Attentive Upsampling module:

We exhibit on two popular benchmarks (Set5 [1] and Set14 [2]) in the task of the single image super-resolution, that the proposed Con-Net performs consistently better than the popular upsampling techniques of pixel-shuffle (PS), strided transpose convolution (DEC) layer and Bilinear (BiL) interpolation followed by Convolution. Table 4.10 clearly demonstrates the efficacy of our proposed Channel Attentive Upsampling (CAU) module surpassing other techniques by huge margin in both SSIM and PSNR. Thus, the proposed CAU could be used as a drop-in replacement of other upsampling layers as it improves the model performance on account of the content adaptive nature of the attention generated kernels with almost same number of parameters. Additional ablation studies and qualitative results are given in the supplementary material.

Effectiveness of Vari-kernel Residual Module:

We further analyse the importance of proposed Vari-kernel residual (VKR) module by comparing it with popular configurations as shown in Table 4.11. All the configurations have been trained on the same dataset and experimental settings. The improvement in performance gain of the proposed VKR can be attributed to the fact the proposed multi-scale approach is more successful at capturing the contextual information as compared to the residual block [26], residual channel attention block [125], and non-local residual block [81]. This full exploitation of local features is clearly visible from the performance gain on PSNR of about 0.30, 0.33, and 0.17 dB of our proposed network when compared to [26], [125], and [81], respectively.

4.3 Summary

In this chapter, we have discussed two different lightweight approaches for super-resolution. In our first solution (Section 4.1), we have proposed an effective multi-scale attention residual network capable of learning the fine image details that are more realistic to the ground truth with relatively less number of parameters. It is composed of multi-scale attention residual (MSAR) block and up and down projection block, for better collection of global and local contextual information. In parallel, for modulation of multi-level features in global and local manner, channel and spatial attention in MSAR blocks is being utilized. In our second solution (Section 4.2), we have proposed Con-Net, a network design capable of exploiting the non-uniformities of the degradations in spatial-domain with limited number of parameters (656k). Our proposed Con-Net comprises of basically two main components, (1) a spatial-degradation aware network for extracting the diverse information inherent in any degraded image, and (2) a holistic attention refinement network for exploiting the knowledge from the degradation aware network to selectively restore the degraded pixels. Extensive qualitative and quantitative comparison with prior arts on benchmark datasets demonstrates the efficacy of our proposed solutions over existing state-of-the-art heavy and lightweight architectures in terms of parameter and FLOPs reduction. In this chapter and the previous chapter we discussed in detail about both lightweight and heavyweight single frame super-resolution approaches, but more research needs to be done on multi-frame super-resolution approaches. A detailed discussion is given in the next chapter.

Chapter 5

A Novel Approach for Burst Restoration and Enhancement

In the recent past, smartphone industry has witnessed a rampant growth on account of the fueling demand of smartphones in day-to-day life. However, there are several hardware barriers that hinder the smartphone industries for manufacturing versatile and small camera sensors for these smartphones. Some of these hardware barriers are with regard to; the small aperture lens and sensor size that limits their spatial resolution and dynamic range [258]. Thus while acquiring images in smartphones, more informative components are generally aliased or lost that hampers their proficiency in regenerating DSLR-alike images. This problem has heightened the focus nowadays towards software solutions for mitigating the hardware limitations in smartphones and to sustain the quality gap pertaining to DSLRs.

Recently, the smartphone camera pipeline has cultivated around the concept of capturing and merging burst images of the same scene for leveraging high quality image than possible through the capture of single image. The burst image processing approaches aims to recover the latent high-quality image by exploiting the multi-frame complementary information. Recent works [75, 259] have validated the potential of burst processing approaches in reconstructing richer details that cannot be recovered from a single image. However, these computationally extensive approaches are unable to effectively model the inherent sub-pixel shifts among multiple frames on account of the camera and scene motion from dynamic moving objects. These sub-pixel shifts often pave the way for ghosting and blurring artefacts in the resultant output image. To tackle these shifts, existing methods employ complex explicit feature alignment [7], deformable convolutions [75], *etc.* Since these approaches target only upon local features at single level, they leave a scope of improvement in feature alignment. Additionally while aggregating multi-frame features, existing approaches either employ late fusion strategy [7, 74] or rigid fusion mechanism (in terms of number of inputs [75]. The former one limits the flexible inter-frame communication, while the later one limits the adaptive multi-frame processing.

To tackle aforementioned problems, we propose two novel solutions for burst processing as:

1. Adaptive Feature Consolidation Network for Burst Super-Resolution.

2. Burst Restoration and Enhancement.

Each solution is discussed in detail in the subsequent sections.

5.1 Adaptive Feature Consolidation Network for Burst Super-Resolution

SISR is the task of generating HR image using a single LR image. Numerous methods have been developed to solve the SISR problem [260, 261]. However, the major hurdle lies in synthesizing high-frequency details in a single input image, consistent to the ground-truth HR image.

On the other hand, Multi-frame super-resolution seeks the reconstruction of HR image by employing numerous degraded LR images of a scene. Critically, capturing LR images under the burst mode results in sub-pixel shifts [262] among the multiple LR burst images and thereby, generates different LR samplings of the underlying scene. However, the process of burst image acquisition brings its own issues. For example, during image capturing, any slight movement in scene objects and/or scene objects arises misalignment issues, thereby generating blurring and ghosting artifacts in the reconstructed image [263]. The existing MFSR approaches utilize pre-trained flow computation [264] or optical-flow [265] for aligning the multi-frame features. This explicit feature alignment causes the resulting errors in the flow estimation stage to be propagated to the image processing and warping stages, thereby negatively affecting the generated outputs.

To mitigate the aforementioned problems, we propose an Adaptive Feature Consolidation Network (AFCNet) for multi-frame super-resolution. The proposed AFCNet comprises of four steps: 1) Feature alignment, 2) Feature extraction, 3) Feature fusion and 4) Feature up-sampling. The features of RAW burst images are initially aligned through deformable convolution [266] followed by feature back-projection approach. This implicit feature alignment limits the error propagation inherent in cascaded explicit alignment approaches [264, 265]. Further, the aligned representations of each burst image are passed through a feature extractor [267] to extract multi-scale local-global representations. The feature fusion mechanism enables the inter-frame communication via abridged pseudo-burst generation such that each and every feature in the pseudo-burst encloses complimentary properties of all input burst images. Furthermore, we adopt an adaptive group up-sampling module [268] to select the reliable and desired information content from each burst image and thus obtain the high-quality HR result.

On account of above modules, our framework efficiently merges the image contents among multiple burst LR RAW frames in a coherent and effective way, generating HR RGB outputs with realistic textures and additional high-frequency details. Highlights of the proposed approach are outlined as follows:

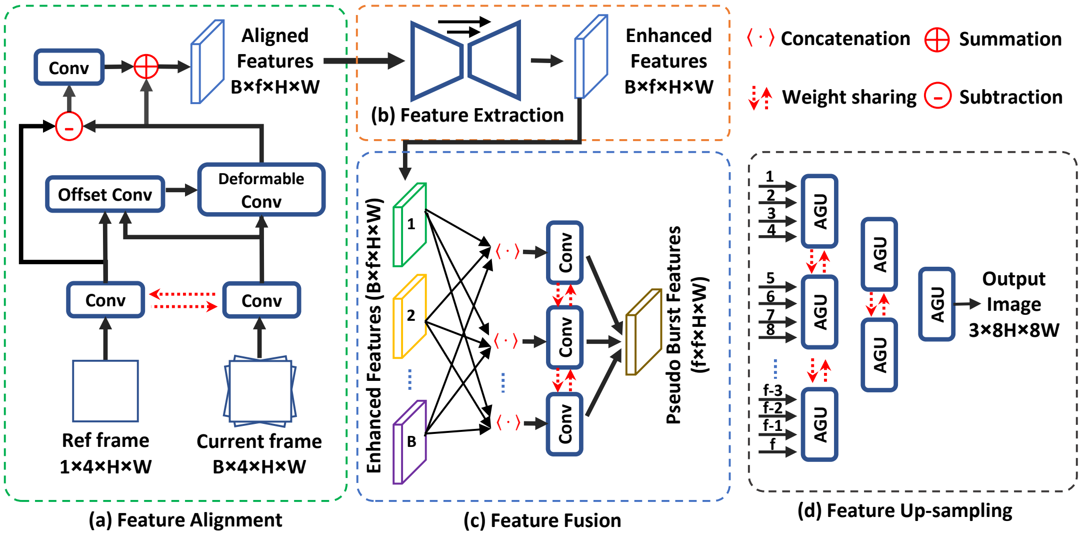


Figure 5.1: Overall pipeline of the proposed adaptive feature consolidation network (AFCNet) for burst SR. The proposed AFCNet processes input RAW burst image and generates a HR RGB image. It is divided into four parts: (a) Feature alignment module, (b) Feature extraction module, (c) Feature fusion module, and (d) Feature up-sampling module to produce HR RGB image.

1. We propose a simple but effective feature alignment module to align the burst image features with the base frame.

2. We utilise encoder-decoder based transformer backbone for feature extraction to enrich the aligned feature representations.

3. An efficient abridged pseudo-burst fusion module is utilized to aid inter-frame information exchange and feature consolidation.

4. Finally, adaptive group up-sampling is performed for progressive fusion and up-scaling of the burst features.

5.1.1 Proposed Method

On account of the rapid capture of images in a burst from a hand-held device, they inherit minute inter-frame offsets. This creates multiple aliased versions of the same scene, thus generating additional signal information for SR. Our proposed AFCNet processes multiple noisy, RAW, LR images to consider the merit of this shifted complementary information from multiple images and combines the information from individual LR images for generating HR RGB image as output. Our first challenge lies in alignment of the slight mismatches between multiple supporting frames and the reference frame. Following this, effective merging of the aligned features is equally important along with the reconstruction of HR image. In subsequent sub-sections, different modules of the proposed AFCNet are discussed.

Feature Alignment

The major hurdle in burst SR is the unknown inter-frame sub-pixel displacement. This displacement, stemming from camera motion and scene variations, results in misalignment among the frames [264]. Thus, to align the burst features with the reference frame, we utilized modulated deformable convolutions [266] as shown in Figure 5.1 (a). Considering, $\{\boldsymbol{x}^b\}_{b\in[1:S]} \in \mathbb{R}^{S\times n\times H\times W}$, as an initial representation of burst having S images and n number of feature channels. Currently, each frame feature x^b is concatenated with the reference frame feature x^{b_r} and passed through convolution layer to get the offsets and modulated scalars required for the deformable convolution layer. With the obtained offsets and modulated scalars, burst features x^b are processed through modulated deformable convolutions which returns the aligned burst features $\bar{x}^b$.
Our alignment module consists of three deformable layers for improving the overall alignment capability to enhance the aligned burst features. Unlike [268], we processed and aligned the burst features without any pre-processing. We combine it with the feature extraction module where we compute the local-global feature representations. This reduces the extra overhead on feature alignment module and simplifies the overall architecture. Further, high-frequency residue is evaluated by calculating the difference between these aligned features and reference frame features followed by its addition to the aligned features [268] to enhance the high-frequency edge information.

Feature Extraction

For further strengthening the feature alignment and to rectify small misalignment errors, we utilize a encoder-decoder based transformer backbone (EDTB) [267] for capturing global context information among various frames. Unlike [268], which employ feature refinement module to capture long-range dependencies for modelling global scene properties prior to aligning the features, we leverage a EDTB, after the aligned features as depicted in Figure 5.1 (b). EDTB processes the aligned features $\bar{x}^b$ and returns its enriched representation y^b. Following [267], we employ a 4-level encoder-decoder architecture with number of transformer blocks as [4, 6, 6, 8], attention heads in multi-head attention block are set to [1, 2, 4, 8], and number of channels are [64, 128, 256, 512], respectively.

Feature Fusion

For generating a merged feature embedding of the enriched aligned features, we designed an abridged pseudo-burst fusion (APBF) module inspired from [268]. It is a well proven fact that simple pooling operations like element-wise average or max pooling across the burst frames generates dissatisfying results [264]. The major reason tends to attribute towards the fact that fusion module requires adaptive merging on the basis of image content and noise levels. Furthermore, considering the benefits, and the indispensable role of inter-frame communication among the channels with multi-path network layout, for fusing the multi-frame features. We, thereby accomplish inter-frame connections through concatenation of the corresponding channel-wise burst feature maps and attain corresponding pseudo-bursts [268] as shown in Figure 5.1 (c). Given the refined features set $y = \{\boldsymbol{y}_c^b\}_{c\in[1:n]}^{b\in[1:S]}$ of burst size S and n number of channels, the pseudo-burst is generated through,

$$\boldsymbol{P}^c = W^\rho \left(\langle \boldsymbol{y}_c^1,\, \boldsymbol{y}_c^2,\, \cdots,\, \boldsymbol{y}_c^S \rangle\right), \quad s.t. \quad c \in [1:n], \tag{5.1}$$

where, $\langle\cdot\rangle$ represents feature concatenation, $\boldsymbol{y}_c^1$ is the c^{th} feature map of 1^{st} aligned burst feature set $\boldsymbol{y}^1$, W^ρ denotes the convolution layer with f output channel, and $\boldsymbol{P} = \{\boldsymbol{P}^c\}_{c\in[1:n]}$ represents the pseudo-burst of size $n \times n \times H \times W$. We have set n = 64 for this module.

Currently, every feature map in the pseudo-bursts embrace complimentary information from all the actual burst frame features. Apart from simplifying the learning task, the inter-frame feature representation merges the required information through decoupling of the burst feature channels. In [268], the aligned burst features are used to obtain the pseudo-bursts followed by the multi-scale feature extractor (encoder-decoder sub-module). In the proposed AFCNet, we abridge this process and directly process the set of enriched features obtained from feature extraction stage (EDTB module) to obtain pseudo-bursts. The proposed abridged pseudo-burst fusion (APBF) scheme serves the dual benefits of, (1) merging the consolidated feature information, and (2) avoiding the computational overhead of processing pseudo-bursts through heavy multi-scale module which is happening in [268].

Feature Up-sampling

The final step for reconstructing HR image is up-sampling. In AFCNet, we utilized the adaptive group up-sampling (AGU) [268] to reconstruct the HR details shown in Figure 5.1(d). AGU takes the feature maps (P^c) produced by the abridged pseudo-burst fusion module as input and generates a super-resolved output via three-level progressive upsampling. In AGU, the pseudo-burst features are sequentially divided into groups of 4. Being mindful of the benefits of applying different fusion weights to texture-less and edge regions, we ought to predict the fusion weights through an attention mechanism. To do so, we initially obtain a dense attention map for each pseudo-burst and subsequently apply element-wise multiplication with the corresponding dense attention map. This adaptively rescaled feature response is further passed through transposed convolution layer to up-sample and thus reconstruct the final HR image.

Since, for burst SR we need to perform ×8 up-sampling[1], we perform three levels, with

[1] The real task is to perform upsampling by ×4, additional ×2 is on account of mosaicked RAW LR

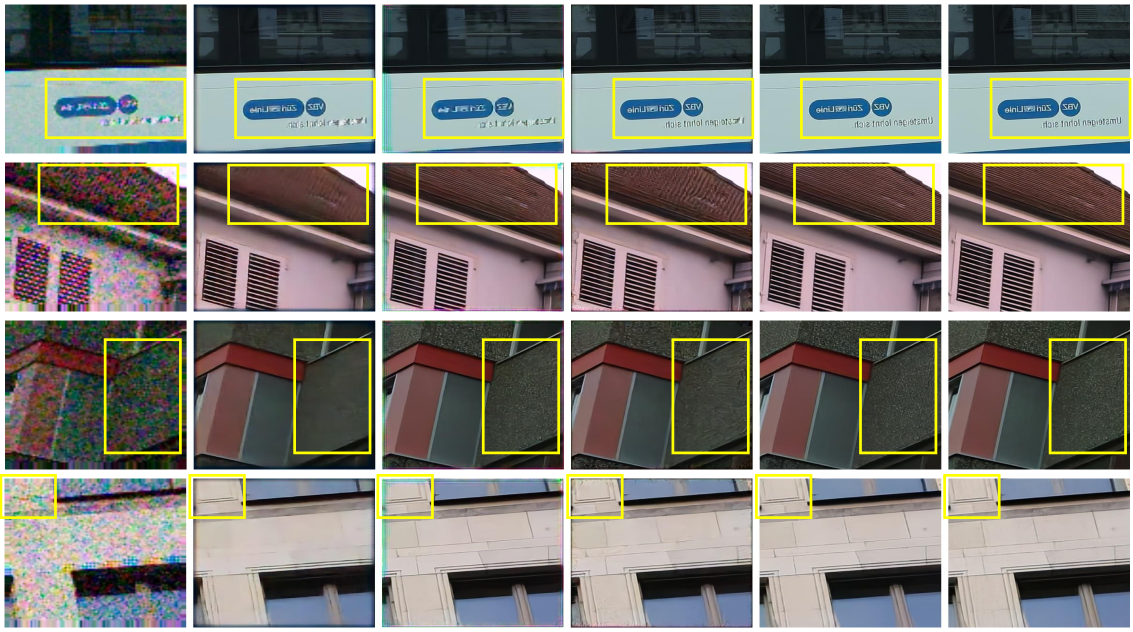


Figure 5.2: Comparisons for ×4 burst super-resolution on SyntheticBurst dataset [71] (NTIRE-21 Track 1).

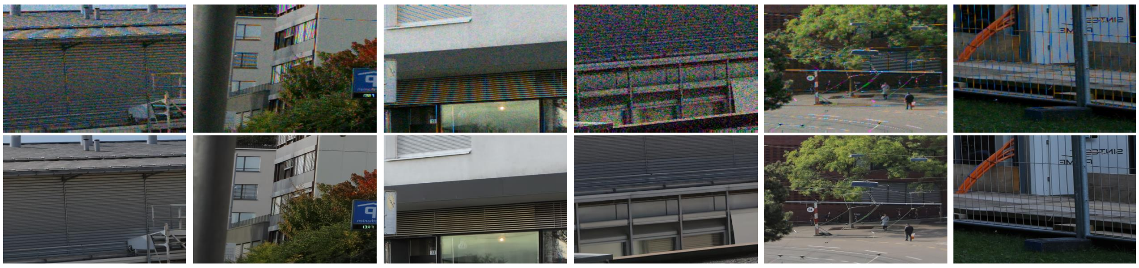


Figure 5.3: Comparisons on SyntheticBurst dataset [72] for ×4 burst super-resolution (NTIRE-22 Track 1). First and second rows depicts results of base frame up-scaled using bilinear interpolation and the proposed AFCNet respectively.

each level performing up-sampling (×2). As we have 64 pseudo-bursts, for three levels of AGU, naturally it forms a group of 16, 4, 1 pseudo-bursts group.

5.1.2 Experimental Analysis

We evaluate the proposed AFCNet for both synthetic as well as real burst SR task. We follow the NTIRE-21 [71] and NTIRE-22 [72] competition guidelines to carry out network training and testing.

Implementation details

Our AFCNet is a single end-to-end trainable network designed for burst SR and requires no pre-training of the proposed module. For overall network efficiency, all burst frames have been processed through shared AFCNet modules. AFCNet has been trained for 100 epochs

frames.

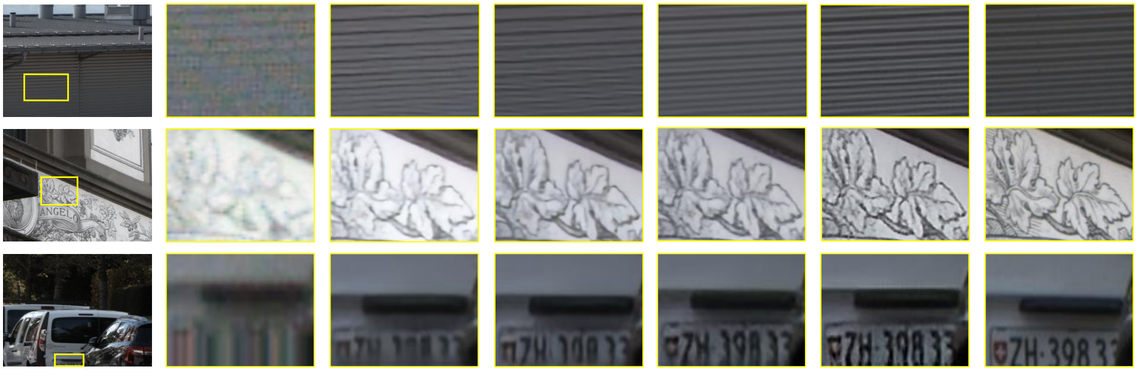


Figure 5.4: Visual Comparisons for ×4 burst super-resolution on Real BurstSR dataset [71] (NTIRE-21 Track 2).

on synthetic bursts generated by utilising 46,839 sRGB images from Zurich-RAW-to-RGB dataset [269]. We train AFCNet for burst SR task using L_1 loss only. While for real burst SR, we fine-tune our AFCNet with pre-trained weights on SyntheticBurst dataset using aligned L_1 loss [270]. The models are trained with Adam optimizer. Cosine annealing strategy [271] is deployed for steadily decreasing the learning rate from 10^{-4} to 10^{-6} during training. We augment our dataset using horizontal and vertical flips. It should be noted that unlike [272], we have not employed any kind of ensemble techniques to boost the evaluation metrics.

SyntheticBurst dataset (NTIRE-21 Track 1)

It consists of 300 RAW bursts for validation. Each burst contains 14 LR RAW images (each of size 48×48 pixels) that are synthetically generated from a single sRGB image [264]. Table 5.1 shows the quantitative evaluation on SyntheticBurst dataset [71]. Also, we have shown the visual comparison between the proposed and existing state-of-the-art (SoTA) methods for ×4 burst SR task in Figure 5.2. From Table 5.1 and Figure 5.2, it is clear that the proposed AFCNet outperforms other existing SoTA methods for ×4 burst SR task.

Real BurstSR dataset (NTIRE-21 Track 2)

It consists of 5,405 and 882 patches for training and validation, respectively cropped from 200 real RAW bursts images. Each input crop has a size of 80×80 pixels. As shown in Table 5.1, the proposed AFCNet performs favorably well when compared to the other existing SOTA for ×4 real burst SR task. Also, Figure 5.9 demonstrates that HR images produced by the AFCNet for ×4 are sharper with vivid details as compared to the other existing SoTA.

Table 5.1: Performance assessment on SyntheticBurst and real BurstSR validation datasets (NTIRE-21) [71] for ×4 burst super-resolution.

Methods	**SyntheticBurst**		**(Real) BurstSR**	
	PSNR ↑	**SSIM ↑**	**PSNR ↑**	**SSIM ↑**
Single Image	36.17	0.91	46.29	0.982
HighRes-net [273]	37.45	0.92	46.64	0.980
DBSR [264]	40.76	0.96	48.05	0.984
LKR [265]	41.45	0.95	-	-
MFIR [270]	41.56	0.96	48.33	0.985
BIPNet [268]	41.93	0.96	48.49	0.985
AFCNet (Ours)	**42.21**	**0.96**	**48.63**	**0.986**

Table 5.2: Performance evaluation on validation and test set of SyntheticBurst dataset (NTIRE-22 Track 1) [72] for ×4 burst super-resolution.

Methods	**Validation set**		**Test set**	
	PSNR ↑	**SSIM ↑**	**PSNR ↑**	**SSIM ↑**
Baseline [268]	42.24	0.97	-	-
AFCNet (Ours)	42.44	0.97	42.08	0.97

SyntheticBurst dataset (NTIRE-22 Track 1)

It consists of 100 and 92 RAW bursts in validation and test set respectively. Each RAW burst contains 14 LR RAW images (each of size 256×256 pixels) synthetically synthesized from a single sRGB image [264]. Table 5.2 summarises the quantitative evaluation on validation and test dataset of the proposed AFCNet in comparison with the baseline approach on SyntheticBurst dataset [72]. While Figure 5.3 display the visual results produced by the proposed AFCNet for RAW bursts from validation set. Figure 5.3 shows the ability of the proposed AFCNet in producing HR images with enriched details. We have not fine-tuned the proposed AFCNet for this experiment and we directly tested the network trained on the training set.

Table 5.3: Significance of AFCNet modules assessed on SyntheticBurst validation set [71] for ×4 burst SR task.

Modules	**A1**	**A2**	**A3**	**A4**	**A5**	**A6**
Baseline	✓	✓	✓	✓	✓	✓
Alignment (§5.1.1)		✓	✓	✓	✓	✓
Back-projection (§5.1.1)			✓	✓	✓	✓
EDTB (§5.1.1)				✓	✓	✓
APBF (§5.1.1)					✓	✓
AGU (§5.1.1)						✓
PSNR	36.38	38.92	39.50	41.20	41.80	42.21

5.1.3 Ablation Study

In this section, we demonstrate the importance of each module in our proposed AFCNet. Commencing from our base network, we introduce different network models systematically, and exhibit its performance for burst SR application. Every network combination has been trained for 100 epochs on the training set discussed in Section 5.1.2. Table 6.3 marks all the ablation experiments conducted for ×4 burst SR task on validation set of Zurich-RAW-to-RGB dataset [269]. For the baseline model, we employ Resblocks [274] as our feature extraction module, simple concatenation operation has been deployed as a fusion module, and we used transposed convolution for upsampling. The baseline network obtains 36.38 dB PSNR. After appending the proposed modules to the baseline, their seem to be a significant and consistent improvement in results. For example, inclusion of alignment module and back projection approach improves the PSNR by 2.54 and 0.58 dB respectively. While, feature extraction stage which is composed of EDTB [267] achieves significant gain of 1.70 dB in the performance. Inclusion of APBF module contributes improvement of 0.60 dB whereas, adaptive group up-sampling block takes the gain to 42.21 dB. Overall, our AFCNet attains a captivating performance gain of 5.83 dB over the baseline.

5.2 Burst Restoration and Enhancement

Three critical factors involved in burst processing are feature alignment, fusion, and subsequent reconstruction of the obtained frames. Generally, any burst processing approach is limited by the accuracy of alignment process on account of the camera and scene motion of dynamically moving objects. Therefore, it is crucial to design a module for facilitating accurate alignment, as the subsequent fusion and reconstruction modules must be robust to misalignment for generating an artifact-free image. We further note that the alignment and fusion modules in existing burst processing approaches [75, 74] do not consider the non-local dependencies and mutual correlation among the frames which hinders the flexible inter-frame information exchange. Moreover, the existing burst up-sampling approaches [7, 75] do not take into account the merits of repeatedly transferring the information across several resolutions. To address these issues, we present a novel burst processing framework named Gated Multi-Resolution Transfer network (GMTNet).

In contrast to the previous works [7, 74] which adopt bulky pre-trained modules for alignment, we propose an implicit Multi-scale Burst Feature Alignment (MBFA) to reduce the inter-frame misalignment. Overall, MBFA module implicitly learns feature alignment at multiple scales through the proposed Attention-Guided Deformable Alignment (AGDA) module and obtains an enriched feature representation via Aligned Feature Enrichment (AFE) module. The proposed AFE module is composed of a back-projection mechanism and capable of extracting long-range pixel interactions that ease the feature alignment in complex motions, where simply aligning the frames does not suffice. Additionally, unlike the recent state-of-the-art (SoTA) algorithm, BIPNet [75] that utilizes a computationally intensive pseudo burst mechanism on the aligned burst for inter-frame communication, we propose a simple Transposed-Attention based Feature Merging (TAFM) module that leverages local and non-local correlations to allow an extensive interaction with the reference frame. Finally, our Resolution Transfer Feature Up-sampler (RTFU) combines the complementary features of both single-stage and progressive up-sampling strategies through deployed conventional and recent feature up-samplers. Such a design enables strong feature embedding of LR and HR images that creates a solid foundation for up-sampling in burst SR tasks. In this work, we validate our GMTNet for popular burst processing tasks such as super-resolution, denoising and low-light image enhancement. Overall, the following are our key contributions:

1. A Multi-scale Burst Feature Alignment (MBFA) is proposed which uses both local and non-local features for alignment at multiple scales, resolving the spatial misalignment within burst images (§5.2.1).
2. A Transposed-Attention Feature Merging (TAFM) is proposed to aggregate the

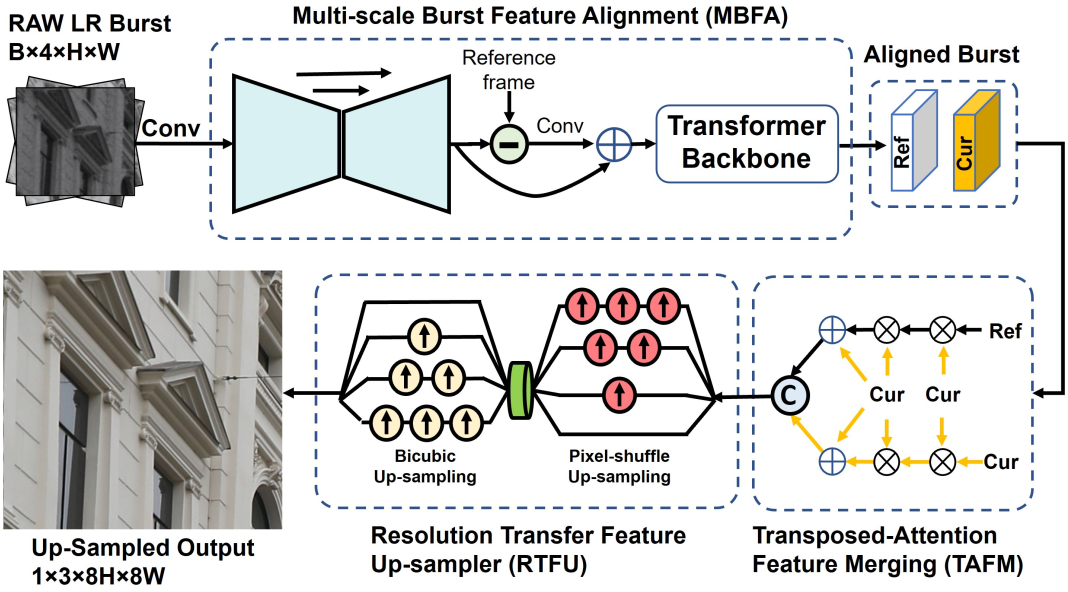

RAW LR Burst
B×4×H×W
Conv
Multi-scale Burst Feature Alignment (MBFA)
Reference frame
Conv
Transformer Backbone
Aligned Burst
Ref
Cur
Ref
Cur
Cur
Cur
C
Bicubic Up-sampling
Pixel-shuffle Up-sampling
Up-Sampled Output
1×3×8H×8W
Resolution Transfer Feature Up-sampler (RTFU)
Transposed-Attention Feature Merging (TAFM)

Figure 5.5: Overview of the proposed GMTNet for burst processing.

features of the aligned and reference frames (§5.2.1).

3. A Resolution Transfer Feature Up-sampler (RTFU) is proposed to upscale the merged features. The proposed RTFU integrates the complementary features extracted by single-stage and progressive up-sampling strategies using the conventional and recent up-samplers (§5.2.1).

5.2.1 Proposed Method

We present the overall pipeline of our burst processing approach in Figure 5.5. Given a raw burst image, the goal of our GMTNet is to reconstruct a clean, high-quality image by exploiting the shifted complementary information from the noisy LR image burst. As shown in Figure 5.5, the input RAW LR burst features are aligned to the reference frame through our proposed **M**ulti-scale **B**urst **F**eature **A**lignment (MBFA) module. Further, aligned burst features are aggregated using the **T**ransposed-**A**ttention **F**eature **M**erging (TAFM) module. Lastly, our **R**esolution **T**ransfer **F**eature **U**p-sampler (RTFU) up-scales the merged features to reconstruct a high-quality image.

Multi-scale Burst Feature Alignment

Generating an artifact-free, high-quality image through burst processing is highly reliant upon the alignment of the mismatched burst frames. However, proper alignment is quite challenging, specifically in low-light and low-resolution images, where noise excessively contaminates the input burst frames. Previous burst restoration methods [75, 275, 7, 74] often seek to alleviate these issues by following alignment on locally extracted features. However, they do not explicitly consider the long-range dependencies which are crucial for

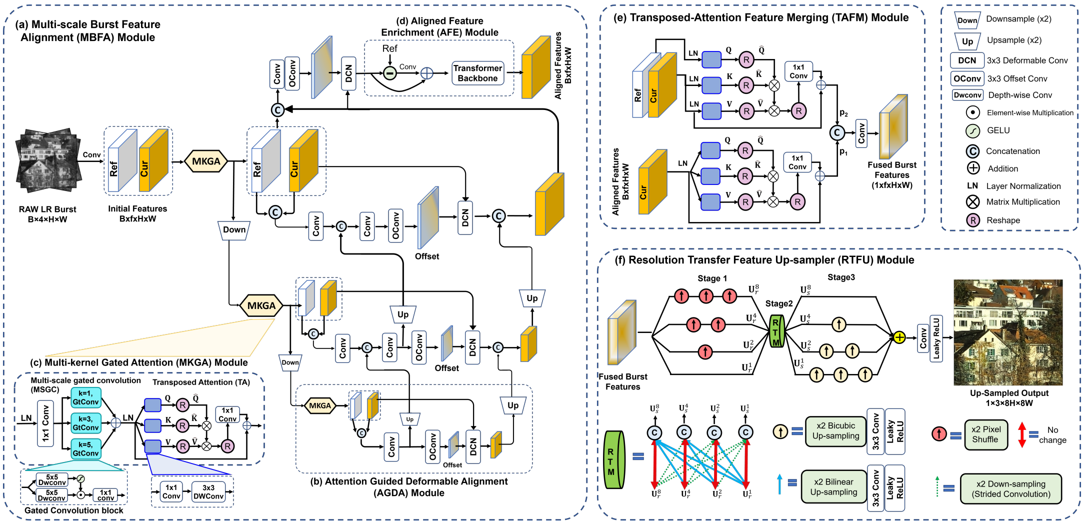


Figure 5.6: Comprehensive representation of each stage of our proposed GMTNet: (a) The proposed Multi-Scale Burst Feature Alignment (MBFA) module, (b) Attention-Guided Deformable Alignment (AGDA), (c) Multi-Kernel Gated Attention (MKGA) module, (d) Aligned Feature Enrichment (AFE) module, (e) Transposed Attention Feature Merging (TAFM) module, and (f) Resolution Transfer Feature Up-sampler (RTFU).

restoration tasks. Consequently, the generated feature maps have limited receptive field making it difficult to align the burst features in case of complex motions. In order to handle the aforementioned issues, we design a new module termed as Multi-scale Burst Feature Alignment (MBFA) that not only aligns burst features at multiple scales but also captures the long-range pixel interactions which eventually ease the alignment process. Figure 5.6 (a) shows our MBFA operates in two stages: Firstly, it aligns the burst features at multiple scales through the proposed Attention-Guided Deformable Alignment (AGDA) module. Secondly, it refines the aligned burst features through Aligned Feature Enrichment (AFE) module.

Attention-Guided Deformable Alignment

As discussed in [276], noise disturbs the prediction of dense correspondences among multiple frames which is the key concern of several alignment methods. However, we find that a well-designed module can easily tackle noisy raw data. Therefore, in order to reduce the noise content in the initial burst features and eventually ease the alignment process, we propose an Attention-Guided Deformable Alignment (AGDA) module that operates at multiple scales to align the burst features as shown in Figure 5.6 (b). The proposed AGDA module is inspired from the deformable alignment proposed in TDAN [277] and EDVR [275]. But, their alignment approaches [277, 275] directly apply deformable convolution on the input features, *making them prone to miss the detailed information in case of noisy RAW burst features.* Additionally, they also lack at extracting long-range pixel interactions which are useful in complex motions. To tackle these problems, *we enhance* *the overall representation of the incoming noisy burst features at each scale* through the proposed Multi-kernel Gated Attention (MKGA) module *(see Figure 5.6 (c))*. Further, the denoised burst features are aligned through the modulated deformable convolution (DCN) as shown in Figure 5.6 (b).

Multi-Kernel Gated Attention.

The proposed MKGA module is designed to emphasize on learning flexible receptive field via Multi-Scale Gated Convolution (MSGC) sub-module and thereafter it learns the non-local spatial and inter-channel dependencies via the transposed-attention (TA) sub-module as demonstrated in Figure 5.6 (c). Given an input tensor $\mathbf{Y} \in \mathbb{R}^{C\times H\times W}$, the overall operation of MSGC, outputting $\hat{\mathbf{Y}}$, is formulated as:

$$\hat{\mathbf{Y}} = W_1 * (G_1(\mathbf{Y})) + W_1 * (G_3(\mathbf{Y})) + W_1 * (G_5(\mathbf{Y})) \tag{5.2}$$

Here, W_1 denotes a convolution filter with size 1×1, and * is a convolution operation. $G_k(\mathbf{Y})$ represent the output of the Gated Convolution block *(See Figure 5.6 (c))*, that is mapped out as the element-wise product of two parallel paths for depth-wise convolution layers with filter size k and formulated as $G_k(\mathbf{Y}) = \lambda(W_k^{dep}) \odot W_k^{dep}$. Here, W_k^{dep} denotes a depth-wise convolution layer, λ and $\odot$ represents the GELU non-linearity, and element-wise multiplication, respectively.

Transposed Attention.

The extracted multi-kernel features from the MSGC module are passed through the transposed attention (TA) sub-module *(see Figure 5.6 (c))* for capturing their long-range pixel interactions. From a layer normalized tensor $\hat{\mathbf{Y}}$, our TA sub-module first generates query $(\mathbf{Q})$, key $(\mathbf{K})$, and value $(\mathbf{V})$ projections by applying 1×1 convolutions followed by 3×3 depth-wise convolutions for encoding the non-local and channel-wise spatial context. Thereafter, we reshape $(\mathbf{Q},\mathbf{K},\mathbf{V})$ into $\hat{\mathbf{Q}}$, $\hat{\mathbf{K}}$ and $\hat{\mathbf{V}}$ projections such that the subsequent dot-product interactions between query and key generate a transposed-attention map of size $\mathbb{R}^{C\times C}$[267], instead of the huge regular attention map of size $\mathbb{R}^{HW\times HW}$ [278]. And, the overall TA process, outputting $\tilde{\mathbf{Y}}$, is defined as:

$$\begin{aligned} \tilde{\mathbf{Y}} &= LN(\hat{\mathbf{Y}}) + W_1 * (TA(\hat{\mathbf{Q}}, \hat{\mathbf{K}}, \hat{\mathbf{V}})); \\ TA(\hat{\mathbf{Q}}, \hat{\mathbf{K}}, \hat{\mathbf{V}}) &= \hat{\mathbf{V}} \otimes S(\hat{\mathbf{K}} \otimes \hat{\mathbf{Q}}) \end{aligned} \tag{5.3}$$

Here, $\hat{\mathbf{Y}}$ is the feature map obtained from the MSGC module, LN denote the layer normalization; TA and S denotes the operation of the TA sub-module and Softmax, respectively, $\hat{\mathbf{Q}} \in \mathbb{R}^{HW\times C}$, $\hat{\mathbf{K}} \in \mathbb{R}^{C\times HW}$, and $\hat{\mathbf{V}} \in \mathbb{R}^{HW\times C}$ matrices are obtained after reshaping the tensors from the original size, $\mathbb{R}^{C\times H\times W}$, and $\otimes$ denotes matrix multiplication. Altogether, the employed MKGA module at each scale allows each pyramidal level to focus on fine details, generating contextualized features that reduce noise and thus ease the subsequent alignment mechanism.

Modulated Deformable Convolution.

After extracting the features from the MKGA module, we implicitly align the current frame features, $\boldsymbol{f}^{b}$ with the reference frame features (*we considered the first frame as reference*), $\boldsymbol{f}^{b^{r}}$ via modulated deformable convolution [266, 277] (learnable offsets for deformable convolution layer are obtained through a 3×3 offset convolution layer) as shown in Figure 5.6 (b). To ensure better learning, the predicted offsets and aligned burst features are shared from the lower-scale to upper-scale in a bottom-up fashion to ensure semantically stronger and cleaner aligned features.

Aligned Feature Enrichment

To fix the remaining minor alignment and noise issues, we embed a novel Aligned Feature Enrichment (AFE) module on the obtained aligned features. The proposed AFE module enhances the aligned burst features by boosting the high-frequency content via a simple back-projection process followed by extracting the local-non-local features. During the back-projection process, we simply compute the high-frequency residue between the aligned burst features and reference frame as shown in Figure 5.6 (d). Thereafter, the local-non-local pixel interactions are enabled by processing the aligned edge boosted burst features through the existing transformer backbone [267]. In a nutshell, besides capturing multi-scale local-global representation among the bursts, the AFE module also bridges the gap between the relevant and irrelevant features of the aligned frames.

Transposed-Attention Feature Merging

In burst processing, temporal relation among the multiple frames plays an indispensable role in feature fusion on account of blurry frames from camera perturbations. Considering the fact, that incoming multiple frames have quite a few similar patterns at the feature level, it is infeasible to directly concatenate or add them as it will naively introduce a large amount of redundancy into the network. Existing DBSR [7] proposed an attention-based fusion approach but it is limited in exploiting the complementary (global and local) relations that can hinder the information exchange among multiple frames. Further, the recently proposed BIPNet [75] tries to merge the relevant information by concatenating channel-wise features from all burst feature maps. Though it is effective in extracting complementary information, it is computationally extensive.

Unlike the aforementioned fusion techniques, we propose a Transposed-Attention Feature Merging (TAFM) to efficiently encode *local and non-local correlations before merging the frames.* As shown in Figure 5.6 (e), TAFM takes queries **(Q)** and a set of key-value (**K**,**V**) pairs as input and outputs the linear combination of values that are determined by correlations between the queries and corresponding keys [216]. The proposed TAFM module has been designed with two parallel blocks (*see Figure 5.6 (e)*), where the lower block (outputting p_1) performs the query-key interactions across channels of the aligned neighboring frames to encode the channel-wise local context. While the upper block (outputting p_2) enhances the feature representations of the reference and current frames by bridging their global correlations. After encoding the feature correlations globally and locally for a given aligned frame, $\bar{\boldsymbol{f}}^{\boldsymbol{b}}$ with b number of burst frames, the overall merged features of TAFM, $\boldsymbol{F_m} \in \mathbb{R}^{1\times C\times H\times W}$ are obtained as follows:

$$\boldsymbol{F_m} = W_3 * (p_1 \textcircled{C} p_2) \tag{5.4}$$

where, W_3 is a convolution layer with filter size 3×3, and $\textcircled{C}$ refers to the concatenation.

Resolution Transfer Feature Up-sampler

The popular up-sampling techniques deployed in SoTA burst SR methods DBSR [7], DRSR [74] perform direct one-stage up-sampling without leveraging the benefits of information exchange between the HR features and their corresponding LR counterparts. Considering the fact that HR features contain abundant global information and LR features are rich in edge information [279], we design a Resolution Transfer Feature Up-sampler (RTFU) module to extract *unique features* of *different resolution spaces.* The proposed RTFU module stems from the observation that the transfer of LR and HR features through a multi-resolution framework can be propitious in adaptively recovering the textural information from the fused frames as shown in the ablation study. In RTFU, we target at exploiting the dual benefits of both direct [154] and progressive up-sampling [280] strategies using the conventional [281] and recent learnable up-sampling layers [282] to adequately get into the HR space. As shown in Figure 5.6 (f), the proposed RTFU achieves its desired HR feature space via a three-stage design: two sets of four parallel progressive multi-resolution streams (Stage1 and Stage3) and a Resolution-Transfer Merging (RTM) module (Stage2).

We first apply progressive up-sampling strategy with pixel-shuffle [282] (*extreme left of Figure 5.6 (f)*) parallelly in Stage1 for generating (×1, ×2, ×4, and ×8) multi-resolution SR feature responses, which are then forwarded to the RTM module (Stage2). RTM module consists of four input representations: U_r^i (output of Stage1), i = 1, 2, 4, and 8 with i being the input resolution index, and the associated output representations are given by U_s^o, o = 1, 2, 4, and 8 with o being the output resolution index. Each output representation (U_s^o) is the concatenation of the transformed representations of the corresponding four inputs *(as shown in the middle of Figure 5.6 (f))*. Thus, the overall operation of Stage2 (RTM module) can be formulated as follows:

$$U_s^o = [[f(U_r^i)]]_{i=1,2,4,8}; \; f = \begin{cases} 1 \;\; \forall \;\; i = o \\ \frac{o}{i} \uparrow \;\; \forall \;\; o > i \\ \frac{i}{o} \downarrow \;\; \forall \;\; o < i \end{cases} \tag{5.5}$$

Here, $o \in \{1, 2, 4, 8\}$, and the mathematical definition of the symbol used in Eq. 5.5 is given as $[[A^j]]_{j=1,2,...n} = A^1 \textcircled{C} A^2 ..... \textcircled{C} A^n$, where $\textcircled{C}$ denotes the concatenation operation among the inputs, and f represents the corresponding transformation operation (upsample or downsample) applied to the input feature U_r and is dependent upon the input resolution index (i), and the output resolution index (o). For instance, as shown above, if $o > i$, then the corresponding input representation U_r^i is up-sampled $(\uparrow)$ by a factor of o/i. In Stage2 (RTM), we deploy bilinear interpolation and strided convolution for feature up-sampling and down-sampling, respectively. Thereafter, the resulting features from each branch of Stage3 are again up-sampled progressively using bicubic interpolation to generate an up-sampled feature map of the size $\mathbb{R}^{1 \times C \times 8H \times 8W}$. Finally, we add the individual branch output of Stage3 to generate the final high-quality image. Thus, for each pixel location, RTFU can leverage the underlying content information from input frames at multiple-scales and utilize it to get better performance than the mainstream up-sampling operations, pixel-shuffle or interpolations.

5.2.2 Experimental Analysis

We validate the proposed GMTNet on real and synthetic datasets for **(a)** Burst Super-resolution, **(b)** Burst denoising, and **(c)** Burst low-light image enhancement tasks.

Implementation Details

We train separate models for all the considered tasks in an end-to-end manner. For better parameter efficiency, we shared each GMTNet module for all burst frames. Our GMTNet has 12.7M parameters with 207 GFLOPs for the burst of size 8×4×128×128 with a run time of 24 fps. All the models are trained with Adam optimizer with L_1 loss function. We employ cosine annealing strategy [271] to decrease the learning rate from 10^{-4} to 10^{-6} during training. For real-world SR, we fine-tune our GMTNet (*with pre-trained weights on SyntheticBurst dataset*) using aligned L_1 loss [7]. We provide the task-specific experimental details in the corresponding sections.

Burst Super-Resolution

We evaluate our proposed GMTNet on synthetic [7] and real-world datasets [7] for scale factor ×4. Following the settings in [7], we utilized **SyntheticBurst** dataset (46,839 and 300 RAW burst sequences for training and validation respectively, where each burst sequence consists of 14 images), and **BurstSR** dataset consisting of 200 RAW burst sequences (5,405 and 882 patches of size 80×80 for training and validation, respectively).

Table 5.4: **Burst super-resolution** results on synthetic and real-world datasets for ×4.

Methods	**SyntheticBurst** [7]		**BurstSR** [7]	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑
SingleImage	36.86	0.919	46.60	0.979
WMKPN [283]	36.56	0.912	41.87	0.958
HighResNet [73]	37.45	0.924	46.64	0.980
DBSR [37]	40.76	0.959	48.05	0.984
MFIR [74]	41.56	**0.964**	48.33	0.985
BIPNet [75]	41.93	0.960	48.49	0.985
Ours	**42.36**	0.961	**48.95**	**0.986**

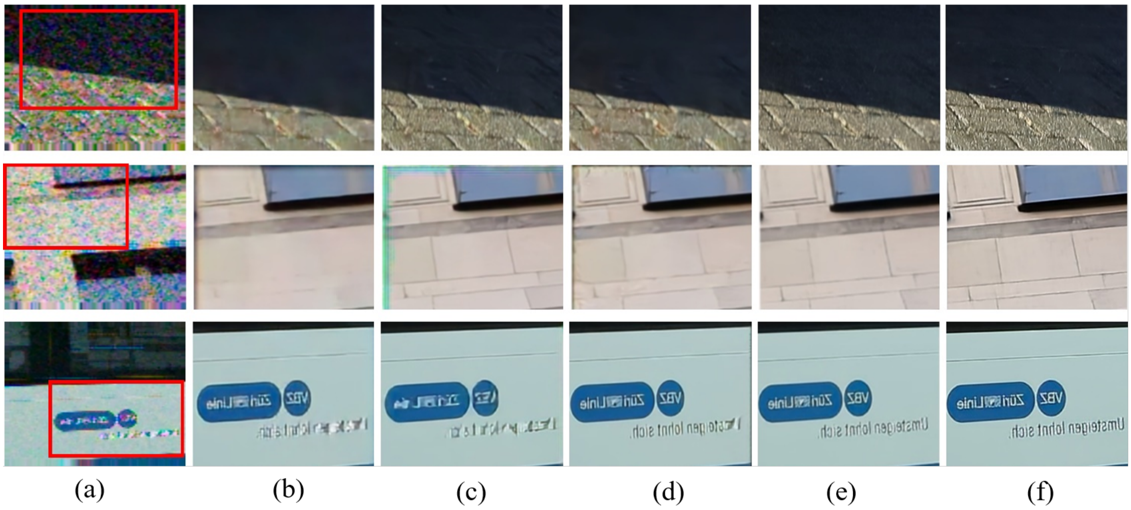


Figure 5.7: Visual results on SyntheticBurst [7] for ×4 burst SR, where (a) Base frame, (b) DBSR [37], (c) LKR [73], (d) MFIR [74], (e) BIPNet [75], and (f) Ours.

SR results on SyntheticBurst dataset for ×4 and ×8.

The proposed GMTNet is trained for 300 epochs on the training split of SyntheticBurst dataset for both ×4, and ×8 up-sampling tasks and evaluated on the validation set of SyntheticBurst dataset [7]. We compared our proposed GMTNet with several SoTA approaches for ×4 as shown in Table 5.4. Particularly, our GMTNet obtains a PSNR gain of about 0.43 dB over the previously best-performing BIPNet [75] and 0.80 dB over the second-best approach [74]. To further prove the potency of our proposed GMTNet on large scale factors, we conduct an experiment for ×8 burst SR. The LR-HR pairs are synthetically generated using the same procedure described for SyntheticBurst dataset [7]. Visual results shown for a few challenging images in Figure 5.7 (×4) and Figure 5.8 (×8) clearly prove that results obtained by GMTNet are sharper and it efficiently reconstructs the structural content and fine textures, without compromising details.

SR results on BurstSR dataset.

Since, the LR-HR pairs for BurstSR dataset are captured using different cameras, they suffer from minor misalignment. Thus we follow the previous work [7] and use aligned L_1 loss for fine-tuning the GMTNet for 25 epochs and evaluate our model by using aligned

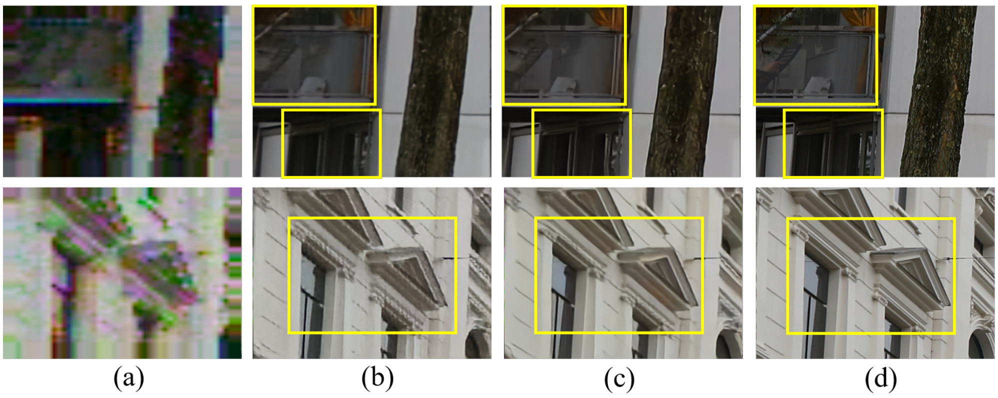


Figure 5.8: Visual results on SyntheticBurst [7] for ×8 burst SR, where (a) denotes base frame, (b) BIPNet [75], (c) Ours, and (d) Ground-truth.

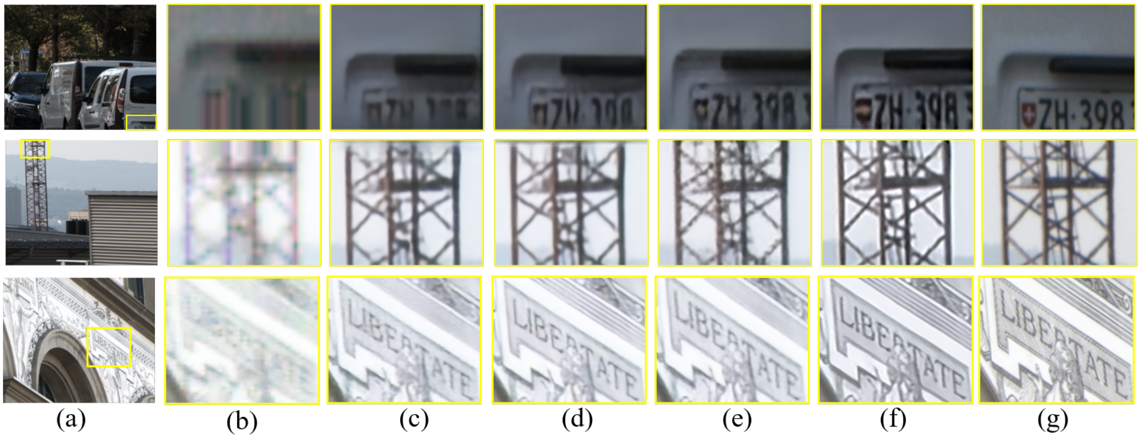


Figure 5.9: Results on real BurstSR dataset [7] for ×4 burst SR, where (a) HR Image, (b) Base frame, (c) DBSR [37], (d) MFIR [74], (e) BIPNet [75], (f) Ours, and (g) Ground-truth.

PSNR/SSIM. Table 5.4 shows that our proposed GMTNet obtain conducive results, outperforming SoTA BIPNet [75] by a substantial gain of 0.46 dB. Visual comparisons in Figure 5.9 depict that unlike other compared methods, the proposed GMTNet is more effective for generating minute details in the reconstructed images, with better color and structure preservation.

Table 5.5: **Gray-scale burst denoising** [12] results with PSNR.

Methods	Gain $\propto$ 1	Gain $\propto$ 2	Gain $\propto$ 4	Gain $\propto$ 8	Average
KPN [12]	36.47	33.93	31.19	27.97	32.19
BPN [11]	38.18	35.42	32.54	29.45	33.90
BIPNet [75]	38.53	35.94	33.08	29.89	34.36
MFIR [7]	**39.37**	**36.51**	33.38	29.69	34.74
Ours	39.07	36.46	**33.52**	**30.46**	**34.87**

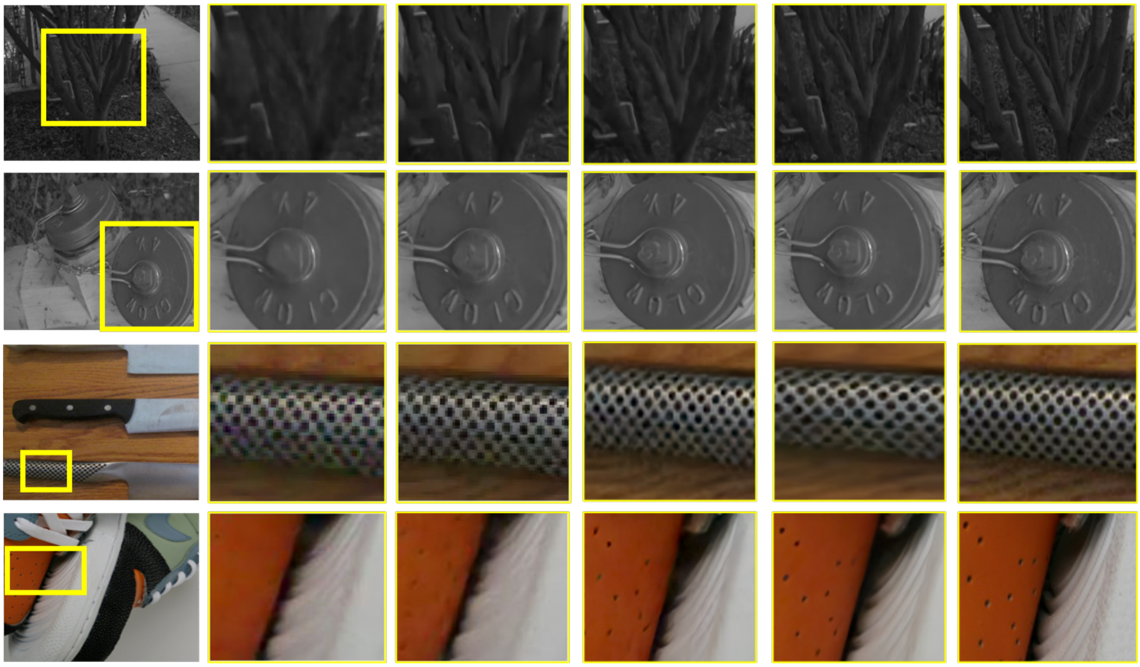


Figure 5.10: Visual results on gray-scale datasets [12] (first two rows) and color [11] (last two rows) for burst denoising, where (a) Noisy image, (b) BPN [11], (c) MFIR [7], (d) BIPNet [75], and (e) GT-Image.

Burst Denoising Results

This section presents the results of burst denoising on color (test split: 100 bursts) [11] as well as gray-scale (test split: 73 bursts) [12] datasets. Both these datasets have four variants with different noise gains (1, 2, 4, 8), corresponding to noise parameters $(\log(\sigma_r), \log(\sigma_s)) \rightarrow$ (-2.2, -2.6), (-1.8, -2.2), (-1.4, -1.8), and (-1.1, -1.5), respectively. We train separate models for grayscale and color burst denoising for 200 epochs on 20k synthetic noisy burst samples generated using the process described in [74].

Denoising results

Table 5.5 shows the results on the gray-scale burst denoising dataset against SoTA methods. Our GMTNet outperforms the recent BIPNet[2] [75] by about 0.57 dB for the highest noise gain (Gain $\propto$ 8). Similarly, for color denoising, our approach outperforms existing MFIR [7] on all four noise levels (except the lowest noise gain) with an average margin of 0.25 dB as shown in Table 5.6. Qualitative comparison in Figure 5.10 clearly proves the efficacy of our approach in recovering the required subtle contextual details, thus generating cleaner denoised outputs.

[2] Existing BIPNet results are collected from their official GitHub repository.

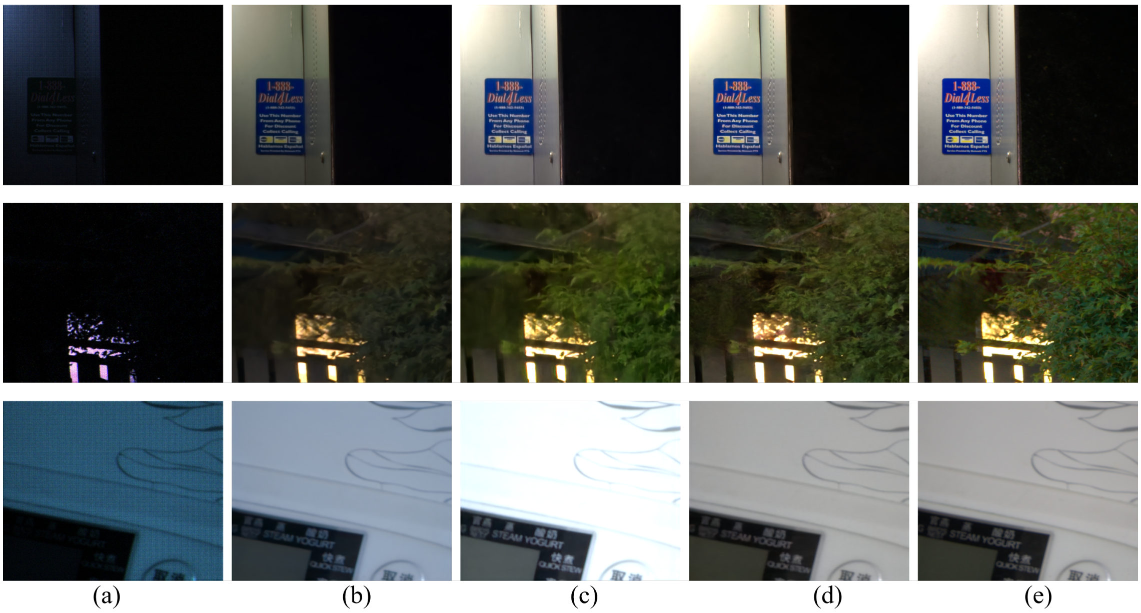


Figure 5.11: Visual results on SONY-subset of SID dataset [76] for burst low-light image enhancement, where (a) Input low-light patch, (b) Kardeniz *et al.* [77], (c) BIPNet [75], (d) Proposed Method, and (e) Ground-truth.

Table 5.6: **Color burst denoising** [11] results with PSNR.

Methods	Gain $\propto$ 1	Gain $\propto$ 2	Gain $\propto$ 4	Gain $\propto$ 8	Average
KPN [12]	38.86	35.97	32.79	30.01	34.40
BPN [11]	40.16	37.08	33.81	31.19	35.56
BIPNet [75]	40.58	38.13	35.30	32.87	36.72
MFIR [7]	**41.90**	38.85	35.48	32.29	37.13
Ours	41.74	**38.91**	**35.74**	**33.09**	**37.38**

Low-Light Enhancement Results

Following other existing works [75, 77], we test the performance of our GMTNet on the SONY-subset from the SID dataset [76]. It contains 161 input RAW burst sequences for training, 36 for validation, and 93 for testing. We train the proposed GMTNet with L_1 loss for 200 epochs on 5000 cropped patches of size 256×256 from the training set of SONY-subset. Table 5.7 gives the image quality scores for several competing approaches. The proposed GMTNet provides 0.26 dB improvement over the existing best BIPNet [75]. Visual comparisons in Figure 5.11 show that the enhanced images are relatively cleaner, sharper and preserves more structural content than other compared approaches.

5.2.3 Ablation Study

Here we analyze the influence of every key component and design choice in our formulation. All models are trained for 100 epochs on SyntheticBurst dataset [7] for ×4 burst SR task. As reported in Table 5.8, the baseline model achieves a PSNR of 36.38 dB. For the baseline

Table 5.7: **Burst low-light enhancement** on Sony-subset [76].

Methods	PSNR ↑	SSIM ↑	LPIPS ↓
Chen *et al.*[76]	29.38	0.89	0.48
Maharjan *et al.* [284]	29.57	0.89	0.48
Zamir *et al.* [285]	29.13	0.88	0.46
Zhao *et al.* [286]	29.49	0.89	0.45
Karadeniz *et al.* [77]	29.80	0.89	0.30
BIPNet [75]	32.87	0.93	**0.30**
Ours	**33.13**	**0.94**	0.31

Table 5.8: Ablation study for GMTNet contributions. PSNR is reported on SyntheticBurst dataset [7] for ×4 burst SR task.

Task	**Modules**								
	Baseline	✓	✓	✓	✓	✓	✓	✓	✓
Align ment (§5.2.1)	w/O MKGA		✓						
	with MKGA			✓	✓	✓	✓	✓	✓
	AFE				✓	✓	✓	✓	✓
Fusi on (§5.2.1)	with p_1					✓			
	with p_2						✓		
	with p_1+p_2							✓	✓
Upsample (§5.2.1)									✓
	PSNR	36.38	38.02	39.12	39.40	39.84	40.23	40.74	**41.82**

model we deploy addition operation for fusion and pixel-shuffle for up-sampling. After adding the proposed modules to the baseline network, the results improve persistently and notably. For instance, we attain a performance gain of 3.02 dB when we incorporate our alignment module into the baseline model. The insertion of the proposed fusion and up-sampling modules in our network further improves the PSNR of the overall network by about 1.34 dB and 1.08 dB, respectively. Overall, GMTNet obtains a compelling gain of 5.44 dB over the baseline model.

Effectiveness of MBFA module.

As reported in Table 5.8, the inclusion of MKGA and AFE modules into our alignment (MBFA) module provides a performance boost of around 1.10 dB and 0.28 dB, respectively which supports the effectiveness of the proposed modules in capturing motion cues. Further, we compare the GMTNet results in Table 6.3 (a) by replacing MBFA with other popular explicit and implicit alignment approaches *(Keeping the rest of the modules same)*. We observe that the MBFA module obtains a performance gain of about 0.83 dB over PCD module proposed in EDVR [275]. To further highlight the ability of MBFA module in aligning burst features, we visualize the features (of few frames) before and after applying it as shown in Figure 5.12. It clearly reveals our MBFA works well without any dedicated supervision.

How to design TAFM module?

A trivial design of our TAFM module is to use a single stream for extracting the information and then concatenating the features. However, from Table 5.8, it is clear

Table 5.9: Impact of the proposed modules in terms of PSNR/SSIM on SyntheticBurst SR dataset for ×4 burst SR task.

Task	Methods	PSNR↑	SSIM↑
(a) Alignment	GMTNet + PCD [275]	40.99	0.953
	GMTNet + Explicit [7]	39.26	0.944
	GMTNet + EBFA [75]	41.10	0.958
	GMTNet + MBFA	**41.82**	**0.960**
(b) Burst Fusion	GMTNet + TSA [275]	39.97	0.947
	GMTNet + DBSR [37]	40.32	0.950
	GMTNet + PBFF [75]	41.60	0.954
	GMTNet + TAFM	**41.82**	**0.960**
(c) Upsampler	GMTNet + Bil	40.22	0.940
	GMTNet + PS [37]	40.41	0.943
	GMTNet + AGU [75]	41.30	0.951
	GMTNet + RTFU	**41.82**	**0.960**

that utilizing both the p_1 and p_2 outputs for subsequent merging results in a performance boost of around 0.90 dB. It clearly signifies that two-stream TAFM performs better than any single-stream.

Impact of TAFM module

The results in Table 5.9 for burst fusion tasks further show that replacing our TAFM module with other popular fusion modules have a detrimental influence on the overall performance of our model, with PSNR drop of around 0.22 dB when utilizing the recently proposed PBFF [75] module in our network.

Effectiveness of the proposed RTFU

To validate the effectiveness of our RTFU, we replace it with the conventional and recent, bilinear interpolation (Bil) and pixel-shuffle (PS), AGU respectively. The accuracy scores in Table 5.9, clearly demonstrate its ability to reconstruct a high-quality image.

How important is the proposed RTM module in RTFU?

To prove the imperativeness of the RTM module, in Figure 5.12 we visualize the feature maps before and after embedding it in RTFU. It clearly proves that our model benefits from the efficient use of both LR and HR information to complete the restoration of sharp regions.

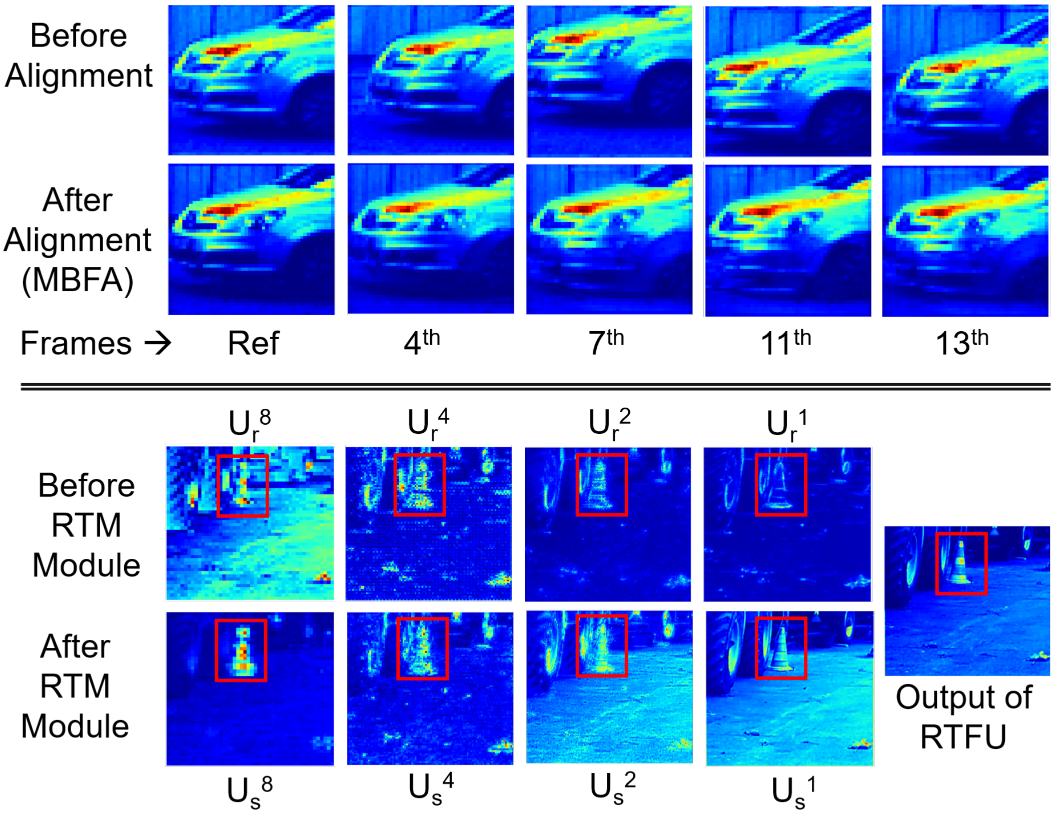


Figure 5.12: Feature map visualizations before and after applying proposed MBFA (Figure 5.6 (b)) and RTM (middle of Figure 5.6 (f)) modules into our GMTNet.

5.3 Summary

In this chapter, we discussed two novel solutions for burst/multi-frame processing approaches. Our first solution we propose an adaptive feature consolidation network (AFCNet) for burst super-resolution. The proposed AFCNet is end-to-end trainable with provision for implicit feature alignment mechanism as well as for inter-frame communication. Additionally, it utilizes adaptive group up-sampling technique to progressively up-scale the multi-frame features.

In our second solution, our proposed Multi-scale Burst Feature Alignment (MBFA) module aligns the noisy burst features at multiple scales using the proposed Attention-Guided Deformable Alignment (AGDA). The inclusion of Aligned Feature Enrichment (AFE) module improves the aligned features by fixing any minor misalignment issue, thus yielding well-refined, denoised and aligned features. To further improve model robustness, Transposed Attention Feature Merging (TAFM) module manifests efficient fusion performance by analyzing the global and local correlations among the incoming frames. Finally, the proposed Resolution Transfer Feature Up-sampler (RTFU) up-scales the merged features by consolidating information from both LR and HR feature spaces to reconstruct a high-quality image.

Experimental analysis of the proposed solutions shows that the proposed networks outperform the existing state-of-the-art methods on all the benchmark datasets for the burst processing task. This chapter and the approaches discussed in the previous chapter are more trained under specific settings (bicubic degradations), and its performance tends to degrade under different settings. Thus, we need to develop a approach that offers good generalization ability in real-world scenarios. A detailed discussion is given in the next chapter.

Chapter 6

A Novel Learning-Based Approach for Blind Super-Resolution

For a deep convolutional network that is trained under fixed conditions, its generalization capability tends to be constrained to that peculiar setting, and its overall performance degrades under different scenarios. This is one of the significant problem in single image super-resolution (SR), where majority of the SR methods presume that blur kernel is fixed and ideal (generally *bicubic kernel*). Naturally, their performance tends to deteriorate when the real kernel diverges from the ideal one, which is quite often in real-world images. Henceforth, recent SR methods opt for *blind* SR, where the true degradation kernels are unknown.

Blind SR that aims to reconstruct the high-resolution image from its low-resolution (LR) counterpart, without knowing the degradation kernel and noise is intrinsically an ill-posed problem as the complex distortions in the LR inputs disrupt many details. The overall degradation process can be expressed as:

$$\mathbf{Y} = (\mathbf{K} * \mathbf{X})\downarrow_s + \mathbf{n} \tag{6.1}$$

where, $\mathbf{Y}$ represents the observed LR counterpart, $\mathbf{X}$ represents the HR image, $\mathbf{K}$ denotes the blur kernel, $*$ is the 2D convolution operator, $\downarrow_s$ denotes the downsampling operation with scale factor $\mathbf{s}$, and $\mathbf{n}$ denotes the additive noise.

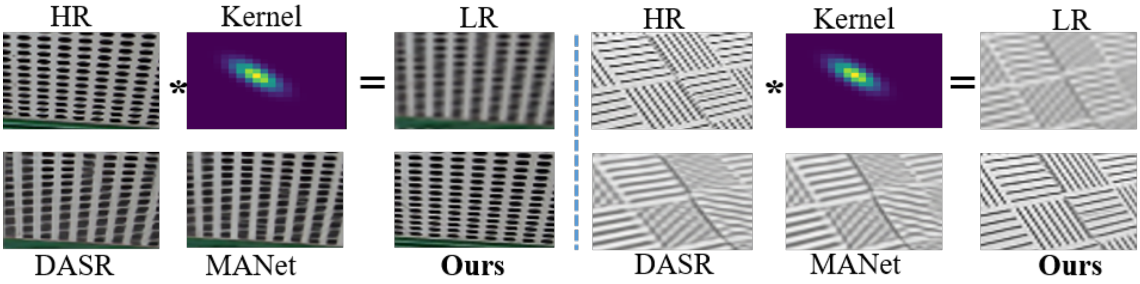


Figure 6.1: Few SR results for ×4 scale factor. The popular methods generate artifacts as they either apply moderate receptive field (MANet) [78], or implicit degradation embedding (DASR) [79]. We incorporate the inherent content information as an important cue to enrich the relationship between the SR network and degradation embedding.

Generally, blind SR is a two step process: degradation kernel estimation, and the subsequent fusion of the kernel prior and content information into the SR network. If the estimated blur kernel diverges from the ground truth, the reconstructed HR image would deteriorate seriously [287, 86]. In light of this, we focus on the degradation estimation problem for blind SR. Additionally, we observe that as the content information can act as a useful cue for the SR network [288], thus estimation of the content-aware degradation can help us in addressing the intrusion arising from the domain gap between the content and degradation spaces. The popular state-of-the-art (SoTA) methods [78, 289] that only deploy a naive encoder or small receptive field are often tumbled by the discrepancy between the above two spaces as these embeddings do not fully exploit the relevant information. To circumvent this problem, we present a transformer-based blind SR framework based on the kernel-oriented adjustment of local and global SR features, called KOADNet. The KOADNet consists of two components: a degradation estimation network for estimating the kernels, and a fusion network for fusing the information via mapping of the predicted degradation kernels to the feature kernel space on SR features. We train on random anisotropic Gaussian degradation settings and our KOADNet is capable of accurately predicting the inherent kernels and leverage this information for SR as shown in Figure 6.1. Additionally, we visualize comparisons on real-world images to demonstrate its good generalization ability in stark comparison to the popular blind SR methods. Our contributions are basically three-fold:

- We design an effective blind kernel-oriented adjustment network for adaptively fusing the degradation-aware embedding and predicted content into the SR network.
- We present an intuitive and efficient network for degradation estimation in blind image super-resolution. It learns both the mean and variance in the latent space of kernels via a dual attention-based information refinement (DAIR) module.
- We experimentally show that the proposed KOADNet outperforms the recently proposed SoTA blind SR models trained under randomized degradation conditions.

6.1 Proposed Methodolgy

Recent works [78, 79] estimate the kernel for the task of blind SR with a moderate receptive field. Furthermore, they do not take into consideration the benefits of leveraging long-range spatial and inter-channel dependencies that play a vital role in the preservation of the required content information. Additionally, unlike previous works [290, 86] that pass the estimated kernel information to each layer of the proposed network thus increasing the overall network complexity, we try to learn the correspondences between the estimated kernels and LR image patches via a novel fusion module. We propose a blind SR network as shown in Fig. 6.2 with two main components: (i) a dual attention-based kernel estimation

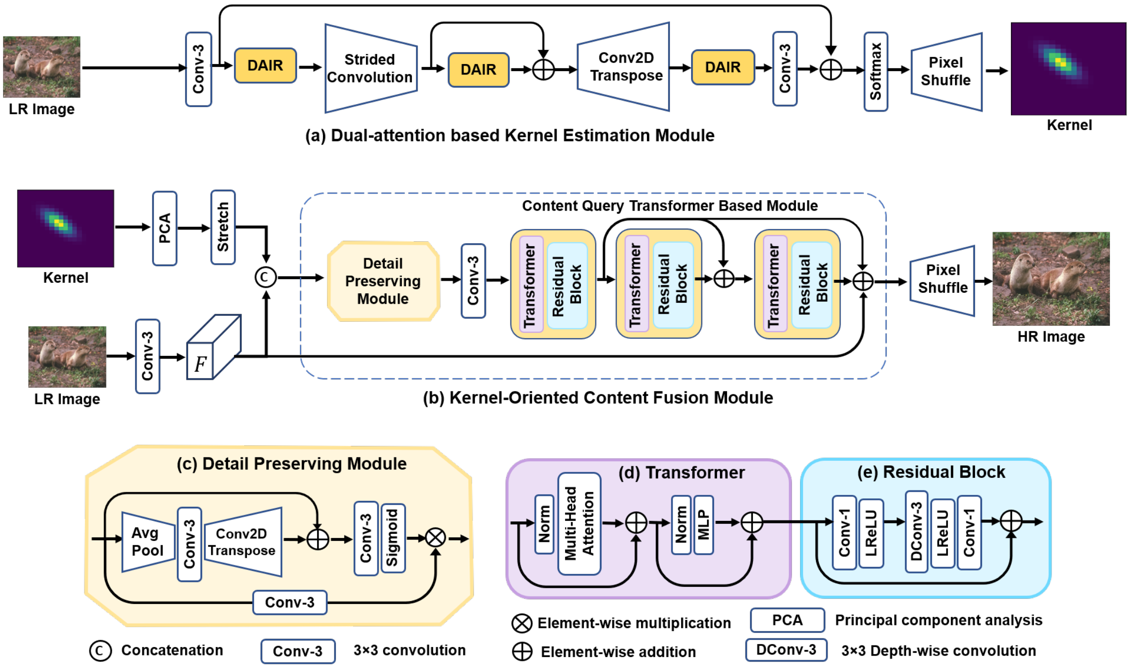


Figure 6.2: The schematic layout of our proposed KOADNet for blind super-resolution. KOADNet has two main components: (a) Dual Attention-Based Kernel Estimation (DAKE) module, (b) Kernel-Oriented Content Fusion (KOCF) module.

module that estimates the degradation kernels, and (ii) a kernel-oriented content fusion module that contains stacked transformer and residual blocks for adaptively fusing the relevant degradation kernel and LR information for enhanced blind SR.

6.1.1 Dual Attention-Based Kernel Estimation Module

It is vital to extract an accurate kernel for the task of blind SR, as kernel mismatch will generate undesirable results [86]. Unlike most previous kernel estimation models [78, 289, 291] that employ naive or small receptive field encoders and often fail to extract the relevant degradation information, we design a dual attention-based kernel estimation module to expand the receptive field. Additionally, other popular SR kernel estimation algorithms [169, 172] are relatively slow and can not be deployed in real-time applications [292]. In light of this, our kernel estimation module, as shown in Fig. 6.2 (a) inputs a degraded LR image and aims to predict the underlying degradation kernel by an efficient architecture that can further provide solid guidance to the kernel-oriented fusion module. Inspired by U-Net [293], the estimation module is composed of convolution layers, dual attention-based information refinement blocks, up-samplers, and down-samplers. The LR image is first passed through a 3×3 convolution layer for extracting image features, which further goes through 3 dual attention-based information refinement blocks. Each refinement block consists of a gradual channel-splitting module with contrast-sensitive and gated attention embedded between them. Before and after the second information

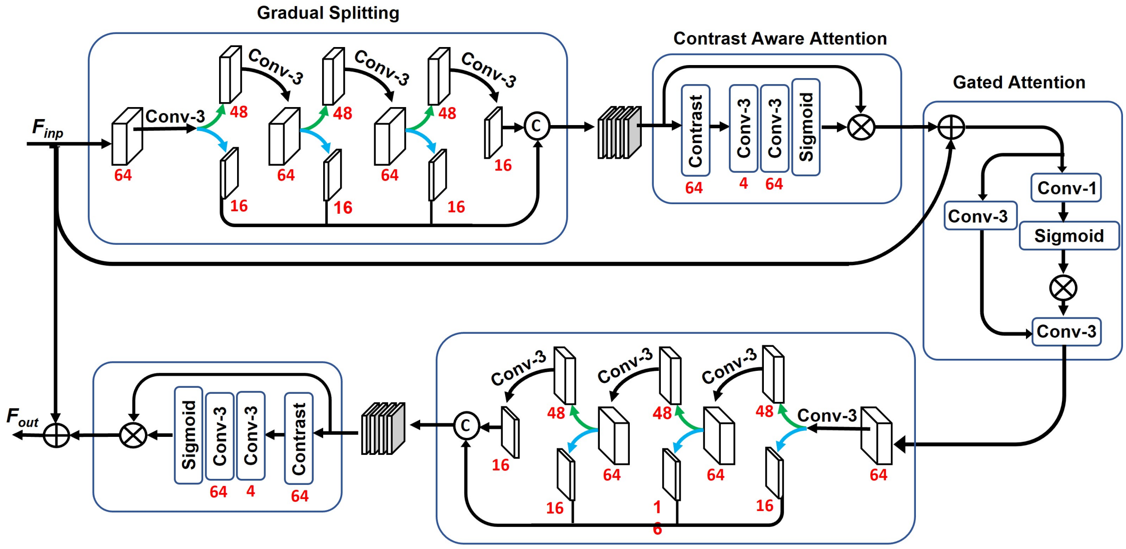


Figure 6.3: Holistic Diagram of Dual-Attention based Information Refinement (DAIR) module.

refinement block, we utilize convolution and a transposed convolution layer (both with a stride of 2) to down-sample and up-sample the features, respectively. We further add skip connections while extracting the features to improve the representation ability and to adaptively utilize the different levels of features. After extracting the features, we reconstruct the kernel by using a 3×3 convolution layer and a softmax layer for predicting the kernels at every LR image pixel. Thereafter, we utilize pixel shuffle for obtaining the kernel predictions for the HR image and the obtained kernel prediction is denoted as $\mathbf{K} \in R^{hw \times H \times W}$, where h, w, H and W denotes the kernel height, width, HR image height, and width, respectively. Its capability to predict kernels by extracting degradation cues is based on the dual attention-based information refinement block described below.

Attention-based Information Refinement Block:

The proposed Dual Attention-based Information Refinement (DAIR) block as demonstrated in Fig. 6.3 extracts the features at a granular level, that preserves the partial information and considers the remaining features of each layer via its gradual splitting module. Concretely, to aggregate the retained features, a contrast-sensitive attention layer is used to enhance the collected refined information. This attention layer helps to adaptively learn the variance and mean of the collected features. To further exploit more useful features (edges, corners, textures) the obtained features are passed through a gated attention layer. Next, we provide more details about each sub-module.

Gradual Splitting Module:

As shown in Fig. 6.3, the DAIR block comprises of two gradual splitting (GS) modules. Each GS module initially adopts a 3×3 convolution layer for extracting the input features for succeeding refinement steps. In every step, channel-wise splitting operation is first employed onto the previous features, that outputs two-part features. From those features, one part is retained (indicated by blue arrow) and the other features are fed into the next calculation part (indicated by green arrow). We consider the retained part as the refined features and the features being fed to the next step as the coarse features. For input features I^{ifeat}, the overall procedure of the each GS module is described as:

$$I_1^{rfeat}, I_1^{cfeat} = Split_1(con_1^{3\times3}(I^{ifeat})), \tag{6.2}$$

$$I_2^{rfeat}, I_2^{cfeat} = Split_2(con_2^{3\times3}(I_1^{cfeat})), \tag{6.3}$$

$$I_3^{rfeat}, I_3^{cfeat} = Split_3(con_3^{3\times3}(I_2^{cfeat})), \tag{6.4}$$

$$I_4^{rfeat} = con_4^{3\times3}(I_3^{cfeat}), \tag{6.5}$$

where, $con_i^{3\times3}$ denotes the i_{th} convolution layer (including LeakyReLU) of each GS module, $Split_i$ represents the i_{th} channel split layer, I_i^{rfeat} represents the i_{th} refined features, and I_i^{cfeat} represents the i_{th} coarse features that require further processing. In the final stage, all the refined features are concatenated and given as:

$$I^{ofeat} = [I_1^{rfeat}, I_2^{rfeat}, I_3^{rfeat}, I_4^{rfeat}] \tag{6.6}$$

where $[\cdot]$ denotes the concatenation operation along the channel dimension.

Contrast Sensitive Attention Layer:

To effectively capture the information from the extracted refined features and improve the overall accuracy of degradation estimation, we utilize a contrast-sensitive attention layer. [20] captures the overall global information in high-level vision via average pooling, but it lacks information about structures, and textures that are quite propitious for enhancing the details in the estimated kernel. Here, we replace the global average pooling operation by the summation of standard mean and deviation to evaluate the overall contrast degree of the kernel map. If we denote $I^{ofeat} = [f_1, f_2, ...f_c]$ as the input of the contrast sensitive attention layer, that has C feature maps of spatial size, $H \times W$. Then, we can acquire the contrast information as:

$$\begin{aligned} z_c &= F_{GC}(f_c) \\ &= \sqrt{\tfrac{1}{HW} \sum_{(i,j) \in f_c} (f_c^{i,j} - \tfrac{1}{HW} \sum_{(i,j) \in f_c} f_c^{i,j})^2} + \\ &\tfrac{1}{HW} \sum_{(i,j) \in f_c} f_c^{i,j}, \end{aligned} \tag{6.7}$$

where, z_c represents the c_{th} element of the output. $F_{GC}(.)$ represents the function for global contrast information extraction.

Gated Attention Layer:

To further improve the representation learning of our kernel estimation network, we propose a gated attention layer as shown in Fig. 6.3. To effectively encode the obtained contrast information from spatially neighboring locations, we formulate this layer as the element-wise product of two parallel paths of transformation layers, one of which is activated through a sigmoid layer. Overall, the gated attention layer is more focused upon enriching features with contextual information by allowing each feature to focus on the fine details complimentary to the other features.

6.1.2 Kernel Oriented Content Fusion Module

Existing blind SR methods usually fuse estimated kernel embeddings into non-blind SR networks without taking into consideration the domain gap. Generally, the degradation features are quite different from the textural features that propagate into the SR network. Thus, we propose the Kernel-Oriented Content Fusion (KOFA) module to mitigate the domain gap by adjusting the intermediate features based upon the estimated degradation kernels. We follow the settings in SRMD [290], which inputs the concatenated LR image and kernel maps of size $(b + C) \times H \times W$, where b represents the batch dimension, C, H and W represent the channels, image height, and width respectively. Thus, given a degraded image and the estimated kernel, KOFA first reshapes the estimated kernel from $H \times W$ to HW, and then the dimensionality is reduced from HW to l via principal component analysis (PCA). Thereafter, the kernel PCA vectors are stretched to a kernel PCA map of size $l \times \frac{H}{s} \times \frac{W}{s}$, where H, W, and s denote HR image height, width, and scale factor, respectively. However, for efficiently exploiting the kernel information, merely concatenating the transformed kernel and LR image as done in [173, 290] is not a better option due to the following reasons: Firstly, the kernel maps are unable to accommodate the actual information from the image. Secondly, simultaneous processing of the kernel maps and LR image introduces interference which is unrelated to the image.

Thus, after concatenation we pass the information through a detail-preserving module as shown in Fig. 6.2 (c) that consists of a down-sampling layer (via average pooling), followed by a 3×3 convolution layer and an up-sampling layer (via transposed Convolution) for extracting content-aware degradation features. We then forward the up-sampled features through a gated attention mechanism to exploit the relevant detail cues. Finally, the extracted features are passed through several cascaded transformer blocks followed by residual blocks for leveraging the content cue.

Content Query Transformer Based Module:

Having learned the content-aware degradation information in LR and HR spaces, we now seek to effectively integrate it among the features. For accomplishing this, we propose a content query transformer based (CQTB) module in Fig. 6.2 (b). Each content query transformer based module consists of transformer and residual blocks, where the transformer module effectively leverages the content cue to query the long-range dependencies among the features, and the residual block enhances the local collaboration among the extracted features. In every transformer block, we employ shifted window based multi-head attention layers and feed-forward networks. Given features F of size, $H \times W \times C$ from the detail preservation module, shifted window mechanism first reshapes the input to $\frac{HW}{M^2} \times M^2 \times C$ features via partitioning the overall input into non-overlapping $M \times M$, w local windows, where $\frac{HW}{M^2}$ denotes the total number of windows. Next, the standard self-attention is applied on each window. For a local window feature, $X \in \mathbb{R}^{M^2 \times C}$, the three learnable weight matrices $\mathbf{W}^Q \in \mathbb{R}^{C \times C}$, $\mathbf{W}^K \in \mathbb{R}^{C \times C}$ and $\mathbf{W}^V \in \mathbb{R}^{C \times C}$ are shared across different windows and projected into the query $\mathbf{Q}$, key $\mathbf{K}$ and value $\mathbf{V}$ via: $\{\mathbf{Q}, \mathbf{K}, \mathbf{V}\} = \{X\mathbf{W}^Q, X\mathbf{W}^K, X\mathbf{W}^V\}$. The attention function computing the dot product of the query with key is defined as:

$$Attention(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = softmax(\frac{\mathbf{QK^T}}{\sqrt{d}})\mathbf{V}, \tag{6.8}$$

Here, d represents the dimensionality of keys. Further, a feedforward network consisting of two multi-layer perceptrons (MLP) layers and GELU activation is employed for refining the features generated by multi-head attention as demonstrated in Fig. 6.2 (a). To effectively capture the global information, residual connection is further applied between the both modules.

$$\begin{aligned} \hat{Z} &= LN(MSA(\mathbf{Q}, \mathbf{K}, \mathbf{V})) \\ Z &= LN(FFN(\hat{Z})) + \hat{Z}; \end{aligned} \tag{6.9}$$

where, $\hat{Z}$ is defined as the output of the MSA unit with X as the input. The feedforward network (FFN) is defined as follows:

$$FFN(X) = GELU(W_1 X + b_1)W_2 + b_2 \tag{6.10}$$

Thereafter, residual block is placed in series after every transformer block for calibrating the anisotropically degraded features before the final reconstruction phase. After passing the features through cascaded transformer and residual blocks, we pass the features through pixel-shuffle layer for up-scaling the extracted features.

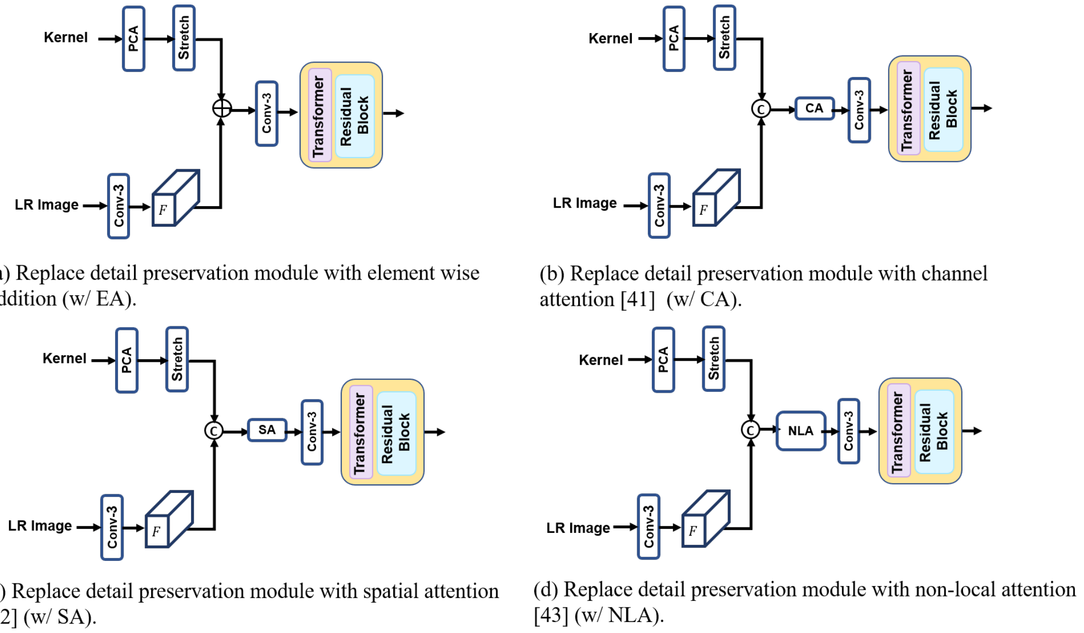
6.2 Experiments

6.2.1 Experimental Setup

As the blur kernels of real-world LR images are generally unimodal [160] and can be effectively depicted by a Gaussian [294] model, most popular blind SR works [295, 86, 296, 297, 157] assume that the SR kernel is either isotropic or anisotropic Gaussian kernel. Following this extensively-embraced assumption, we perform all the experiments upon anisotropic Gaussian kernels and 800 training images from DIV2K and 2650 images from Flickr2K [298] datasets are used to train the network. The training degradation makes use of 21×21 anisotropic Gaussian kernels and the noise level is set to 0 except for the ×4 noisy experiment that is set to 15. For each scale factor $s \in 2, 3, 4$, the kernel width and size ranges are set to [0.175s, 2.5s] and (4s + 3)×(4s + 3), respectively. For all s, the rotation angle range is $[0, \Pi]$. According to the settings in [83, 155], we shifted the blur kernel and the upper left pixels are left downsampled fo avoid subpixel misalignment. For quantitative evaluation, we compare the estimated kernels using PSNR and compare the resulting SR images using PSNR and SSIM on the Y channel in YCbCr space. Our model is trained for 300,000 iterations. We use Adam [207] with $\beta 1 = 0.9$, and $\beta 2 = 0.999$ as the optimizer. The inital learning rate is initialized as 2e-4 and reduced by half every 50,000 iterations.

Table 6.1: Quantitative comparison of the all the variants in detail preserving module on Set5 for ×4

Variant	Explanation	PSNR
w/o DP module	Baseline model, removing DIP	31.42
w/ EA	Replacing DP with element-wise addition	31.46
w/ CA	Replacing DP with channel attention [44]	31.52
w/ SA	Replacing DP with spatial attention [80]	31.49
w/ NLA	Replacing DP with non-local attention [81]	31.58
w/ DP (Our full model)	Replacing DP with non-local attention	**31.63**



(a) Replace detail preservation module with element wise addition (w/ EA).

(b) Replace detail preservation module with channel attention [41] (w/ CA).

(c) Replace detail preservation module with spatial attention [42] (w/ SA).

(d) Replace detail preservation module with non-local attention [43] (w/ NLA).

Figure 6.4: The structure of the four different variants: (a) w/ EA; (b) w/ CA [44]; (c) w/ SA [80]; (d) w/ NLA [81] which are used to substitute the detail preservation module (DP) shown in Fig. 6.2 (c).

Table 6.2: Ablation study on popular kernel estimation and fusion modules.

Method	PSNR/SSIM
RRDB-SFT [78] + IKC [86]	31.08/0.8781
RRDB-SFT [78] + MANet [78]	31.54/0.8876
RRDB-SFT [78] + DAKE	31.47/0.8875
KOCF + IKC [86]	31.38/0.8789
KOCF + MANet [78]	31.51/0.8875
KOCF + DAKE	**31.63/0.8883**

Table 6.3: Investigation of the transformer and residual blocks in our Content-Query Transformer Based (CQTB) module. Here, we notice the quantitative performance (PSNR/SSIM) on Set5 for ×4.

Metrics	CQTB w/ transformer	CQTB w/ residual	CQTB w/ all
PSNR	31.48	31.34	**31.63**
SSIM	0.8849	0.8855	**0.8883**

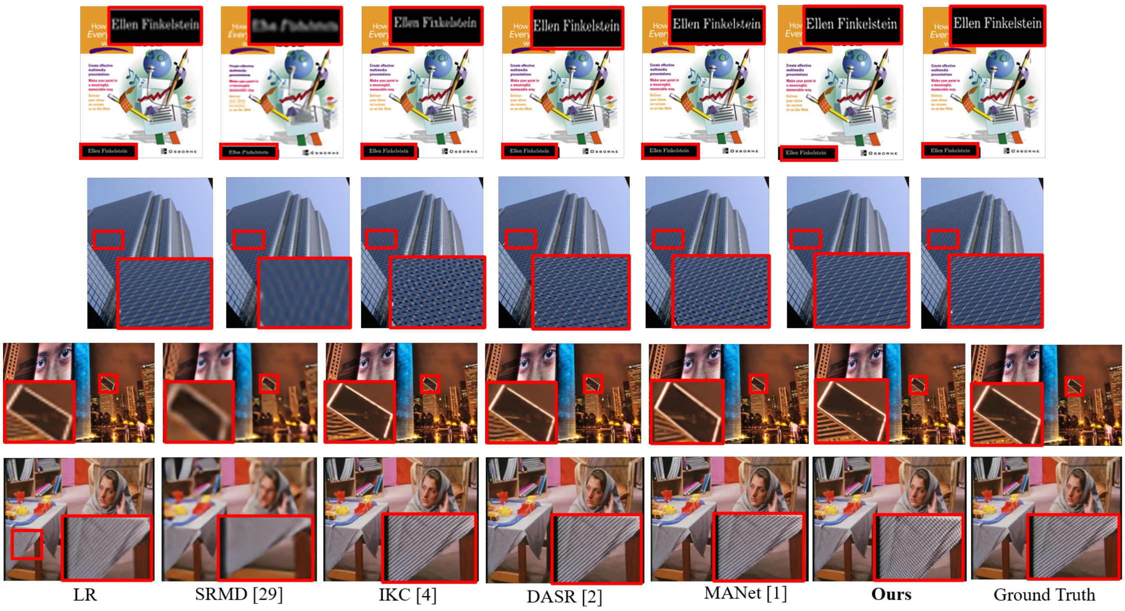


Figure 6.5: Visual results of challenging images from different benchmark datasets for scale ×3 (first two rows) and ×4 (last two rows).

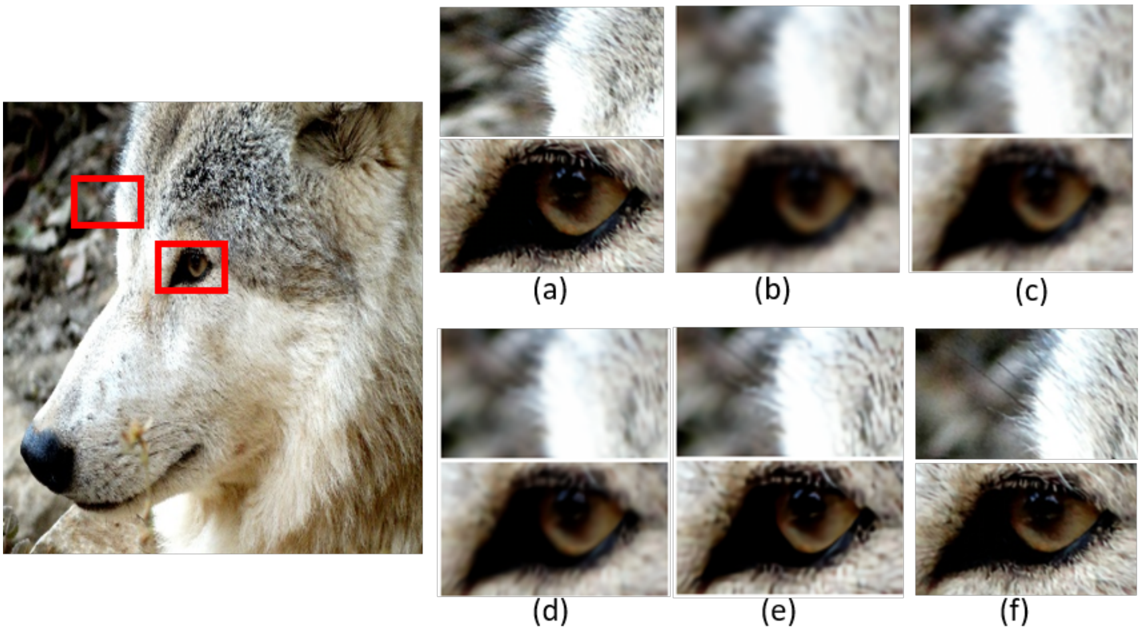


Figure 6.6: Visual results on an image in DIV2KRK dataset for ×4 under anisotropic gaussian kernel. Here (a) Ground-truth HR patch, (b) RCAN [44], (c) WDSR [82], (d) KernelGAN + ZSSR [83], (e) DASR [79], and (f) Ours.

Table 6.4: Quantitative comparison on popular benchmark testing sets with anisotropic gaussian kernels.

Method	Scale Factor	Noise Level	Set5 [38]	Set14 [31]	BSD100 [21]	Urban100 [84]
HAN+Correction [299]	×2	0	28.61/0.8013	26.22/0.7292	26.88/0.7116	25.31/0.7109
SRSVD [300]	×2	0	34.51/0.8787	31.10/0.8581	29.71/0.7993	28.08/0.7965
IKC [86]	×2	0	35.30/0.9381	31.48/0.8797	30.50/0.8545	28.62/0.8689
DASR [79]	×2	0	35.30/0.9360	31.30/0.8683	30.46/0.8507	28.66/0.8654
CDSR [288]	×2	0	36.17/0.9428	32.14/0.8841	31.02/0.8643	29.57/0.8851
Ours	×2	0	35.85/0.9411	31.97/0.8849	30.99/0.8648	29.84/0.8855
HAN [42]	×3	0	23.71/0.6171	22.31/0.5878	23.21/0.5653	20.34/0.5311
DIP [301]	×3	0	27.51/0.7740	25.03/0.6674	24.60/0.6499	22.23/0.6450
IKC [86]	×3	0	32.94/0.9104	29.14/0.8162	28.36/0.7814	26.34/0.8049
DASR [79]	×3	0	33.43/0.9151	29.57/0.8187	28.58/0.7846	26.83/0.8174
MANet [78]	×3	0	33.69/0.9184	29.81/0.8270	28.81/0.7932	27.39/0.8331
CDSR [288]	×3	0	33.81/0.9192	29.95/0.8275	28.81/0.7922	27.44/0.8329
Ours	×3	0	33.97/0.9222	30.08/0.8336	28.90/0.7991	26.98/0.8204
HAN+Correction [299]	×4	0	24.31/0.6357	24.44/0.6341	24.01/0.6005	22.32/0.6368
IKC [86]	×4	0	31.08/0.8781	27.83/0.7663	27.12/0.7233	25.16/0.7609
DASR [79]	×4	0	31.45/0.8859	28.12/0.7703	27.24/0.7284	25.28/0.7636
MANet [78]	×4	0	31.54/0.8876	28.28/0.7727	27.36/0.7307	25.66/0.7660
CDSR [288]	×4	0	31.62/0.8885	28.31/0.7746	27.38/0.7311	25.55/0.7783
Ours	×4	0	31.63/0.8888	28.36/0.7741	27.36/0.7315	25.82/0.7793
HAN+Correction [299]	×4	15	19.21/0.2281	19.25/0.4231	19.25/0.4231	19.01/0.3500
IKC [86]	×4	15	27.23/0.7877	25.55/0.6717	25.15/0.6236	23.31/0.6697
DASR [79]	×4	15	27.48/0.7907	25.56/0.6723	25.25/0.6261	23.30/0.6663
MANet [78]	×4	15	27.58/0.7915	25.75/0.6744	25.30/0.6262	23.57/0.6760
CDSR [288]	×4	15	27.70/0.7947	25.81/0.6757	25.33/0.6277	23.60/0.6761
Ours	×4	15	27.93/0.8020	26.02/0.6859	25.53/0.6367	23.76/0.6832

Table 6.5: Quantitative comparison of popular SR methods for anisotropic gaussian kernels on DIV2KRK [86] dataset.

	Method	PSNR/SSIM
Category 1: trained upon bicubic downsampled images	RCAN [44]	25.66/0.6936
Category 2: winners of NTIRE blind SR competition	WDSR [82]	25.64/0.7144
Category 3: degradation kernel estimation + blind SR method	KernelGAN + ZSSR [83]	26.81/0.7316
Category 4: Ground-truth kernel + blind SR method	Ground-truth kernel + ZSSR [83]	27.53/0.7446
Category 5: end-to-end trainable blind SR method(kernel estimation + SR)	DASR [79]	28.15/0.7722
	DSSR [302]	28.78/0.7905
	Ours	**30.25/0.8340**

6.2.2 Ablation Study

Here, we perform several experiments for investigating the overall effectiveness of every proposed component in our KOADNet. For all the experiments, we trained our model for 1×10^5 iterations.

Effect of detail preservation module

For investigating the effect of our proposed detail preservation module, we quantitatively compare performance of KOADNet with 5 other variants. For unbiased comparison, we just substitute the detail preservation (DP) module in Fig. 6.2) (c) with the following models:

1. The baseline model (represented by w/o DP module), that removes the detail preserving module.

2. w/ EA module that consolidates the degradation and low-resolution information by element-wise addition.

3. w/ SA and w/ CA, that replaces the DP module with spatial attention [80] and channel attention [44] modules, respectively.

4. w/ NLA, that replaces the DP module with non-local attention operation [81].

The corresponding designs of all these variants are shown in Fig. 6.4. TABLE 6.1 and Fig. 6.6 demonstrate the comparison of the quantitative and qualitative results on all the considered variants. As clearly visible from the TABLE and Fig., our DP module generates the best results, thus proving its effectiveness in modulating the structural contextual information conditioned on the degradation kernels for improving the SR performance.

Effect of kernel estimation and fusion module:

To further prove the effectiveness of our proposed kernel estimation and fusion modules, we provide a quantitative comparison among popular kernel estimation and fusion modules. As shown in TABLE 6.2, we replace our proposed dual-attention based kernel estimation (DAKE) module with popular kernel estimation modules, Iterative Kernel Correction (IKC) [86] and Mutual affine convolution layer (MANet) [78], and thereafter their estimated kernels are embedded into fusion modules, RRDB-SFT [78] and our proposed Kernel Oriented Content Fusion (KOCF) module. We can conclude from TABLE 6.2 that a combination of our dual-attention based kernel estimation (DAKE) module and KOCF gives an average 0.14 dB improvement in PSNR.

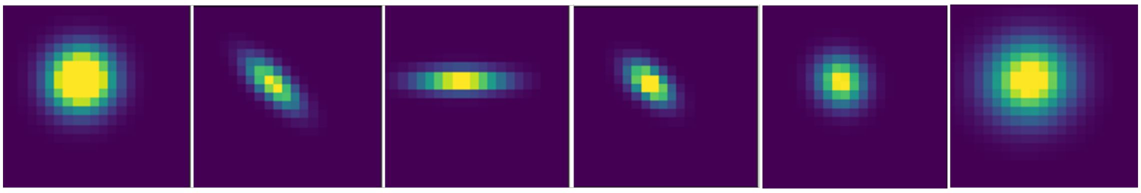


Figure 6.7: Diverse kernel samples estimated by our dual attention based kernel estimation module.

Effect of content query based transformer module

Additionally, we research the influence of content query based transformer module in kernel oriented content fusion module. As demonstrated in TABLE 6.3, after modulating the content information upon both the transformer and residual blocks (w/ all), the model performs better than as compared to when it is conditioned on either of these blocks.

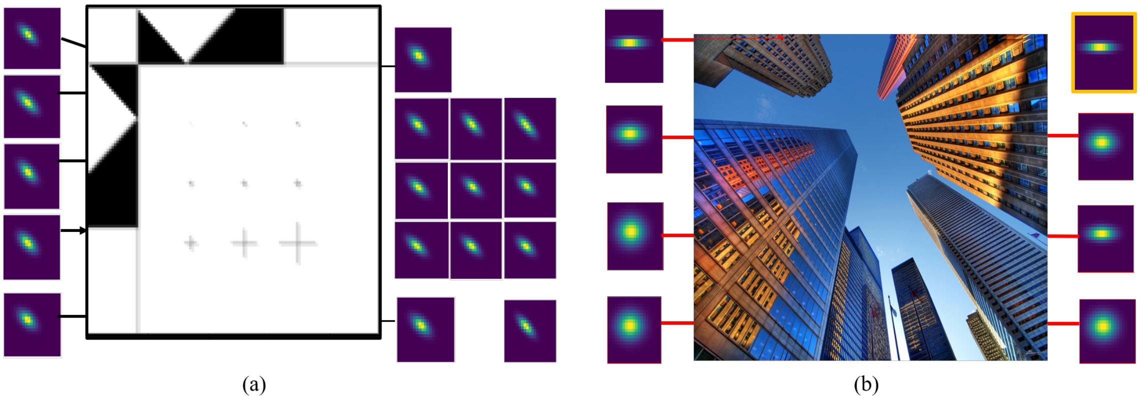


Figure 6.8: (a) Kernel estimation results of our KOADNet on "*img*012" of Urban100 [84] for scale factor ×4 whose matching HR image has been blurred by an anisotropic kernel as demonstrated in the top right yellow rectangle, and (b) Multiple position kernel estimation results of our proposed KOADNet on a synthetic image for scale factor 4 that is synthesized via blurring through a Gaussian kernel with $\sigma_1 = 6$, $\sigma_2 = 1$, and $\theta = \pi/4$.

Comparison with state-of-the-art methods

We perform several experiments upon degradations from anisotropic Gaussian kernels and noise. As demonstrated in TABLE 7.1, we compare KOADNet with existing blind SR models. For fair comparison, we retrained few methods in the same experimental settings. As demonstrated in TABLE, the proposed KOADNet outperforms the popular DASR and MANet by 0.54 dB and 0.28 dB, respectively on ×3 and attains competitive performance on PSNR and SSIM. Though SRSVD [300] and IKC [86] struggle in estimating the kernel information by incorporating adversarial network and iteratative kernel refinement, respectively, they perform poorly when compared to KOADNet. The implicit kernel prediction in KernelGAN is incapable of capturing the relevant information in cases of severe degradation. HAN [42] extracts the refinement feature in the SR network by employing the attention mechanism, but this method ignores the degradation prior that results in poor performance. DIP [301] suffers on account of multiple degradations and is not able to generate a suitable prior. DASR [79] improves the overall performance by employing a discriminative degradation encoder. However, the domain gap between the contextual and degradation space limits the quantitative results, thus performing inferior in comparison to proposed KOADNet. It is worth mentioning that KOADNet attains dominant performance in comparison to the SoTA methods for multiple degradations (*i.e.* ×4 with noise level 15) on all the benchmark datasets.

Table 6.6: Kernel estimation results under more complex noise corruptions. Here, q and η denotes the different JPEG compression and gaussian noise level. We have reported the PSNR/SSIM on BSD100 [84] for ×4 scale factor.

$\downarrow \eta/q \rightarrow$	70	80	100
0	44.23/0.9937	45.97/0.9955	46.78/0.9968
5	44.44/0.9934	44.89/0.9934	44.32/0.9936
15	39.11/0.9746	40.44/0.9843	42.76/0.9911

In Fig. 6.5, we visualize few challenging images from the benchmark datasets for comparison. It is clearly observed that popular methods SRMD and IKC cannot recover the textural and contextual information. Though DASR and MANet performs better than the other methods, but the resultant image contains several obvious blurs. In contrast, our KOADNet generates visually pleasing image with sharper edges and clearer textures, that resembles the ground-truth.

Following [86], we also compare our KOADNet with 5 types of SR algorithms on a more challenging, DIV2KRK dataset:

1. Category 1: Non-blind SoTA SR method that is trained upon LR images that are downsampled bicubically, *e.g.* RCAN [44].

2. Category 2: The winner of the NTIRE Blind SR competition, WDSR [82].
3. Category 3: The method that consolidates the degradation kernel estimation and blind SR framework, KernelGAN + ZSSR [83].
4. Category 4: On a similar note to Category 3 method, the ground truth blur kernels are provided and considered as input to the complete network.
5. Category 5: The methods that unify the kernel estimation and SR reconstruction in an end-to-end trainable framework, *e.g.* DASR [157], DSSR [302], and Ours.

Table 6.7: Average PSNR/SSIM results on SoTA methods for spatially variant blind SR on BSD100 [21] dataset for ×4.

Method	Spatial Variant kernel type		
	1	2	3
	$\sigma_1 = a + b$ $\sigma_2 = ax + b$ $\theta = 0$	$\sigma_1 = a + b$ $\sigma_2 = ax + b$ $\theta = 0$	$\sigma_1 = a + b$ $\sigma_2 = ax + b$ $\theta = \pi x$
HAN [42]	22.19/0.5111	21.83/0.5066	21.66/0.4989
DIP [301]	25.24/0.6174	25.30/0.6242	24.01/0.5813
KernelGAN [83]	19.90/0.4317	18.32/0.6697	17.62/0.3517
HAN + Correction [299]	25.13/0.6151	25.51/0.6156	24.41/0.6017
MANet [78] + IKC [86]	**26.46/0.6952**	26.03/**0.6880**	25.58/**0.6759**
Ours	26.23/0.6790	**26.30**/0.6832	**25.70**/0.6655

Table 6.8: The no-reference NIQE [87] , and NRQM [88] results of SoTA methods on the RealSRSet [85] Dataset.

Metrics	Methods		
	DASR	MANet	Ours
NIQE ↓	6.22	6.42	**6.04**
NRQM ↑	3.23	3.39	**4.01**

The quantitative and qualitative results (PSNR/SSIM) on DIV2KRK dataset are illustrated in TABLE 6.5 and Fig. 6.6. As clear from the TABLE and Fig,, on account of the kernel divergence, the method that is trained upon bicubic kernel (Category 1) have limited capability in solving the anisotropic Gaussian kernels. Though the methods that are trained on synthesized images of NTIRE competition (Category 2) achieve better results than Category 1, but they still have some limitation on irregular blur kernels. From the quantitative results of Category 3, we observe that the sequential estimation of kernel and subsequent SR reconstruction cannot generate better HR images. Similar phenomenon is observed in Category 5. The main reason being that these methods generate results that are quite sensitive to kernel estimation errors, and larger estimation gap can worsen the results. For Category 4 method, under the provision of ground-truth kernels, ZSSR can perform much better than compared to the estimated degradation kernels (Category 3). However, unlike DASR, and DSSR, our KOADNet attains the SoTA performance on ×4 factor for Set5 dataset.

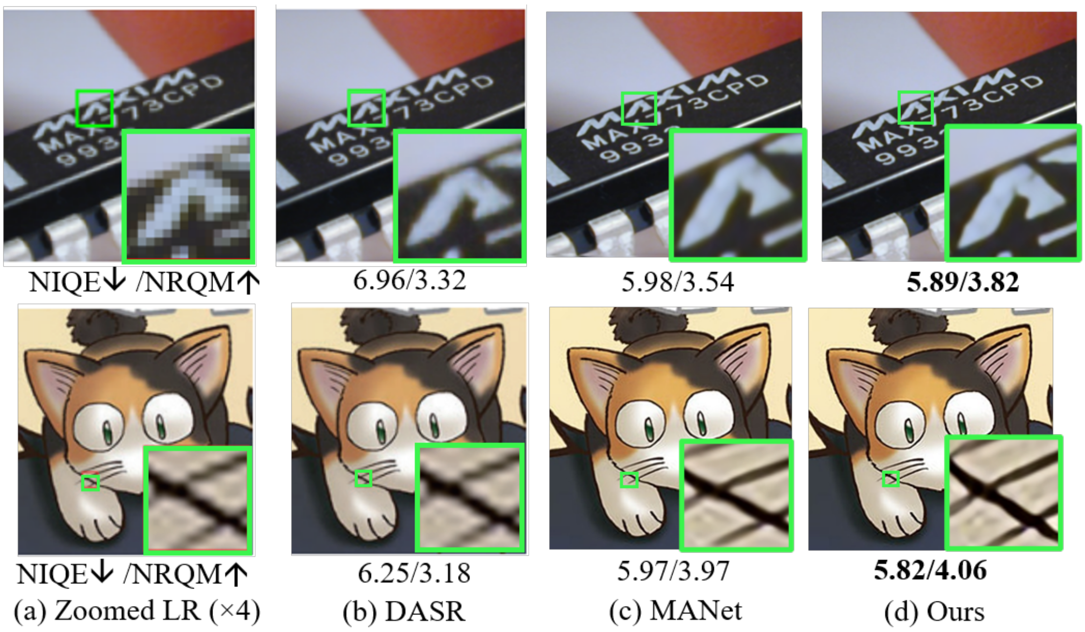


Figure 6.9: Few Visual Results on challenging images from RealSRSet [85] for scale factor 4.

6.2.3 Experimental Analysis on Kernel Estimation

In Fig. 6.7, we demonstrated few samples of the kernels estimated by our dual-attention based kernel estimation module. In Fig. 6.8, we plot the results of kernel estimation on a testing image. As clearly visible, KOADNet tends to accurately estimate the kernels from the non-flat patches (*e.g.*, the building structures) and predict fixed kernels for the flat patches (*such as*, the blue sky) in an image, that may be the average of all the possible predicted kernels. The estimated kernels are quite variable and may not resemble the ground-truth kernel, but majority of these are the correctly estimated kernels as specified by the high image LR PSNR. For a clear understanding, we further test our proposed network upon a synthetic image. As demonstrated in Fig. 6.8 (b), KOADNet can accurately estimate kernels from a small patch of size 9×9. The overall performance continues to ameliorate with increase in the patch size. With only edges in a small patch, KOADNet is unable to accurately estimate the kernels on account of insufficient information. In case of flat patches without any edges and corners, KOADNet infers a fixed isotropic-like kernel.

In real-world scenes, images are likely to be affected on account of compression artifacts or noisy degradations. For testing the performance of kernel estimation in more complex cases, we perform another experiment where we add JPEG compression and Gaussian noises while training and consequently test the performance of the network under different compression and noise levels. As one can see in TABLE 6.6 that the LR image PSNR varies from 39.11 dB to 46.78 dB, which clearly reveals the potential of our network to estimate kernels under heavy noisy corruptions.

6.2.4 Experimental Analysis on Spatially Invariant SR

Though the existing SoTA blind SR methods have attained outstanding performance, they presume that the inherent blur kernels are spatially invariant and generally estimate a single kernel for the complete image, causing some inherent problems. Firstly, the blur kernels in real-world images are spatially variant and on account of several environmental factors such as object motion and depth difference, blur kernels at multiple locations in the image are generally variable.

Additionally, estimating a single kernel for the complete image is quite prone to the adverse results of flat patches, even assuming the spatial invariant case. To prove the efficacy of our approach, we show some additional results for spatially variant blind SR approach in TABLE 6.7 where each testing image is first divided into $m \times n$ patches (every single patch is of size 40×40), that are then degraded by variable kernels. Following the settings in [78], for scale factor s, the minimum kernel width range, a and the Gaussian kernel width range, b and are set to $0.175s$ and $2.325s$, respectively. Particularly, for patch (i, j), the matching kernel is decided by a, b, $x = \frac{i}{m}$, and $y = \frac{j}{n}$ as demonstrated in the header of the TABLE 6.7.

6.2.5 Experimental Analysis on Real-world SR

For further demonstrating the overall effectiveness of our method, we test our model on some real-world images. As the ground-truth for the RealSRSet [85] is unavailable, we utilize the non-reference image quality assessment metrics including NIQE [87], and NRQM [88] for quantitative evaluation. As observed from TABLE 6.8, in comparison to the SoTA methods our KOADNet shows impressive results. It certainly prove that our KOADNet generalizes well upon images with a broad range of degradations.

As shown in Fig. 6.9, MANet and DASR method produce blurry images with ringing artifacts. However, our KOADNet produce sharp edges with lesser artifacts, thus corroborating the efficiency of our proposed components in terms of adding high-frequency details in the edges and low-frequency information in the flat areas.

6.3 Summary

Blind SR is a crucial problem to generalize the learning-based SR networks for handling diverse types of content and degradations of LR data. In our work, we have proposed a transformer based network for blind SR which leverages the degradation information efficiently for subsequent modulation of the SR features. For achieving this goal, we designed a dual-attention based kernel estimation (DAKE) module that predicts spatially invariant kernels. Then, a kernel-oriented content fusion (KOCF) module is proposed to leverage this estimated kernel information efficiently for enhancing the model expressiveness. Consequently, our proposed KOADNet accurately anticipates the HR images under real-world settings. It also exhibits good performance on degradation kernel estimation, that leads to SoTA performance on blind image SR when fused with popular blind SR networks. In future, our main intention is to work upon enhancing the SR ability for much more complex degradations, like rainy and hazy scenes, low-light images, and low-resolution surveillance videos.

Chapter 7

Conclusion and Future Scope

7.1 Conclusion

The main aim of this thesis work is to design and develop novel approaches for image super-resolution. The major challenges like unknown degradations, high dependency on prior edge information, huge computational complexity, lack of research on multi-frame super-resolution and upsampling approaches need to be tackled for accurate super-resolution. This work mainly focuses on analyzing and designing different solutions for image super-resolution in the context of providing the solution to the above-mentioned challenges.

Accurate estimation of frequency information is a key step to super-resolve the low-quality images. Most of the existing state-of-the-art methods utilize explicit edge prior information for extracting the relevant frequency information, which increases the overall complexity of the network. To overcome this, an end-to-end learning based frequency extraction network is proposed to generate visually plausible results.

The imperfection of existing heavy-weight SR approaches in terms of longer training times, limited flexibility, huge time and power consumption inspired us to propose computationally efficient architectures with less computational cost. In light of this, an end-to-end lightweight network is proposed that maintains a proper trade-off between accuracy and speed.

Burst Super-resolution is quite a challenging task since individual burst images often have inter-frame misalignments that usually leads to ghosting, and zipper artifacts and subsequently affecting the fusion and reconstruction stages. To mitigate this, we have developed two novel approaches for burst image processing that focuses solely on the relevant information exchange between burst frames and filter-out the inherent degradations while preserving and enhancing the actual scene details.

Most of the existing techniques proposed for SR are highly dependent upon simple degradations, that leads to limited practical serviceability of the algorithm in real-world scenarios. To address these issues, a simple, and effective novel transformer-based blind approach has been proposed for the task of super-resolution.

The proposed single image and blind super-resolution approaches are evaluated on the current state-of-the-art SR databases such as Set5, Set14, BSD100, Urban100, and

Manga109. And, the proposed burst super-resolution approaches are evaluated on testing datasets of SyntheticBurst and real BurstSR. The qualitative and quantitative results of proposed methods are examined and compared with SoTA learning-based methods. Standard quantitative evaluation parameters such as PSNR, and SSIM are used to evaluate the proposed approaches.

To summarize, the existing problems for super-resolution are scrutinized, and attempts are made to contribute for resolving the existing problems. The quantitative comparison of all the proposed approaches in terms of parameters, PSNR, and SSIM on the benchmark datasets of single image and blind image super-resolution (Set5 and Urban100), and Synthetic BurstSR and Real BurstSR datasets of burst image super-resolution for scale factor ×4 factor is shown in Table7.1.

Table 7.1: The quantitative comparison between the proposed approaches in terms of parameters, PSNR and SSIM for Set5, Urban100, SyntheticBurst SR, and Real BurstSR datasets for × 4 scale factor. The best results for each contributions are highlighted in red color.

	Single Image SR			
Modalities	I:A	I:B	II:A	II:B
	(Sec 3.1)	(Sec 3.2)	(Sec 4.1)	(Sec 4.2)
Methods	MBUP-Net	MLEAUNet	MSARNet	Con-Net
Parameters	9.1M	8.2M	1.5M	0.67M
Set5	32.67/0.9005	**32.80/0.9100**	32.29/0.8989	**32.57/0.9001**
Urban100	**27.67**/0.8113	27.05/**0.8165**	26.25/0.7907	**26.44/0.7972**
	Burst Image SR		Blind Image SR	
Modalities	III:A	III:B	IV	
	(Sec 5.1)	(Sec 5.2)	(Sec 6.1)	
Methods	AFCNet	GMTNet	KOADNet	
Parameters	35M	13M	11M	
Set5	-	-	31.63/0.8888	
Urban100	-	-	25.82/0.7793	
SyntheticBurst SR	42.21/0.9600	**42.36/0.9610**	-	
RealBurst SR	48.63/0.9860	**48.95/0.9860**	-	

7.2 Future Scope

The main aim of the work is to propose novel solutions for major super-resolution categories - single image super-resolution, burst image super-resolution and blind image super-resolution by tackling the major challenges of each task. In future, we intend to propose an end-to-end trainable network that can handle arbitrary (non-integer) scale factors. Also, we can extend our approach for much larger scale factors like ×16, ×32 and ×64. Additionally, we can propose a lightweight solution for burst super-resolution that can be easily deployed on mobile devices. Additionally, we can try to enhance the reconstruction ability of SR for more complex degradations, like hazy, rainy scenes, and low-resolution surveillance videos. Additionally, meta-learning can be considered to make our model more adaptable and flexible on real-world scenes.

List of the Publications

International Journals

1. Nancy Mehta and Subrahmanyam Murala, “MLE2A2U-Net: Image Super-Resolution via Multi-Level Edge Embedding and Aggregated Attentive Upsampler Network,” in ***IEEE Transactions on Emerging Topics In Computational Intelligence (2022)*** (**Impact factor- 4.85**).

2. Nancy Mehta and Subrahmanyam Murala, “Con-Net: A Consolidated Lightweight Image Super-Resolution Network”, in ***IEEE Transactions on Broadcasting (2022)*** (**Impact factor- 5.194**).

3. Nancy Mehta and Subrahmanyam Murala, “Image Super-Resolution with Content-Aware Feature Processing,” in ***IEEE Transactions on Artificial Intelligence (2022)***.

4. Nancy Mehta and Subrahmanyam Murala, “MSAR-Net: Multi-scale Attention based Light-Weight Image Super-Resolution,” ***Pattern Recognition Letters (2022) (Impact factor- 4.757).***

5. Nancy Mehta and Subrahmanyam Murala, “KOADNet: Kernel Oriented Adjustment Network for Blind Image Super-Resolution,” ***IEEE Transactions on Image Processing. (Under review)***

International Conferences

1. N. Mehta, A. Dudhane, S.Murala, S. Zamir, S. Khan and F. Khan, “Adaptive Feature Consolidation Network for Burst Super-Resolution”, In Proceedings of the ***IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshop, 2022, pp. 1279-1286*** (**h-index- 106**).

2. Mehta, N., Dudhane, A., Murala, S., Zamir, S. W., Khan, S., and Khan, F. S. (2023). Gated Multi-Resolution Transfer Network for Burst Restoration and Enhancement. In Proceedings of the **IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 22201-22210)** (**h-index- 389**).

References

[1] Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie Line Alberi-Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. 2012.

[2] Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In *International conference on curves and surfaces*, pages 711–730. Springer, 2010.

[3] Pablo Arbelaez, Michael Maire, Charless Fowlkes, and Jitendra Malik. Contour detection and hierarchical image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 33(5):898–916, 2010.

[4] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5197–5206, 2015.

[5] Radu Timofte, Shuhang Gu, Jiqing Wu, and Luc Van Gool. Ntire 2018 challenge on single image super-resolution: Methods and results. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 852–863, 2018.

[6] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 126–135, 2017.

[7] Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. Deep burst super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9209–9218, 2021.

[8] Yusuke Matsui, Kota Ito, Yuji Aramaki, Azuma Fujimoto, Toru Ogawa, Toshihiko Yamasaki, and Kiyoharu Aizawa. Sketch-based manga retrieval using manga109 dataset. *Multimedia Tools and Applications*, 76(20):21811–21838, 2017.

[9] Tobias Plotz and Stefan Roth. Benchmarking denoising algorithms with real photographs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1586–1595, 2017.

[10] Abdelrahman Abdelhamed, Stephen Lin, and Michael S Brown. A high-quality denoising dataset for smartphone cameras. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1692–1700, 2018.

[11] Zhihao Xia, Federico Perazzi, Michaël Gharbi, Kalyan Sunkavalli, and Ayan Chakrabarti. Basis prediction networks for effective burst denoising with large kernels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11844–11853, 2020.

[12] Ben Mildenhall, Jonathan T Barron, Jiawen Chen, Dillon Sharlet, Ren Ng, and Robert Carroll. Burst denoising with kernel prediction networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2502–2510, 2018.

[13] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. *arXiv preprint arXiv:1808.04560*, 2018.

[14] Vladimir Bychkovsky, Sylvain Paris, Eric Chan, and Frédo Durand. Learning photographic global tonal adjustment with a database of input/output image pairs. In *CVPR 2011*, pages 97–104. IEEE, 2011.

[15] Tan Kean Lai, Aymen F Abbas, Aliyu M Abdu, Usman U Sheikh, Musa Mokji, and Kamal Khalil. Super resolution of car plate images using generative adversarial networks. In *2019 IEEE 15th International Colloquium on Signal Processing & Its Applications (CSPA)*, pages 80–85. IEEE, 2019.

[16] Jin Chen, Jun Chen, Zheng Wang, Chao Liang, and Chia-Wen Lin. Identity-aware face super-resolution for low-resolution face recognition. *IEEE Signal Processing Letters*, 27: 645–649, 2020.

[17] Jacob Shermeyer and Adam Van Etten. The effects of super-resolution on object detection performance in satellite imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[18] Yuchong Gu, Zitao Zeng, Haibin Chen, Jun Wei, Yaqin Zhang, Binghui Chen, Yingqin Li, Yujuan Qin, Qing Xie, Zhuoren Jiang, et al. Medsrgan: medical images super-resolution using generative adversarial networks. *Multimedia Tools and Applications*, 79(29):21815–21840, 2020.

[19] Biao Xu, Zhiqiang Wang, and Jinping He. Super-resolution imaging via aperture modulation and intensity extrapolation. *Scientific reports*, 8(1):1–9, 2018.

[20] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 286–301, 2018.

[21] Pablo Arbelaez, Michael Maire, Charless Fowlkes, and Jitendra Malik. Contour detection and hierarchical image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 33(5):898–916, 2010.

[22] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *European conference on computer vision*, pages 184–199. Springer, 2014.

[23] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 624–632, 2017.

[24] Ying Tai, Jian Yang, Xiaoming Liu, and Chunyan Xu. Memnet: A persistent memory network for image restoration. In *Proceedings of the IEEE international conference on computer vision*, pages 4539–4547, 2017.

[25] Zheng Hui, Xiumei Wang, and Xinbo Gao. Fast and accurate single image super-resolution via information distillation network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 723–731, 2018.

[26] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 136–144, 2017.

[27] Kai Zhang, Wangmeng Zuo, and Lei Zhang. Learning a single convolutional super-resolution network for multiple degradations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3262–3271, 2018.

[28] Namhyuk Ahn, Byungkon Kang, and Kyung-Ah Sohn. Fast, accurate, and lightweight super-resolution with cascading residual network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 252–268, 2018.

[29] Juncheng Li, Faming Fang, Kangfu Mei, and Guixu Zhang. Multi-scale residual network for image super-resolution. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 517–532, 2018.

[30] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2472–2481, 2018.

[31] Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In *International conference on curves and surfaces*, pages 711–730. Springer, 2010.

[32] Juncheng Li, Faming Fang, Kangfu Mei, and Guixu Zhang. Multi-scale residual network for image super-resolution. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 517–532, 2018.

[33] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018.

[34] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1874–1883, 2016.

[35] Chao Dong, Chen Change Loy, and Xiaoou Tang. Accelerating the super-resolution convolutional neural network. In *European conference on computer vision*, pages 391–407. Springer, 2016.

[36] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 286–301, 2018.

[37] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4): 834–848, 2017.

[38] Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie Line Alberi-Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. 2012.

[39] Yiqun Mei, Yuchen Fan, Yuqian Zhou, Lichao Huang, Thomas S Huang, and Honghui Shi. Image super-resolution with cross-scale non-local attention and exhaustive self-exemplars mining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5690–5699, 2020.

[40] Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. Deep back-projection networks for super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1664–1673, 2018.

[41] Saeed Anwar and Nick Barnes. Densely residual laplacian super-resolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[42] Ben Niu, Weilei Wen, Wenqi Ren, Xiangde Zhang, Lianping Yang, Shuzhen Wang, Kaihao Zhang, Xiaochun Cao, and Haifeng Shen. Single image super-resolution via a holistic attention network. In *European Conference on Computer Vision*, pages 191–207. Springer, 2020.

[43] Jie Liu, Wenjie Zhang, Yuting Tang, Jie Tang, and Gangshan Wu. Residual feature aggregation network for image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2359–2368, 2020.

[44] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 286–301, 2018.

[45] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2472–2481, 2018.

[46] Zhen Li, Jinglei Yang, Zheng Liu, Xiaomin Yang, Gwanggil Jeon, and Wei Wu. Feedback network for image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3867–3876, 2019.

[47] Kai Zhang, Luc Van Gool, and Radu Timofte. Deep unfolding network for image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3217–3226, 2020.

[48] Yanyang Yan, Wenqi Ren, Xiaobin Hu, Kun Li, Haifeng Shen, and Xiaochun Cao. Srgat: Single image super-resolution with graph attention network. *IEEE Transactions on Image Processing*, 30:4905–4918, 2021.

[49] Yusuke Matsui, Kota Ito, Yuji Aramaki, Azuma Fujimoto, Toru Ogawa, Toshihiko Yamasaki, and Kiyoharu Aizawa. Sketch-based manga retrieval using manga109 dataset. *Multimedia Tools and Applications*, 76(20):21811–21838, 2017.

[50] Namhyuk Ahn, Byungkon Kang, and Kyung-Ah Sohn. Fast, accurate, and lightweight super-resolution with cascading residual network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 252–268, 2018.

[51] Jinjin Gu and Chao Dong. Interpreting super-resolution networks with local attribution maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9199–9208, 2021.

[52] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-order attention network for single image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11065–11074, 2019.

[53] Zheng Hui, Xinbo Gao, Yunchu Yang, and Xiumei Wang. Lightweight image super-resolution with information multi-distillation network. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 2024–2032, 2019.

[54] Kui Jiang, Zhongyuan Wang, Peng Yi, and Junjun Jiang. Hierarchical dense recursive network for image super-resolution. *Pattern Recognition*, 107:107475, 2020.

[55] Tao Lu, Yu Wang, Jiaming Wang, Wei Liu, and Yanduo Zhang. Single image super-resolution via multi-scale information polymerization network. *IEEE Signal Processing Letters*, 2021.

[56] Longguang Wang, Xiaoyu Dong, Yingqian Wang, Xinyi Ying, Zaiping Lin, Wei An, and Yulan Guo. Exploring sparsity in image super-resolution for efficient inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4917–4926, 2021.

[57] Jingwei Xin, Nannan Wang, Xinrui Jiang, Jie Li, Heng Huang, and Xinbo Gao. Binarized neural network for single image super resolution. In *European Conference on Computer Vision*, pages 91–107. Springer, 2020.

[58] Namhyuk Ahn, Byungkon Kang, and Kyung-Ah Sohn. Fast, accurate, and lightweight super-resolution with cascading residual network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 252–268, 2018.

[59] Zheng Hui, Xinbo Gao, Yunchu Yang, and Xiumei Wang. Lightweight image super-resolution with information multi-distillation network. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 2024–2032, 2019.

[60] Zheng Hui, Xiumei Wang, and Xinbo Gao. Fast and accurate single image super-resolution via information distillation network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 723–731, 2018.

[61] Xiaotong Luo, Yuan Xie, Yulun Zhang, Yanyun Qu, Cuihua Li, and Yun Fu. Latticenet: Towards lightweight image super-resolution with lattice block. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16*, pages 272–289. Springer, 2020.

[62] Xueyang Fu, Delu Zeng, Yue Huang, Xiao-Ping Zhang, and Xinghao Ding. A weighted variational model for simultaneous reflectance and illumination estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2782–2790, 2016.

[63] Zhenqiang Ying, Ge Li, Yurui Ren, Ronggang Wang, and Wenmin Wang. A new image contrast enhancement algorithm using exposure fusion framework. In *International Conference on Computer Analysis of Images and Patterns*, pages 36–46. Springer, 2017.

[64] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. *arXiv preprint arXiv:1808.04560*, 2018.

[65] Chunle Guo, Chongyi Li, Jichang Guo, Chen Change Loy, Junhui Hou, Sam Kwong, and Runmin Cong. Zero-reference deep curve estimation for low-light image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1780–1789, 2020.

[66] Yifan Jiang, Xinyu Gong, Ding Liu, Yu Cheng, Chen Fang, Xiaohui Shen, Jianchao Yang, Pan Zhou, and Zhangyang Wang. Enlightengan: Deep light enhancement without paired supervision. *IEEE Transactions on Image Processing*, 30:2340–2349, 2021.

[67] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE transactions on image processing*, 26(7):3142–3155, 2017.

[68] Shi Guo, Zifei Yan, Kai Zhang, Wangmeng Zuo, and Lei Zhang. Toward convolutional blind denoising of real photographs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1712–1722, 2019.

[69] Zongsheng Yue, Hongwei Yong, Qian Zhao, Lei Zhang, and Deyu Meng. Variational denoising network: Toward blind noise modeling and removal. *arXiv preprint arXiv:1908.11314*, 2019.

[70] Zongsheng Yue, Qian Zhao, Lei Zhang, and Deyu Meng. Dual adversarial network: Toward real-world noise removal and noise generation. In *European Conference on Computer Vision*, pages 41–58. Springer, 2020.

[71] Goutam Bhat, Martin Danelljan, and Radu Timofte. Ntire 2021 challenge on burst super-resolution: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 613–626, 2021.

[72] Goutam Bhat, Martin Danelljan, Radu Timofte, et al. NTIRE 2022 burst super-resolution challenge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2022.

[73] Michel Deudon, Alfredo Kalaitzis, Israel Goytom, Md Rifat Arefin, Zhichao Lin, Kris Sankaran, Vincent Michalski, Samira E Kahou, Julien Cornebise, and Yoshua Bengio. Highres-net: Recursive fusion for multi-frame super-resolution of satellite imagery. *arXiv preprint arXiv:2002.06460*, 2020.

[74] Goutam Bhat, Martin Danelljan, Fisher Yu, Luc Van Gool, and Radu Timofte. Deep reparametrization of multi-frame super-resolution and denoising. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2460–2470, 2021.

[75] Akshay Dudhane, Syed Waqas Zamir, Salman Khan, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Burst image restoration and enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5759–5768, 2022.

[76] Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. Learning to see in the dark. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3291–3300, 2018.

[77] Ahmet Serdar Karadeniz, Erkut Erdem, and Aykut Erdem. Burst photography for learning to enhance extremely dark images. *IEEE Transactions on Image Processing*, 30:9372–9385, 2021.

[78] Jingyun Liang, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Mutual affine network for spatially variant kernel estimation in blind image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4096–4105, 2021.

[79] Longguang Wang, Yingqian Wang, Xiaoyu Dong, Qingyu Xu, Jungang Yang, Wei An, and Yulan Guo. Unsupervised degradation representation learning for blind super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10581–10590, 2021.

[80] Kim Jun-Hyuk, C Jun-Ho, C Manri, and J Lee. Ram: Residual attention module for single image super-resolution. In *CVPR*, 2019.

[81] Yulun Zhang, Kunpeng Li, Kai Li, Bineng Zhong, and Yun Fu. Residual non-local attention networks for image restoration. *arXiv preprint arXiv:1903.10082*, 2019.

[82] Jiahui Yu, Yuchen Fan, Jianchao Yang, Ning Xu, Zhaowen Wang, Xinchao Wang, and Thomas Huang. Wide activation for efficient and accurate image super-resolution. *arXiv preprint arXiv:1808.08718*, 2018.

[83] Sefi Bell-Kligler, Assaf Shocher, and Michal Irani. Blind super-resolution kernel estimation using an internal-gan. *Advances in Neural Information Processing Systems*, 32, 2019.

[84] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5197–5206, 2015.

[85] Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4791–4800, 2021.

[86] Jinjin Gu, Hannan Lu, Wangmeng Zuo, and Chao Dong. Blind super-resolution with iterative kernel correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1604–1613, 2019.

[87] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012.

[88] Chao Ma, Chih-Yuan Yang, Xiaokang Yang, and Ming-Hsuan Yang. Learning a no-reference quality metric for single-image super-resolution. *Computer Vision and Image Understanding*, 158:1–16, 2017.

[89] Michal Irani and Shmuel Peleg. Improving resolution by image registration. *CVGIP: Graphical models and image processing*, 53(3):231–239, 1991.

[90] Raanan Fattal. Image upsampling via imposed edge statistics. In *ACM SIGGRAPH 2007 papers*, pages 95–es. 2007.

[91] Jian Sun, Zongben Xu, and Heung-Yeung Shum. Image super-resolution using gradient profile prior. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008.

[92] Qi Shan, Zhaorong Li, Jiaya Jia, and Chi-Keung Tang. Fast image/video upsampling. *ACM Transactions on Graphics (TOG)*, 27(5):1–7, 2008.

[93] Jinggang Huang and David Mumford. Statistics of natural images and models. In *Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149)*, volume 1, pages 541–547. IEEE, 1999.

[94] Kwang In Kim and Younghee Kwon. Single-image super-resolution using sparse regression and natural image prior. *IEEE transactions on pattern analysis and machine intelligence*, 32(6):1127–1133, 2010.

[95] Zhiwei Xiong, Xiaoyan Sun, and Feng Wu. Robust web image/video super-resolution. *IEEE transactions on image processing*, 19(8):2017–2028, 2010.

[96] Xinbo Gao, Kaibing Zhang, Dacheng Tao, and Xuelong Li. Image super-resolution with sparse neighbor embedding. *IEEE Transactions on Image Processing*, 21(7):3194–3205, 2012.

[97] Hong Chang, Dit-Yan Yeung, and Yimin Xiong. Super-resolution through neighbor embedding. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 1, pages I–I. IEEE, 2004.

[98] Daniel Glasner, Shai Bagon, and Michal Irani. Super-resolution from a single image. In *2009 IEEE 12th international conference on computer vision*, pages 349–356. IEEE, 2009.

[99] Jianchao Yang, Zhe Lin, and Scott Cohen. Fast image super-resolution based on in-place example regression. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1059–1066, 2013.

[100] Jian Sun, Nan-Ning Zheng, Hai Tao, and Heung-Yeung Shum. Image hallucination with primal sketch priors. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, volume 2, pages II–729. IEEE, 2003.

[101] Hong Chang, Dit-Yan Yeung, and Yimin Xiong. Super-resolution through neighbor embedding. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 1, pages I–I. IEEE, 2004.

[102] Karl S Ni and Truong Q Nguyen. Image superresolution using support vector regression. *IEEE Transactions on Image Processing*, 16(6):1596–1610, 2007.

[103] He He and Wan-Chi Siu. Single image super-resolution using gaussian process regression. In *CVPR 2011*, pages 449–456. IEEE, 2011.

[104] Kwang In Kim and Younghee Kwon. Single-image super-resolution using sparse regression and natural image prior. *IEEE transactions on pattern analysis and machine intelligence*, 32(6):1127–1133, 2010.

[105] Jianchao Yang, John Wright, Thomas S Huang, and Yi Ma. Image super-resolution via sparse representation. *IEEE transactions on image processing*, 19(11):2861–2873, 2010.

[106] Shuyuan Yang, Min Wang, Yiguang Chen, and Yaxin Sun. Single-image super-resolution reconstruction via learned geometric dictionaries and clustered sparse coding. *IEEE Transactions on Image Processing*, 21(9):4016–4028, 2012.

[107] Jianchao Yang, Zhe Lin, and Scott Cohen. Fast image super-resolution based on in-place example regression. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1059–1066, 2013.

[108] Qiang Wang, Xiaoou Tang, and Harry Shum. Patch based blind image super resolution. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 1, pages 709–716. IEEE, 2005.

[109] William T Freeman, Thouis R Jones, and Egon C Pasztor. Example-based super-resolution. *IEEE Computer graphics and Applications*, 22(2):56–65, 2002.

[110] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307, 2015.

[111] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Deeply-recursive convolutional network for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1637–1645, 2016.

[112] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[113] Ying Tai, Jian Yang, and Xiaoming Liu. Image super-resolution via deep recursive residual network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3147–3155, 2017.

[114] Zheng Hui, Xiumei Wang, and Xinbo Gao. Fast and accurate single image super-resolution via information distillation network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 723–731, 2018.

[115] Juncheng Li, Faming Fang, Jiaqian Li, Kangfu Mei, and Guixu Zhang. Mdcn: Multi-scale dense cross network for image super-resolution. *IEEE Transactions on Circuits and Systems for Video Technology*, 2020.

[116] Armin Mehri, Parichehr B Ardakani, and Angel D Sappa. Mprnet: Multi-path residual network for lightweight image super resolution. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2704–2713, 2021.

[117] Zheng Hui, Jie Li, Xinbo Gao, and Xiumei Wang. Progressive perception-oriented network for single image super-resolution. *Information Sciences*, 546:769–786, 2021.

[118] Anqi Liu, Sumei Li, and Yongli Chang. Image super-resolution using progressive residual multi-dilated aggregation network. *Signal, Image and Video Processing*, pages 1–9, 2022.

[119] Kalpesh Prajapati, Vishal Chudasama, Heena Patel, Kishor Upla, Kiran Raja, Raghavendra Ramachandra, and Christoph Busch. Direct unsupervised super-resolution using generative adversarial network (dus-gan) for real-world data. *IEEE Transactions on Image Processing*, 30:8251–8264, 2021.

[120] Lidong Song, Yiyan Li, and Ning Lu. Profilesr-gan: A gan based super-resolution method for generating high-resolution load profiles. *IEEE Transactions on Smart Grid*, 2022.

[121] Yuxiang Yang, Qi Cao, Jing Zhang, and Dacheng Tao. Codon: On orchestrating cross-domain attentions for depth super-resolution. *International Journal of Computer Vision*, pages 1–18, 2022.

[122] Xiang Lv, Changzhong Wang, Xiaodong Fan, Qiangkui Leng, and Xiaoli Jiang. A novel image super-resolution algorithm based on multi-scale dense recursive fusion network. *Neurocomputing*, 489:98–111, 2022.

[123] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

[124] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[125] Yanting Hu, Jie Li, Yuanfei Huang, and Xinbo Gao. Channel-wise and spatial feature modulation network for single image super-resolution. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(11):3911–3927, 2019.

[126] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018.

[127] Xiaobiao Du, Jie Niu, and Chongjin Liu. Expectation-maximization attention cross residual network for single image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 888–896, 2021.

[128] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12299–12310, 2021.

[129] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1833–1844, 2021.

[130] Yunjin Chen and Thomas Pock. Trainable nonlinear reaction diffusion: A flexible framework for fast and effective image restoration. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1256–1272, 2016.

[131] Nicola Asuni and Andrea Giachetti. Accuracy improvements and artifacts removal in edge based image interpolation. *VISAPP (1)*, 8:58–65, 2008.

[132] Chao Dong, Chen Change Loy, and Xiaoou Tang. Accelerating the super-resolution convolutional neural network. In *European conference on computer vision*, pages 391–407. Springer, 2016.

[133] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1874–1883, 2016.

[134] Augustus Odena, Vincent Dumoulin, and Chris Olah. Deconvolution and checkerboard artifacts. *Distill*, 1(10):e3, 2016.

[135] Zhuangzi Li. Image super-resolution using attention based densenet with residual deconvolution. *arXiv preprint arXiv:1907.05282*, 2019.

[136] Dongliang Xiong, Kai Huang, Siang Chen, Bowen Li, Haitian Jiang, and Wenyuan Xu. Noucsr: Efficient super-resolution network without upsampling convolution. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 3378–3387. IEEE, 2019.

[137] Xuecai Hu, Haoyuan Mu, Xiangyu Zhang, Zilei Wang, Tieniu Tan, and Jian Sun. Meta-sr: A magnification-arbitrary network for super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1575–1584, 2019.

[138] Yutong Dai, Hao Lu, and Chunhua Shen. Learning affinity-aware upsampling for deep image matting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6841–6850, 2021.

[139] Athanasios Papoulis. Generalized sampling expansion. *IEEE transactions on circuits and systems*, 24(11):652–654, 1977.

[140] R Tsai. Multiframe image restoration and registration. *Advance Computer Visual and Image Processing*, 1:317–339, 1984.

[141] Michael Elad and Yacov Hel-Or. A fast super-resolution reconstruction algorithm for pure translational motion and common space-invariant blur. *IEEE Transactions on image Processing*, 10(8):1187–1193, 2001.

[142] Michal Irani and Shmuel Peleg. Improving resolution by image registration. *CVGIP: Graphical models and image processing*, 53(3):231–239, 1991.

[143] Shmuel Peleg, Danny Keren, and Limor Schweitzer. Improving image resolution using subpixel motion. *Pattern recognition letters*, 5(3):223–226, 1987.

[144] Benedicte Bascle, Andrew Blake, and Andrew Zisserman. Motion deblurring and super-resolution from an image sequence. In *European conference on computer vision*, pages 571–582. Springer, 1996.

[145] Michael Elad and Arie Feuer. Restoration of a single superresolution image from several blurred, noisy, and undersampled measured images. *IEEE transactions on image processing*, 6(12):1646–1658, 1997.

[146] Russell C Hardie, Kenneth J Barnard, John G Bognar, Ernest E Armstrong, and Edward A Watson. High-resolution image reconstruction from a sequence of rotated and translated frames and its application to an infrared imaging system. *Optical Engineering*, 37(1):247–260, 1998.

[147] Assaf Zomet, Alex Rav-Acha, and Shmuel Peleg. Robust super-resolution. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 1, pages I–I. IEEE, 2001.

[148] Sina Farsiu, Michael Elad, and Peyman Milanfar. Multiframe demosaicing and super-resolution from undersampled color images. In *Computational Imaging II*, volume 5299, pages 222–233. SPIE, 2004.

[149] Esmaeil Faramarzi, Dinesh Rajan, and Marc P Christensen. Unified blind method for multi-image super-resolution and single/multi-image blur deconvolution. *IEEE Transactions on Image Processing*, 22(6):2101–2114, 2013.

[150] Yu He, Kim-Hui Yap, Li Chen, and Lap-Pui Chau. A nonlinear least square technique for simultaneous image registration and super-resolution. *IEEE Transactions on Image Processing*, 16(11):2830–2841, 2007.

[151] Sina Farsiu, Michael Elad, and Peyman Milanfar. Multiframe demosaicing and super-resolution from undersampled color images. In *Computational Imaging II*, volume 5299, pages 222–233. SPIE, 2004.

[152] Bartlomiej Wronski, Ignacio Garcia-Dorado, Manfred Ernst, Damien Kelly, Michael Krainin, Chia-Kai Liang, Marc Levoy, and Peyman Milanfar. Handheld multi-frame super-resolution. *ACM Transactions on Graphics (TOG)*, 38(4):1–18, 2019.

[153] Andrea Bordone Molini, Diego Valsesia, Giulia Fracastoro, and Enrico Magli. Deepsum: Deep neural network for super-resolution of unregistered multitemporal images. *IEEE Transactions on Geoscience and Remote Sensing*, 58(5):3644–3656, 2019.

[154] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *European conference on computer vision*, pages 184–199. Springer, 2014.

[155] Kai Zhang, Luc Van Gool, and Radu Timofte. Deep unfolding network for image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3217–3226, 2020.

[156] Jinjin Gu, Hannan Lu, Wangmeng Zuo, and Chao Dong. Blind super-resolution with iterative kernel correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1604–1613, 2019.

[157] Yu-Syuan Xu, Shou-Yao Roy Tseng, Yu Tseng, Hsien-Kai Kuo, and Yi-Min Tsai. Unified dynamic convolutional network for super-resolution with variational degradations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12496–12505, 2020.

[158] Yu He, Kim-Hui Yap, Li Chen, and Lap-Pui Chau. A soft map framework for blind super-resolution image reconstruction. *Image and Vision Computing*, 27(4):364–373, 2009.

[159] Wen-Ze Shao and Michael Elad. Simple, accurate, and robust nonparametric blind super-resolution. In *Image and Graphics*, pages 333–348. Springer, 2015.

[160] Tomer Michaeli and Michal Irani. Nonparametric blind super-resolution. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 945–952, 2013.

[161] Anat Levin, Yair Weiss, Fredo Durand, and William T Freeman. Understanding and evaluating blind deconvolution algorithms. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1964–1971. IEEE, 2009.

[162] Anat Levin, Yair Weiss, Fredo Durand, and William T Freeman. Efficient marginal likelihood optimization in blind deconvolution. In *CVPR 2011*, pages 2657–2664. IEEE, 2011.

[163] Yunxuan Wei, Shuhang Gu, Yawei Li, Radu Timofte, Longcun Jin, and Hengjie Song. Unsupervised real-world image super resolution via domain-distance aware training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13385–13394, 2021.

[164] Yuan Yuan, Siyuan Liu, Jiawei Zhang, Yongbing Zhang, Chao Dong, and Liang Lin. Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 701–710, 2018.

[165] Andreas Lugmayr, Martin Danelljan, and Radu Timofte. Unsupervised learning for real-world super-resolution. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 3408–3416. IEEE, 2019.

[166] Shunta Maeda. Unpaired image super-resolution using pseudo-supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 291–300, 2020.

[167] Assaf Shocher, Nadav Cohen, and Michal Irani. "zero-shot" super-resolution using deep internal learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3118–3126, 2018.

[168] Jae Woong Soh, Sunwoo Cho, and Nam Ik Cho. Meta-transfer learning for zero-shot super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3516–3525, 2020.

[169] Sefi Bell-Kligler, Assaf Shocher, and Michal Irani. Blind super-resolution kernel estimation using an internal-gan. *Advances in Neural Information Processing Systems*, 32, 2019.

[170] Jingyun Liang, Kai Zhang, Shuhang Gu, Luc Van Gool, and Radu Timofte. Flow-based kernel prior with application to blind super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10601–10610, 2021.

[171] Guangpin Tao, Xiaozhong Ji, Wenzhuo Wang, Shuo Chen, Chuming Lin, Yun Cao, Tong Lu, Donghao Luo, and Ying Tai. Spectrum-to-kernel translation for accurate blind image super-resolution. *Advances in Neural Information Processing Systems*, 34:22643–22654, 2021.

[172] Jinjin Gu, Hannan Lu, Wangmeng Zuo, and Chao Dong. Blind super-resolution with iterative kernel correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1604–1613, 2019.

[173] Yan Huang, Shang Li, Liang Wang, Tieniu Tan, et al. Unfolding the alternating optimization for blind super resolution. *Advances in Neural Information Processing Systems*, 33: 5632–5643, 2020.

[174] Jingyun Liang, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Mutual affine network for spatially variant kernel estimation in blind image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4096–4105, 2021.

[175] Hongyi Zheng, Hongwei Yong, and Lei Zhang. Unfolded deep kernel estimation for blind image super-resolution. In *European Conference on Computer Vision*, pages 502–518. Springer, 2022.

[176] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-order attention network for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11065–11074, 2019.

[177] Yiqun Mei, Yuchen Fan, and Yuqian Zhou. Image super-resolution with non-local sparse attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3517–3526, 2021.

[178] Faming Fang, Juncheng Li, and Tieyong Zeng. Soft-edge assisted network for single image super-resolution. *IEEE Transactions on Image Processing*, 29:4656–4668, 2020.

[179] Alireza Esmaeilzehi, M Omair Ahmad, and MNS Swamy. Srnssi: A deep light-weight network for single image super resolution using spatial and spectral information. *IEEE Transactions on Computational Imaging*, 7:409–421, 2021.

[180] Augustus Odena, Vincent Dumoulin, and Chris Olah. Deconvolution and checkerboard artifacts. *Distill*, 1(10):e3, 2016.

[181] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1874–1883, 2016.

[182] Xiang Li, Wenhai Wang, Xiaolin Hu, and Jian Yang. Selective kernel networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 510–519, 2019.

[183] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[184] Andrew Aitken, Christian Ledig, Lucas Theis, Jose Caballero, Zehan Wang, and Wenzhe Shi. Checkerboard artifact free sub-pixel convolution: A note on sub-pixel convolution, resize convolution and convolution resize. *arXiv preprint arXiv:1707.02937*, 2017.

[185] Zhen Li, Jinglei Yang, Zheng Liu, Xiaomin Yang, Gwanggil Jeon, and Wei Wu. Feedback network for image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3867–3876, 2019.

[186] Yuqing Liu, Xinfeng Zhang, Shanshe Wang, Siwei Ma, and Wen Gao. Sequential hierarchical learning with distribution transformation for image super-resolution. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 2022.

[187] Mingjin Zhang, Qianqian Wu, Jing Zhang, Xinbo Gao, Jie Guo, and Dacheng Tao. Fluid micelle network for image super-resolution reconstruction. *IEEE Transactions on Cybernetics*, 2022.

[188] Mingjin Zhang, Jingwei Xin, Jing Zhang, Dacheng Tao, and Xinbo Gao. Curvature consistent network for microscope chip image super-resolution. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[189] Dongliang Xiong, Kai Huang, Siang Chen, Bowen Li, Haitian Jiang, and Wenyuan Xu. Noucsr: Efficient super-resolution network without upsampling convolution. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 3378–3387. IEEE, 2019.

[190] Zhuangzi Li. Image super-resolution using attention based densenet with residual deconvolution. *arXiv preprint arXiv:1907.05282*, 2019.

[191] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu 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.

[192] Ying Tai, Jian Yang, and Xiaoming Liu. Image super-resolution via deep recursive residual network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3147–3155, 2017.

[193] Xinyan Zhang, Peng Gao, Sunxiangyu Liu, Kongya Zhao, Guitao Li, Liuguo Yin, and Chang Wen Chen. Accurate and efficient image super-resolution via global-local adjusting dense network. *IEEE Transactions on Multimedia*, 2020.

[194] Haoyu Ren, Mostafa El-Khamy, and Jungwon Lee. Image super resolution based on fusing multiple convolution neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 54–61, 2017.

[195] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[196] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016.

[197] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *Advances in neural information processing systems*, pages 5767–5777, 2017.

[198] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1520–1528, 2015.

[199] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[200] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[201] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[202] Chong Mou, Jian Zhang, Xiaopeng Fan, Hangfan Liu, and Ronggang Wang. Cola-net: Collaborative attention network for image restoration. *arXiv preprint arXiv:2103.05961*, 2021.

[203] Yajun Qiu, Ruxin Wang, Dapeng Tao, and Jun Cheng. Embedded block residual network: A recursive restoration model for single-image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4180–4189, 2019.

[204] Xiangtai Li, Xia Li, Li Zhang, Guangliang Cheng, Jianping Shi, Zhouchen Lin, Shaohua Tan, and Yunhai Tong. Improving semantic segmentation via decoupled body and edge supervision. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*, pages 435–452. Springer, 2020.

[205] Xin Luo, Wen Qin, Ani Dong, Khaled Sedraoui, and MengChu Zhou. Efficient and high-quality recommendations via momentum-incorporated parallel stochastic gradient descent-based learning. *IEEE/CAA Journal of Automatica Sinica*, 8(2):402–411, 2020.

[206] Xiang Li, Wenhai Wang, Xiaolin Hu, and Jian Yang. Selective kernel networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 510–519, 2019.

[207] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[208] Xiangyu He, Zitao Mo, Peisong Wang, Yang Liu, Mingyuan Yang, and Jian Cheng. Ode-inspired network design for single image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1732–1741, 2019.

[209] Jiqing Zhang, Chengjiang Long, Yuxin Wang, Haiyin Piao, Haiyang Mei, Xin Yang, and Baocai Yin. A two-stage attentive network for single image super-resolution. *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

[210] Yuanfei Huang, Jie Li, Xinbo Gao, Yanting Hu, and Wen Lu. Interpretable detail-fidelity attention network for single image super-resolution. *IEEE Transactions on Image Processing*, 30:2325–2339, 2021.

[211] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 624–632, 2017.

[212] Kai Zhang, Wangmeng Zuo, and Lei Zhang. Learning a single convolutional super-resolution network for multiple degradations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3262–3271, 2018.

[213] Xiaotong Luo, Yuan Xie, Yulun Zhang, Yanyun Qu, Cuihua Li, and Yun Fu. Latticenet: Towards lightweight image super-resolution with lattice block. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16*, pages 272–289. Springer, 2020.

[214] Rushi Lan, Long Sun, Zhenbing Liu, Huimin Lu, Cheng Pang, and Xiaonan Luo. Madnet: A fast and lightweight network for single-image super resolution. *IEEE transactions on cybernetics*, 51(3):1443–1453, 2020.

[215] Biao Li, Bo Wang, Jiabin Liu, Zhiquan Qi, and Yong Shi. s-lwsr: Super lightweight super-resolution network. *IEEE Transactions on Image Processing*, 29:8368–8380, 2020.

[216] Yulun Zhang, Kunpeng Li, Kai Li, Bineng Zhong, and Yun Fu. Residual non-local attention networks for image restoration. *arXiv preprint arXiv:1903.10082*, 2019.

[217] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 136–144, 2017.

[218] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[219] Yanting Hu, Xinbo Gao, Jie Li, Yuanfei Huang, and Hanzi Wang. Single image super-resolution with multi-scale information cross-fusion network. *Signal Processing*, 179: 107831, 2021.

[220] Chunwei Tian, Yong Xu, Wangmeng Zuo, Chia-Wen Lin, and David Zhang. Asymmetric cnn for image superresolution. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2021.

[221] Chunwei Tian, Yong Xu, Wangmeng Zuo, Bob Zhang, Lunke Fei, and Chia-Wen Lin. Coarse-to-fine cnn for image super-resolution. *IEEE Transactions on Multimedia*, 23: 1489–1502, 2020.

[222] Chunwei Tian, Ruibin Zhuge, Zhihao Wu, Yong Xu, Wangmeng Zuo, Chen Chen, and Chia-Wen Lin. Lightweight image super-resolution with enhanced cnn. *Knowledge-Based Systems*, 205:106235, 2020.

[223] Ryo Nakagaki and Aggelos K Katsaggelos. A vq-based blind image restoration algorithm. *IEEE transactions on image processing*, 12(9):1044–1053, 2003.

[224] Jianbo Jiao, Wei-Chih Tu, Ding Liu, Shengfeng He, Rynson WH Lau, and Thomas S Huang. Formnet: Formatted learning for image restoration. *IEEE Transactions on Image Processing*, 29:6302–6314, 2020.

[225] Kuldeep Purohit, Maitreya Suin, AN Rajagopalan, and Vishnu Naresh Boddeti. Spatially-adaptive image restoration using distortion-guided networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2309–2319, 2021.

[226] Panqu Wang, Pengfei Chen, Ye Yuan, Ding Liu, Zehua Huang, Xiaodi Hou, and Garrison Cottrell. Understanding convolution for semantic segmentation. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 1451–1460. IEEE, 2018.

[227] Mohammadreza Mostajabi, Payman Yadollahpour, and Gregory Shakhnarovich. Feedforward semantic segmentation with zoom-out features. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3376–3385, 2015.

[228] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[229] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Deeply-recursive convolutional network for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1637–1645, 2016.

[230] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu 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.

[231] Ying Tai, Jian Yang, Xiaoming Liu, and Chunyan Xu. Memnet: A persistent memory network for image restoration. In *Proceedings of the IEEE international conference on computer vision*, pages 4539–4547, 2017.

[232] Rushi Lan, Long Sun, Zhenbing Liu, Huimin Lu, Cheng Pang, and Xiaonan Luo. Madnet: A fast and lightweight network for single-image super resolution. *IEEE Transactions on Cybernetics*, 2020.

[233] Parichehr Behjati, Pau Rodriguez, Armin Mehri, Isabelle Hupont, Carles Fernandez Tena, and Jordi Gonzalez. Overnet: Lightweight multi-scale super-resolution with overscaling network. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2694–2703, 2021.

[234] Jongchan Park, Joon-Young Lee, Donggeun Yoo, and In So Kweon. Distort-and-recover: Color enhancement using deep reinforcement learning. In *Proceedings of the IEEE Conference on computer vision and pattern recognition*, pages 5928–5936, 2018.

[235] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

[236] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. *arXiv preprint arXiv:1808.04560*, 2018.

[237] Wenjing Wang, Chen Wei, Wenhan Yang, and Jiaying Liu. Gladnet: Low-light enhancement network with global awareness. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 751–755. IEEE, 2018.

[238] Yonghua Zhang, Jiawan Zhang, and Xiaojie Guo. Kindling the darkness: A practical low-light image enhancer. In *Proceedings of the 27th ACM international conference on multimedia*, pages 1632–1640, 2019.

[239] Wenhan Yang, Shiqi Wang, Yuming Fang, Yue Wang, and Jiaying Liu. From fidelity to perceptual quality: A semi-supervised approach for low-light image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3063–3072, 2020.

[240] Ke Xu, Xin Yang, Baocai Yin, and Rynson WH Lau. Learning to restore low-light images via decomposition-and-enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2281–2290, 2020.

[241] Ruixing Wang, Qing Zhang, Chi-Wing Fu, Xiaoyong Shen, Wei-Shi Zheng, and Jiaya Jia. Underexposed photo enhancement using deep illumination estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6849–6857, 2019.

[242] Long Ma, Risheng Liu, Jiaao Zhang, Xin Fan, and Zhongxuan Luo. Learning deep context-sensitive decomposition for low-light image enhancement. *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[243] Xinxu Wei, Xianshi Zhang, Shisen Wang, Cheng Cheng, Yanlin Huang, Kaifu Yang, and Yongjie Li. Blnet: A fast deep learning framework for low-light image enhancement with noise removal and color restoration. *arXiv preprint arXiv:2106.15953*, 2021.

[244] Risheng Liu, Long Ma, Jiaao Zhang, Xin Fan, and Zhongxuan Luo. Retinex-inspired unrolling with cooperative prior architecture search for low-light image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10561–10570, 2021.

[245] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE transactions on image processing*, 26(7):3142–3155, 2017.

[246] Daniel Zoran and Yair Weiss. From learning models of natural image patches to whole image restoration. In *2011 International Conference on Computer Vision*, pages 479–486. IEEE, 2011.

[247] Yunjin Chen and Thomas Pock. Trainable nonlinear reaction diffusion: A flexible framework for fast and effective image restoration. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1256–1272, 2016.

[248] Kai Zhang, Wangmeng Zuo, and Lei Zhang. Ffdnet: Toward a fast and flexible solution for cnn-based image denoising. *IEEE Transactions on Image Processing*, 27(9):4608–4622, 2018.

[249] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. Image denoising by sparse 3-d transform-domain collaborative filtering. *IEEE Transactions on image processing*, 16(8):2080–2095, 2007.

[250] Marc Lebrun, Miguel Colom, and Jean-Michel Morel. The noise clinic: a blind image denoising algorithm. *Image Processing On Line*, 5:1–54, 2015.

[251] Michal Aharon, Michael Elad, and Alfred Bruckstein. K-svd: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on signal processing*, 54(11):4311–4322, 2006.

[252] Saeed Anwar and Nick Barnes. Real image denoising with feature attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3155–3164, 2019.

[253] Yiyun Zhao, Zhuqing Jiang, Aidong Men, and Guodong Ju. Pyramid real image denoising network. In *2019 IEEE Visual Communications and Image Processing (VCIP)*, pages 1–4. IEEE, 2019.

[254] Yuda Song, Yunfang Zhu, and Xin Du. Dynamic residual dense network for image denoising. *Sensors*, 19(17):3809, 2019.

[255] Yang Liu, Saeed Anwar, Liang Zheng, and Qi Tian. Gradnet image denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 508–509, 2020.

[256] Yoonsik Kim, Jae Woong Soh, Gu Yong Park, and Nam Ik Cho. Transfer learning from synthetic to real-noise denoising with adaptive instance normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3482–3492, 2020.

[257] Yang Liu, Zhenyue Qin, Saeed Anwar, Pan Ji, Dongwoo Kim, Sabrina Caldwell, and Tom Gedeon. Invertible denoising network: A light solution for real noise removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13365–13374, 2021.

[258] Mauricio Delbracio, Damien Kelly, Michael S Brown, and Peyman Milanfar. Mobile computational photography: A tour. *arXiv:2102.09000*, 2021.

[259] Ahmet Serdar Karadeniz, Erkut Erdem, and Aykut Erdem. Burst photography for learning to enhance extremely dark images. *IEEE Transactions on Image Processing*, 30:9372–9385, 2021.

[260] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 286–301, 2018.

[261] Pengxu Wei, Ziwei Xie, Hannan Lu, Zongyuan Zhan, Qixiang Ye, Wangmeng Zuo, and Liang Lin. Component divide-and-conquer for real-world image super-resolution. In *European Conference on Computer Vision*, pages 101–117. Springer, 2020.

[262] Bartlomiej Wronski, Ignacio Garcia-Dorado, Manfred Ernst, Damien Kelly, Michael Krainin, Chia-Kai Liang, Marc Levoy, and Peyman Milanfar. Handheld multi-frame super-resolution. *ACM Transactions on Graphics (TOG)*, 38(4):1–18, 2019.

[263] Bartlomiej Wronski, Ignacio Garcia-Dorado, Manfred Ernst, Damien Kelly, Michael Krainin, Chia-Kai Liang, Marc Levoy, and Peyman Milanfar. Handheld multi-frame super-resolution. *ACM Transactions on Graphics (TOG)*, 38(4):1–18, 2019.

[264] Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. Deep burst super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9209–9218, 2021.

[265] Bruno Lecouat, Jean Ponce, and Julien Mairal. Lucas-kanade reloaded: End-to-end super-resolution from raw image bursts. In *ICCV*, 2021.

[266] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9308–9316, 2019.

[267] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. *arXiv preprint arXiv:2111.09881*, 2021.

[268] Akshay Dudhane, Syed Waqas Zamir, Salman Khan, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Burst image restoration and enhancement. In *CVPR*, 2022.

[269] Andrey Ignatov, Luc Van Gool, and Radu Timofte. Replacing mobile camera isp with a single deep learning model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 536–537, 2020.

[270] Goutam Bhat, Martin Danelljan, Fisher Yu, Luc Van Gool, and Radu Timofte. Deep reparametrization of multi-frame super-resolution and denoising. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2460–2470, 2021.

[271] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

[272] Ziwei Luo, Lei Yu, Xuan Mo, Youwei Li, Lanpeng Jia, Haoqiang Fan, Jian Sun, and Shuaicheng Liu. Ebsr: Feature enhanced burst super-resolution with deformable alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 471–478, 2021.

[273] Michel Deudon, Alfredo Kalaitzis, Israel Goytom, Md Rifat Arefin, Zhichao Lin, Kris Sankaran, Vincent Michalski, Samira E Kahou, Julien Cornebise, and Yoshua Bengio. Highres-net: Recursive fusion for multi-frame super-resolution of satellite imagery. *arXiv preprint arXiv:2002.06460*, 2020.

[274] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *CVPRW*, 2017.

[275] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[276] Chen Chen, Qifeng Chen, Minh N Do, and Vladlen Koltun. Seeing motion in the dark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3185–3194, 2019.

[277] Yapeng Tian, Yulun Zhang, Yun Fu, and Chenliang Xu. Tdan: Temporally-deformable alignment network for video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3360–3369, 2020.

[278] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[279] Zhenli Zhang, Xiangyu Zhang, Chao Peng, Xiangyang Xue, and Jian Sun. Exfuse: Enhancing feature fusion for semantic segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 269–284, 2018.

[280] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 624–632, 2017.

[281] Ralph E Carlson and Frederick N Fritsch. Monotone piecewise bicubic interpolation. *SIAM journal on numerical analysis*, 22(2):386–400, 1985.

[282] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1874–1883, 2016.

[283] Wooyeong Cho, Sanghyeok Son, and Dae-Shik Kim. Weighted multi-kernel prediction network for burst image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 404–413, 2021.

[284] Paras Maharjan, Li Li, Zhu Li, Ning Xu, Chongyang Ma, and Yue Li. Improving extreme low-light image denoising via residual learning. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 916–921. IEEE, 2019.

[285] Syed Waqas Zamir, Aditya Arora, Salman Khan, Fahad Shahbaz Khan, and Ling Shao. Learning digital camera pipeline for extreme low-light imaging. *Neurocomputing*, 452:37–47, 2021.

[286] Di Zhao, Lan Ma, Songnan Li, and Dahai Yu. End-to-end denoising of dark burst images using recurrent fully convolutional networks. *arXiv preprint arXiv:1904.07483*, 2019.

[287] Netalee Efrat, Daniel Glasner, Alexander Apartsin, Boaz Nadler, and Anat Levin. Accurate blur models vs. image priors in single image super-resolution. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2832–2839, 2013.

[288] Yifeng Zhou, Chuming Lin, Donghao Luo, Yong Liu, Ying Tai, Chengjie Wang, and Mingang Chen. Joint learning content and degradation aware feature for blind super-resolution. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 2606–2616, 2022.

[289] Dongwei Ren, Kai Zhang, Qilong Wang, Qinghua Hu, and Wangmeng Zuo. Neural blind deconvolution using deep priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3341–3350, 2020.

[290] Kai Zhang, Wangmeng Zuo, and Lei Zhang. Learning a single convolutional super-resolution network for multiple degradations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3262–3271, 2018.

[291] Yosef Gandelsman, Assaf Shocher, and Michal Irani. "double-dip": Unsupervised image decomposition via coupled deep-image-priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11026–11035, 2019.

[292] Mehmet Yamac, Baran Ataman, and Aakif Nawaz. Kernelnet: A blind super-resolution kernel estimation network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 453–462, 2021.

[293] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[294] Gernot Riegler, Samuel Schulter, Matthias Ruther, and Horst Bischof. Conditioned regression models for non-blind single image super-resolution. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 522–530, 2015.

[295] Sefi Bell-Kligler, Assaf Shocher, and Michal Irani. Blind super-resolution kernel estimation using an internal-gan. *Advances in Neural Information Processing Systems*, 32, 2019.

[296] Wen-Ze Shao and Michael Elad. Simple, accurate, and robust nonparametric blind super-resolution. In *Image and Graphics*, pages 333–348. Springer, 2015.

[297] Assaf Shocher, Nadav Cohen, and Michal Irani. "zero-shot" super-resolution using deep internal learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3118–3126, 2018.

[298] Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, and Lei Zhang. Ntire 2017 challenge on single image super-resolution: Methods and results. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 114–125, 2017.

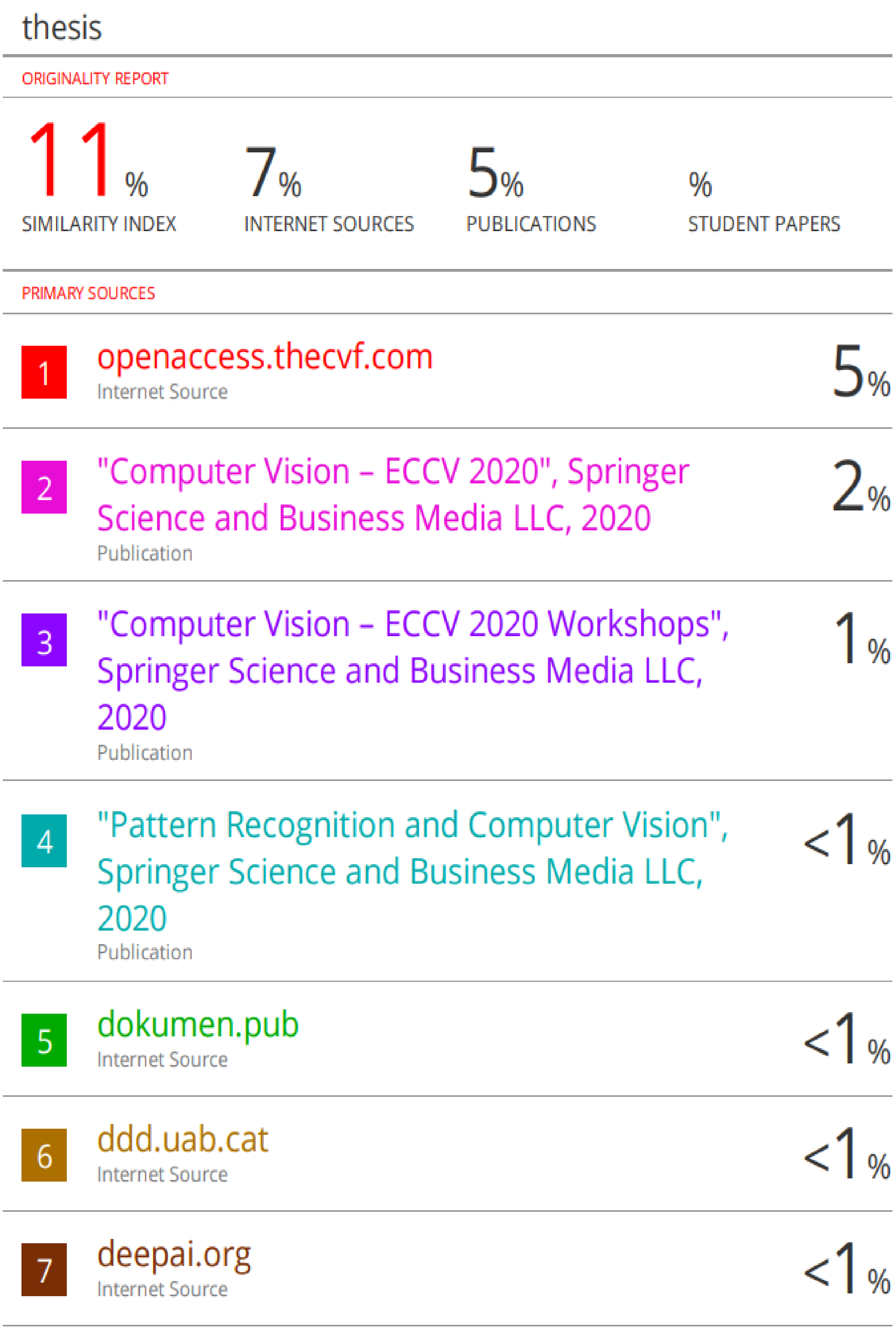
[299] Shady Abu Hussein, Tom Tirer, and Raja Giryes. Correction filter for single image super-resolution: Robustifying off-the-shelf deep super-resolvers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1428–1437, 2020.

[300] Victor Cornillere, Abdelaziz Djelouah, Wang Yifan, Olga Sorkine-Hornung, and Christopher Schroers. Blind image super-resolution with spatially variant degradations. *ACM Transactions on Graphics (TOG)*, 38(6):1–13, 2019.

[301] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9446–9454, 2018.

[302] Feng Li, Yixuan Wu, Huihui Bai, Weisi Lin, Runmin Cong, and Yao Zhao. Learning detail-structure alternative optimization for blind super-resolution. *arXiv preprint arXiv:2212.01624*, 2022.

Thesis Plagiarism



thesis

ORIGINALITY REPORT

11%	7%	5%	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	openaccess.thecvf.com Internet Source	5%
2	"Computer Vision – ECCV 2020", Springer Science and Business Media LLC, 2020 Publication	2%
3	"Computer Vision – ECCV 2020 Workshops", Springer Science and Business Media LLC, 2020 Publication	1%
4	"Pattern Recognition and Computer Vision", Springer Science and Business Media LLC, 2020 Publication	<1%
5	dokumen.pub Internet Source	<1%
6	ddd.uab.cat Internet Source	<1%
7	deepai.org Internet Source	<1%