Universitat Autònoma de Barcelona

Deep learning based architectures for cross-domain image processing

Armin Mehri

Universitat Autònoma
de Barcelona

Deep learning based architectures for cross-domain image processing

A dissertation submitted by **Armin Mehri** at Universitat Autònoma de Barcelona to fulfil the degree of **Doctor of Philosophy**.

Bellaterra, December 5, 2022

Director

Dr. Angel Domingo Sappa
Dept. Ciències de la computació & Centre de Visió per Computador
Universitat Autònoma de Barcelona (UAB), Spain

Thesis committee

Dr. Petia Ivanova Radeva
Dept. Matemáticas e Informática
Universidad de Barcelona, Spain

Dr. Daniel Ponsa Mussarra
Dept. Ciències de la computació & Centre de Visió per Computador
Universitat Autònoma de Barcelona (UAB), Spain

Dr. David Masip Rodó
Dept. Informàtica Multimedia y Telecomunicación
Universidad Oberta de Catalunya, Barcelona Spain

This document was typeset by the author using LaTeX2$_\varepsilon$.

Printed by Ediciones Gráficas Rey, S.L.

To my family, and friends…

Acknowledgments

Working as a Ph.D. student was both a wonderful and difficult experience for me. It has been a stage of challenges, fun and discovery. In all these years, several people have directly or indirectly influenced my academic career. Without their invaluable assistance, it would have been nearly impossible for me to flourish in my PhD study. I hope not to forget anyone.

First and foremost, my deepest appreciation and many thanks go to my supervisor, Angel Sappa, for believing in me from the beginning and allowing me to fulfill my dream. Not many Ph.D. candidates are not lucky as I am to have a supportive learning environment, intellectual freedom, and valuable advice that he has provided. Without his help, advice, expertise, and encouragement this research and dissertation would not have happened. Thank you for your support, discussions, and the freedom you gave me during this journey.

I would like to thank my friends at the Computer Vision Center: Jacopo, Ali, Andres, Sounak, and Sanket, with whom I shared many moments and made my way through the CVC memorable.

I also would like to thank everyone in the CVC administration for their hard work, commitment, help, and sympathy throughout the years. Thank you very much for providing hospitality. I also would like to express my sincere gratitude, particularly to Montse, Jordi, Gisele, and Encarna for all their kind help during this research.

I must thank my best friends for all their support: Mohammad Reza, Shahram, Arash, Parichehr, Javad, Alireza, Masoud, and Mohammad Amin, I am grateful for your friendship and hope that our friendship will continue through good times and bad.

Last but not least, my warm and heartfelt thanks go to my beloved parents and brother for the tremendous support and hope they have given to me during all these years of my study. Without that hope and support, this research would not have been possible. Thank you for the strength you gave me. I love you all!

Abstract

Human vision is restricted to the visual-optical spectrum. Machine vision is not. Cameras sensitive to diverse infrared spectral bands can improve the capacities of autonomous systems and provide a comprehensive view. Relevant scene content can be made visible, particularly in situations when sensors of other modalities, such as a visual-optical camera, require a source of illumination. As a result, increasing the level of automation not only avoids human errors but also reduces machine-induced errors. Furthermore, multi-spectral sensor systems with infrared imagery as one modality are a rich source of information and can conceivably increase the robustness of many autonomous systems. Robotics, automobiles, biometrics, security, surveillance, and the military are some examples of fields that can profit from the use of infrared imagery in their respective applications. Although multimodal spectral sensors have come a long way, there are still several bottlenecks that prevent us from combining their output information and using them as comprehensive images. The primary issue with infrared imaging is the lack of potential benefits due to their cost influence on sensor resolution, which grows exponentially with greater resolution. Due to the more costly sensor technology required for their development, their resolutions are substantially lower than those of regular digital cameras.

This thesis aims to improve beyond-visible-spectrum machine vision by integrating multi-modal spectral sensors. The emphasis is on transforming the produced images to enhance their resolution to match expected human perception, bring the color representation close to human understanding of natural color, and improve machine vision application performance. This research focuses mainly on two tasks, image Colorization and Image Super resolution for both single- and cross-domain problems. We first start with an extensive review of the state of the art in both tasks, point out the shortcomings of existing approaches, and then present our solutions to address their limitations. Our solutions demonstrate that low-cost channel information (i.e., visible image) can be used to improve expensive channel information (i.e., infrared image), resulting in images of higher quality and closer to human perception at a lower cost than a high-cost infrared camera.

Key words: *cross-domain image processing, deep learning, computer vision, image restoration*

Resumen

La visión humana está restringida al espectro visual-óptico. La visión artificial no lo es. Las cámaras sensibles a diversas bandas espectrales de infrarrojos pueden mejorar las capacidades de los sistemas autónomos y proporcionar una visión integral. El contenido relevante de la escena se puede hacer visible, particularmente en situaciones en las que los sensores de otras modalidades, como una cámara visual-óptica, requieren una fuente de iluminación. Como resultado, aumentar el nivel de automatización no solo evita los errores humanos, sino que también reduce los errores inducidos por las máquinas. Además, los sistemas de sensores multiespectrales con imágenes infrarrojas como una modalidad son una rica fuente de información y posiblemente pueden aumentar la solidez de muchos sistemas autónomos. La robótica, los automóviles, la biometría, la seguridad, la vigilancia y el ejército son algunos ejemplos de campos que pueden beneficiarse del uso de imágenes infrarrojas en sus respectivas aplicaciones. Aunque los sensores espectrales multimodales han recorrido un largo camino, todavía hay varios cuellos de botella que nos impiden combinar su información de salida y utilizarlos como imágenes completas. El problema principal con las imágenes infrarrojas es la falta de beneficios potenciales debido a la influencia de su costo en la resolución del sensor, que crece exponencialmente con una mayor resolución. Debido a la tecnología de sensor más costosa requerida para su desarrollo, sus resoluciones son sustancialmente más bajas que las de las cámaras digitales normales.

Esta tesis tiene como objetivo mejorar la visión artificial más allá del espectro visible mediante la integración de sensores espectrales multimodales. El énfasis está en transformar las imágenes producidas para mejorar su resolución para que coincida con la percepción humana esperada, acercar la representación del color a la comprensión humana del color natural y mejorar el rendimiento de la aplicación de visión artificial. Esta investigación se centra principalmente en dos tareas, la colorización de imágenes y la superresolución de imágenes, tanto para problemas de un solo dominio como de dominio cruzado. Primero comenzamos con una revisión extensa del estado del arte en ambas tareas, señalamos las deficiencias de los enfoques existentes y luego presentamos nuestras soluciones para abordar sus limitaciones. Nuestras soluciones demuestran que la información de canal de bajo costo (es decir, la imagen visible) se puede utilizar para mejorar la información de canal costosa (es decir, la imagen infrarroja), lo que da como resultado imágenes de mayor calidad y más cercanas a la percepción humana a un costo menor que una imagen de alto costo cámara infrarroja.

Palabras clave: *procesamiento de imágenes entre dominios, aprendizaje profundo, visión artificial, restauración de imágenes*

Resum

La visió humana està restringida a l'espectre visual-òptic. La visió artificial no ho és. Les càmeres sensibles a diverses bandes espectrals d'infrarojos poden millorar les capacitats dels sistemes autònoms i proporcionar una visió completa. El contingut rellevant de l'escena es pot fer visible, especialment en situacions en què els sensors d'altres modalitats, com ara una càmera visual-òptica, requereixen una font d'il·luminació. Com a resultat, augmentar el nivell d'automatització no només evita errors humans sinó que també redueix els errors induïts per la màquina. A més, els sistemes de sensors multiespectrals amb imatges infrarojes com una modalitat són una font rica d'informació i poden augmentar la robustesa de molts sistemes autònoms. La robòtica, els automòbils, la biometria, la seguretat, la vigilància i l'exèrcit són alguns exemples de camps que poden beneficiar-se de l'ús d'imatges infrarojes en les seves respectives aplicacions. Tot i que els sensors espectrals multimodals han recorregut un llarg camí, encara hi ha diversos colls d'ampolla que ens impedeixen combinar la seva informació de sortida i utilitzar-los com a imatges completes. El problema principal amb la imatge infraroja és la manca de beneficis potencials a causa de la seva influència en el cost en la resolució del sensor, que creix exponencialment amb una resolució més gran. A causa de la tecnologia de sensors més costosa necessària per al seu desenvolupament, les seves resolucions són substancialment inferiors a les de les càmeres digitals normals.

Aquesta tesi té com a objectiu millorar la visió artificial de l'espectre més enllà del visible mitjançant la integració de sensors espectrals multimodals. L'èmfasi està en transformar les imatges produïdes per millorar-ne la resolució perquè coincideixi amb la percepció humana esperada, apropar la representació del color a la comprensió humana del color natural i millorar el rendiment de l'aplicació de visió artificial. Aquesta investigació se centra principalment en dues tasques, la coloració d'imatges i la superresolució d'imatges, tant per a problemes d'un sol domini com per a problemes entre dominis. Primer comencem amb una revisió extensa de l'estat de l'art en ambdues tasques, assenyalem les deficiències dels enfocaments existents i després presentem les nostres solucions per abordar les seves limitacions. Les nostres solucions demostren que la informació del canal de baix cost (és a dir, la imatge visible) es pot utilitzar per millorar la informació del canal cara (és a dir, la imatge infraroja), donant lloc a imatges de major qualitat i més properes a la percepció humana a un cost més baix que un cost elevat càmera infraroja.

Paraules clau: *processament d'imatges entre dominis, aprenentatge profund,*

visió per computador, restauració d'imatges

Contents

List of Figures

List of Tables

1 Introduction

When we describe light in common terms, most often we are talking about Visible light, which our eyes can naturally see. This, however, represents a very narrow portion of what we call the Electromagnetic Spectrum (see Fig. 1.1). The human eye is one of nature's most complicated systems. Our eyesight can distinguish between 8 million colors with around 126 million light-sensitive cells. We have a great imaging device, however, this device has certain limitations. For instance, it cannot see in certain conditions or in full darkness. What humans perceive as color is limited to a tiny window of wavelengths inside the electromagnetic spectrum. Our visual perception is dominated by reflected light with wavelengths ranging from 380 to 760 nm in the electromagnetic spectrum [165] This wavelength range is also called the visual-optical (VIS) spectrum and it is also what a basic digital camera captures. VIS sensitive cameras capture the visible light either separately in three channel Red-Green-Blue (RGB) color images or entirely in gray-value image.

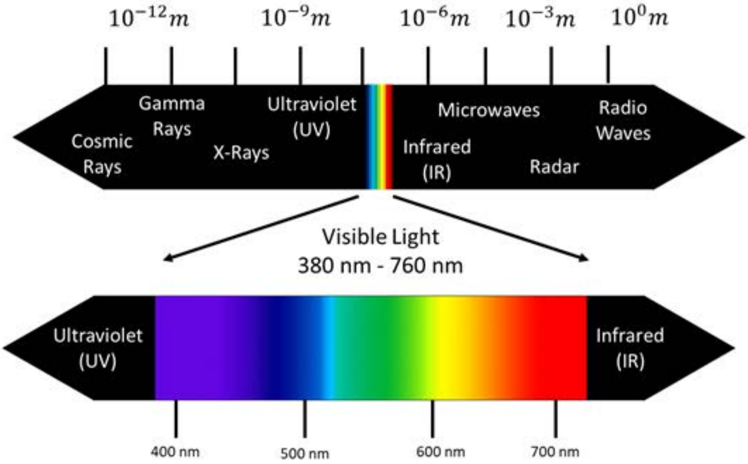


Figure 1.1: Overview of the different range in the electromagnetic spectrum.

There are situations when the visible spectrum fails to show items in the scene due to darkness or occlusion, whereas other wavelengths beyond the visible spectrum can detect and expose them. Non-visible light sensors, such as infrared, depth imaging, and other sensors can provide images where the visible spectrum fails to provide images. This is where vision beyond the visible spectrum plays a crucial role with the help of non-visible light sensors such as infrared and depth imaging sensors.

Astronomer William Herschel discovered the invisible rays that produce heat (also known as dark heat or infrared radiation) in the year 1800. This invention was initially used by British scientists in 1929, when they constructed the first infrared-sensitive electronic camera for an anti-aircraft defense system. [149]. The infrared (IR) spectral range is close to the visible range, with wavelengths from 760nm to 1mm. Other waves come after the IR range such as microwaves and RADAR/TV waves. Small part of IR spectrum belongs to thermal IR imaging, which also called thermography. Unlike visible images, thermal images don't need visible light to work, therefore they can work just fine in total darkness since objects produce heat radiation based on their temperature. Also the amount of light around the object does not matter since thermal imaging can show covered objects at different temperatures [2].

However, imaging with electromagnetic waves (EM) is limited by the characteristics of matter between the object and the imaging sensor, which can influence the imaging sensor's capacity to acquire an accurate image when the item is not obscured, such as behind a glass or a bright object, thermal imaging works well. Glass can completely obscure the thermal sensor or reflect the heat of nearby objects. The environment in which the thermal image is acquired has a significant impact on its accuracy, because the values given in the thermal image represent fluctuations in temperature in the context of the field of view. Warm items stand out well in the recorded image versus colder objects in the surrounding area. Sunlight during the warmer days can also produce substantial noise in the thermal image.

Infrared imaging is increasingly being used in development research and a range of different fields in industry, which has lately led to the manufacturing of low-cost vision sensors. Low-cost sensors are fast becoming available, and they are making their way into applications other than heavy industrial usage, such as surveillance, criminal investigation, military use, medical research, and building maintenance (see Fig. 1.2). Exploiting these alternate perspectives has the potential to play a significant role in computer vision by improving the accuracy of our existing conventional digital vision. A machine vision built on all of these modalities can see in the optical, infrared, audible spectrum, and can catch physical object details that the human eye cannot see. A massive quantity of data is generated when all of these sensor modalities are combined in a single camera. As a result, it is critical to discover ways to leverage the information from all of the sensors and combine them together to maximize advantages.

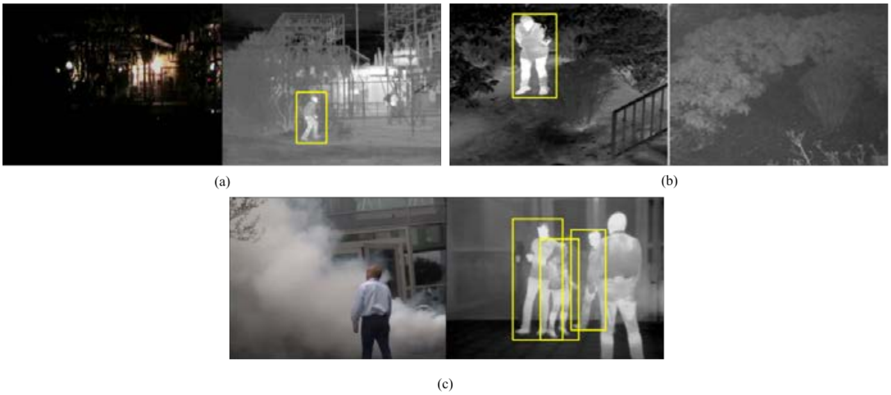


Figure 1.2: Thermal IR imaging usage in security applications from [33, 150].

Despite all of the developments, there are several challenges that prevent the fusing of their output features for use as integral images. The fundamental disadvantage of infrared imaging (SWIR, MWIR, and LWIR) is the high cost of high resolution IR sensors. In this modality sensor, the cost effect by sensor resolution, which grows exponentially with resolution. For instance, the common resolution for thermal sensors varies from 40 x 60 to 640 x 480, however greater resolutions are possible, limiting their use in particular applications where precise high-resolution images are required.

In this thesis, we aim to improve beyond visible spectrum machine vision by enhancing the quality of an expensive channel (IR image) by integrating the information of its low-cost channel (visible image) counterpart. The emphasis is on reconstructing the low-resolution IR images with better visual quality, enhancing the resolutions to match the expected human perception, and improving the performance of different computer vision tasks. Furthermore, human perception is better at understanding true colors than shades of gray or pseudo-colors since infrared images are well suited for night vision and darkness. Thus, the color enhancement is applied to transfer the color representation to the infrared image and map it as closely as possible to the human understanding of natural color.

1.1 Image Restoration and Enhancement

Recent years have witnessed of increased interest in low-level vision task from the vision and graphics communities. Image restoration and enhancement are key components of computer vision tasks, which aim to restore a degraded image content, filling in of missing information, or the needed transformation and/or manipulation to achieve a desired target with respect to perceptual quality, contents, or performance of apps working on such images. It is often used to assist individuals in analyzing video and image content or to provide visually appealing images and videos to human. It can also be used as a preprocessing approach to simplify the work and enhance the performance of later automatic image content analysis algorithms.

Image restoration has been a long-standing research topic in digital image processing since last century [130], which aim is to recover clean latent images from degraded observations. Generally, image restoration is an inverse problem, in which infinite possible mappings between multi-dimensional degraded observations and restored images, determine the ill-posed nature of such inverse problems. In cases where mappings are known and invertible, corresponding solutions are easy to obtain, but such mappings are unique and lack generality. In practice, the inverse mappings are unknown, therefore the solution space is infinite and requires regularization techniques to be applied in order to derive feasible and optimal solutions. Therefore most researches in image restoration are devoted to resorting effective analytical models and learning schemes, such that approximations of exact mappings can be found to restore degraded images[143].

Conventional methods for image restoration rely on advanced mathematics and probabilistic models to solve inverse problems, which are mostly based on the maximum likelihood or Bayesian approaches in iterative algorithms [22, 23]. In the past decade, the rapid rise of deep learning techniques has greatly impacted various computer vision tasks, from recognition and classification to regression and generation. Convolutional Neural Networks (CNNs) firstly boosted the performance of classification and detection [81], with numerous network architectures proposed to tackle benchmark research tasks. VGGNet [140] points out that deep network architecture is beneficial, whereas previous studies mostly focused on shallow networks. ResNet [55] provides the baseline structure of image restoration and becomes the basic structure of several following methods, like EDSR [95] (for super-resolution), DeepDeblur [121] (for image deblurring), DnCNN [184] (for image denoising). See Fig. 1.3.

Deep learning approaches bring many benefits to image restoration, such as learning-based methods that can boost the performance of different tasks. On most benchmark datasets, deep learning-based methods outperform traditional methods significantly. By using parallel processing units such as graphic processing units (GPUs), deep learning algorithms naturally fit with computer hardware, leading to high efficiency compared to using CPUs. Nowadays, many applications in different fields rely on image restoration and enhancement methods, such as digital display technologies, medical imaging analysis, security and surveillance, remote sensing satellite imagery, and many others.

In this work, we divide the image restoration methods into two groups based on the type of domain information (visible spectrum and infrared spectrum) in order to meet the goal of the research. Overall, a single-domain image restoration is when the original and improved image are in the same domain. On the other hand, when different domains are involved in the processing, it is called cross-domain image restoration. All the image restoration approaches discussed concern only the spatial transformation.

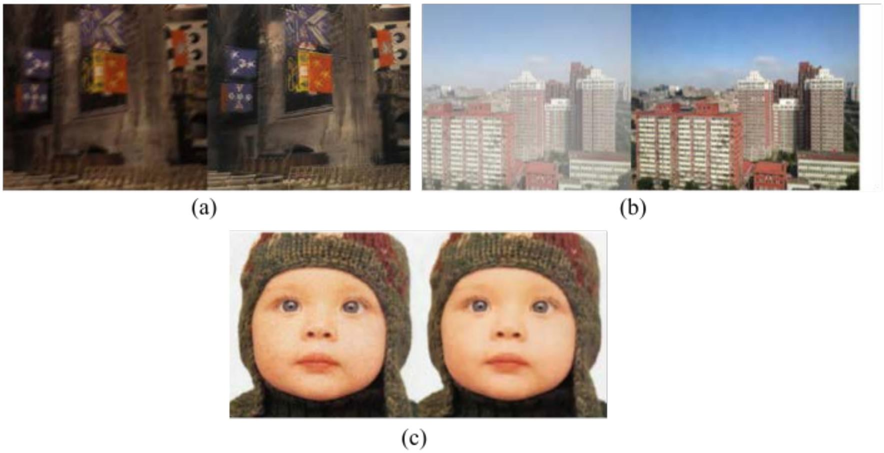
1.1.1 Single Domain Image Restoration Approaches

High image quality is essential for all tasks, whether performed by a human or a computer [34]. This image quality is characterized by displaying the observed scene's content with clear, well-defined edges, textures, and structures. Additionally, factors that degrade image quality, such as blur, noise or low contrast, should be reduced as they impact the image content analysis.

Image Denoising

During the process of acquiring, compressing, and transmitting an image, the images can unavoidably polluted by noise, which causes distortion and information loss. This noise might be an influence of the environment, the transmission channel, or other reasons. Thus, removing noise from a noisy image play an important role since noise has negative effects in analysis applications or performance of various low-level vision tasks from many aspects. Image denoising aim is to remove noise from a noisy image, therefore a clean image without noise can be restore. Deep learning based denoising can be seen mathematically as follow, where noised image Y can be expressed as $Y = X + N$ where X denotes clean image and N additive noise corrupted with X. Noise can also be multiplicative in nature. Based on the type of noise, image denoising can be divided into four categories: additive white noise image (AWNI) denoising, real noisy image denoising, blind denoising, and hybrid image denoising. Among these categories, AWNI attracts most attention [143]. However, the popularity of AWNI does not reflect real noisy images. As a result, although AWNI denosing includes Gaussian, Poisson, salt, pepper and multiplicative noise, there are still gaps with actual application scenarios.

Image Debluring



(a)

(b)

(c)

Figure 1.3: Example of image restoration tasks [143]: (a) deblurring, (b) dehazing, and (c) denoising.

Blurry images are common in practice, and restoration of these degraded images are intractable due to a wide range of factors, such as inevitable motions during long exposure time, physical limitations of imaging devices, unknown degradation process, and many others. Researchers have paid many efforts to develop efficient and novel methods to solve these challenging problems. Dynamic scene blurs are ubiquitous in real life image capturing. Blurs can be caused by a mixture of camera motion, object motion, and scene depth variation. Camera motion has six degrees of freedom in two categories, translational and rotational motions. Translational motion relates to depth variation [51, 65], while rotational camera motion and object motion are independent factors that also lead to non-uniform blurs in the image. Since these motion blurs are spatially variant, it is not a trivial task to model imaging and degradation process, especially when there is only a single blurry image that is available. Numerous attempts contribute to building models that approximate real blur kernels by using prior knowledge and additional observations on images. The deblurring process has attracted widespread attention in the field of image processing, since it is needed in many applications such as, image segmentation [129], astronomy [39], and microscopy [53].

Image Dehazing

Single image dehazing aims to estimate a haze-free image from a hazy image. It is a classical image processing problem, which has been an active research topic

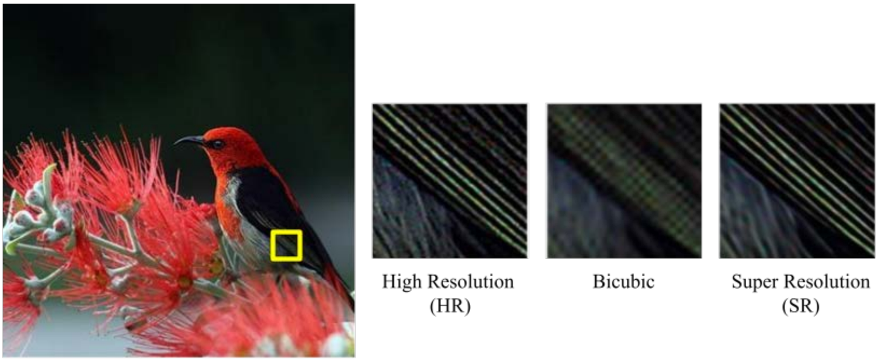


Figure 1.4: Example of a high-resolution image compared to a bicubic interpolation and super-resolution sample of the same image [95].

in the vision and graphics communities within the last decade. As numerous real-world tasks such as, traffic detection and environmental monitoring require high-quality images, and the hazy environment usually leads to deprecated images. Thus, it is of great interest to develop an effective algorithm to recover haze-free images [103].

Haze is a complex atmospheric phenomenon. Images with haze may lose color fidelity and visual contrast as a result of light scattering through the haze particles. Mathematically, the hazing process can be simplified by $I(x) = J(x)t(x) + A(1 - t(x))$, where I is the hazy image, J is the haze-free image, and t is the medium transmission map which describes the relative portion of the light that reaches the camera sensor from scene surfaces without being scattered. While A is the atmospheric light and x denotes pixel coordinate. This problem is highly ill-posed because many different pairs of A, t and J give rise to the same I. Image dehazing is essential in many real-world applications that demand a high-quality image, as well as in areas where fog and haze are common, such as archaeology, traffic detection, and satellite imaging,.

Single Image Super Resolution

The image-based computer graphics models lack resolution independence [40] as the images cannot be zoomed beyond the image sample resolution without compromising the quality of images. Thus, simple image interpolation will lead to blurring of features and edges within a sample image. Also, in some imaging systems, a high-quality sensor is too expensive to utilize or not feasible such as, remote sensing satellite imagery or thermal IR imaging.

The concept of super-resolution was first used by [42] to improve the resolution of an optical system beyond the diffraction limit. In the past two decades, the concept of super-resolution (SR) is defined as the method of producing high-resolution (HR) images from a corresponding low-resolution (LR) image. The applications of super-resolution include computer graphics [78], medical imaging [13], security, and surveillance [48], which shows the importance of this topic in recent years. Other than improving image perceptual quality, it also helps to improve other computer vision tasks such as detection, recognition, segmentation, and many other vision tasks.

The image super-resolution, although being explored for decades, remains a challenging task in computer vision, and this problem is fundamentally ill-posed because, for any given LR image, there can be several HR images with slight variations in camera angle, color, brightness, and other variables. Furthermore, there are fundamental uncertainties among the LR and HR data since the downsampling of different HR images may lead to a similar LR image, which makes this conversion a many-to-one process [172].

In the past, classical SR methods such as statistical methods, prediction-based methods, patch-based methods, edge-based, and sparse representation methods were used to achieve super-resolution. However, recently the advances in computational power and big data have made researchers use deep learning to address the problem of SR. In the past decade, deep learning-based SR studies have reported superior performance than the classical methods, and DL methods have been used frequently to achieve SR image. A range of methods has been used by researchers to explore SR, ranging from the first method of Convolutional Neural Network (CNN) [35] to Generative Adversarial Nets (GAN) [86] and recently Vision Transformer (ViT) [93] (see Fig. 1.4). This problem is one of the objective of this research and will be discussed in details in Chapter 2.

1.1.2 Cross Domain Image Restoration Approaches

Humans are capable of adapting between different domains and integrating the transfer of knowledge from one domain to another one easily due to the fact that, they can relate previously acquired knowledge in different domains and infer the unknown knowledge of the unknown domain based on the prior information of the other learned domain. Humans are also excellent at cross-domain inference because of their extensive previous knowledge of both domains, which they have acquired via the information they have gathered throughout their everyday lives. In addition, human self-learning and error correction, such as predicting an oncoming sound are daily practices that expand our knowledge. However, with machines, transitioning across domains becomes extremely challenging. For a machine to learn cross-domain transformation, pairs images of both domains must be available.

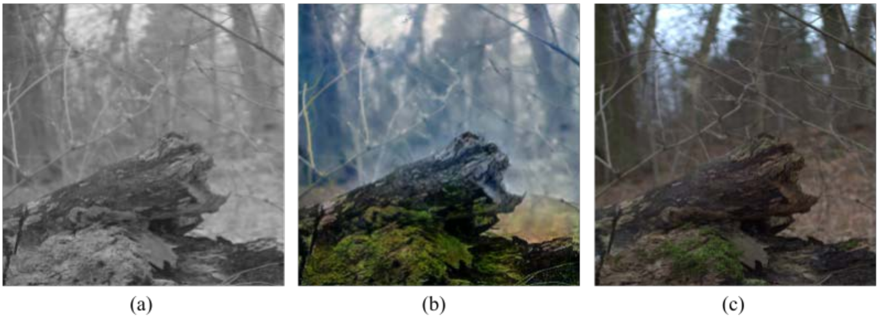


Figure 1.5: Images from the CycleGAN based approach presented in [116] NIR input image (a), result from [179] (b), and ground truth RGB image (c).

However, there are circumstances in which paired images cannot be obtained or it is a costly and time-consuming operation.

In cross-domain image restoration, the model uses the images from two domains as input to improve the image of one domain, such as in multispectral satellite imaging, image super resolution, and multimodal image fusion. In this problem, image restoration tasks require the existence of a roughly aligned pair of images. However, multimodal sensors must have identical geometrical registration, and in the majority of situations, their fields of view are distinct and their alignment is poor. It is possible to build a cross-domain model that learns the joint distribution of data from the marginal distributions of each of its specific domains without the need for paired images. However, this has its own limitation because training image sets have limited variation and cannot generalize as in the human scenario by self-learning and error correction. The inference process in humans uses a multidimensional input of the environments, whereas it uses a reduction of reality in machines. Therefore, samples in the two different domains may not fully overlap because of the different nature of the information or because of outliers, which makes it difficult to infer the target domain given only the source domain.

Image Colorization

The color distribution of the image has a significant impact on the appearance of the image; therefore, color enhancement methods can play an important role in improving the visual quality of images. Image colorization is an essential image processing task that has been largely studied in recent years in the context of the VIS spectrum to automatically colorize black and white photos or classic movies, for instance [72]. It is a very challenging problem, as it is severely ill-posed since two out of the three image channels are lost. Furthermore, changes in illumination, viewpoint variations, shadows in the scene, and occlusions all have an impact on the colorization problem [149].

A night vision application is designed to give humans the ability to see in low-light conditions or total darkness. IR images are shown in shades of gray or in pseudo-colored images, but human vision is superior at identifying and understanding actual colors. Therefore, image colorization is essential for enhancing nighttime or dark scene visibility and understanding. Coarsely speaking, colorization techniques can be classified into parametric and non-parametric approaches. Parametric methods learn prediction functions from large datasets of color images at training time, posing the problem as either regression onto continuous color space or classification of quantified color values. On the other hand, non-parametric methods utilize an input gray-scale image and firstly define one or more color reference images (provided by a user or automatically retrieved) as source data. Then, following the image analogy framework, color is transferred onto the input image from analogous regions of the reference image(s). The aforementioned classification is based on visible spectrum image colorization approaches.

IR image colorization somehow shares common properties and problems with these monochromatic image colorization approaches. There are different motivations to colorize IR images, like if an operator has to analyze and evaluate the scene content of a NIR image, a visual representation he is familiar with may help him to fulfill his task. Also, IR image colorization can improve the detection and recognition tasks when these algorithms are failing to work with IR images. This problem is part of this research and is discussed in Chapter 3.

Guidance Image Super Resolution

The use of infrared images has grown over the last two decades as the cost and availability of infrared cameras has decreased. However, in spite of the continuous increase in the usage of IR cameras, there is still a limitation on image resolution. This limitation is imposed by the technology needed for these cameras. For example, there are some high-resolution thermal cameras on the market, but they are generally based on a more expensive technology called actively cooled thermal cameras; hence, most of the applications are based on utilizing uncooled thermal cameras, which are available on the market at a significantly lower price.

As previously stated, various image super resolution techniques have been presented over the last few decades. Although most of them are intended for the VIS spectrum, in recent years some adaptations or novel approaches have been proposed for the IR image domain. In spite of these contributions, the difference between the resolution of VIS spectrum and IR images, in particular thermal IR images, is still considerable due to the nature and the market of the sensors. This significant difference in resolution prompted researchers working on cross-spectral computer vision to devise strategies for using high resolution VIS spectrum images to generate super resolution thermal IR images at a lower cost.

Thermal IR and visible images can both show the same scene, but the texture of the images will be different because of the differences in how light reflection and temperature are captured. When there is enough light, visible images have more information than their thermal counterparts [2]. However, objects may be missed in areas with less light. Therefore, cross-domain image integration can improve the performance of the system by using the finer details captured in the visible image domain to improve the resolution of the thermal IR image. This is especially true when the nature of the problem calls for integration and when the environment is not ideal for a one-sensor approach. These contributions can be considered as a guidance approaches. This problem is discussed in detail in Chapter 7.

1.2 Thesis Outline and Contributions

This thesis is divided into three main parts. In first part, we present the problem of image Colorization and our contributions to this task. In the second part, we present the Single image super-resolution and our propose new solutions. In the last part, we present the problem of Guidance Image Super-Resolution and our propose approach. Each chapter corresponds to an article either published or submitted in a journal or conference.

- **Chapter 1: Introduction.**
- **Chapter 2: Background.** In this chapter a general overview and problem definitions are being provided, as well as a review of the mainstream datasets and evaluation metrics which have been used for performance comparison by the community. Following with a literature review including state-of-the-art models in relation to the proposed methods of each main part.
- **Chapter 3: Colorizing Near Infrared Images Through a Cyclic Adversarial Approach of Unpaired Samples**.

  Objectives: The objective of this chapter is to do a thorough examination of the image colorization problems by introducing a new solution which can colorize near-inferred images when no paired dataset is available to increase the applicability of near-inferred images.

  Contributions: The focus lays on colorizing the near-infrared images when no paired dataset is available by using a Generative Adversarial Network. This will be done by proposing a new CycleGAN variant with a completely redesigned the generator and discriminator of the original CycleGAN, which allows for producing of more realistic colors for infrared images, as well as better visual quality compared to other network.

Publication: Mehri, Armin, and Angel D. Sappa. "Colorizing near infrared images through a cyclic adversarial approach of unpaired samples." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2019.

- **Chapter 4: MPRNet: Multi-Path Residual Network for Lightweight Image Super Resolution**.

  Objectives: The main objective of this chapter is to propose a novel lightweight single image super resolution model for visible images (RGB images), which is suitable for real-world applications, and to determine if the already existing models can be applied to super-resolution thermal image problems as a novel experiment and test their applicability to this domain.

  Contributions: An efficient and fast CNN-based network introduced, named MPRNet, to solve the SISR problem. MPRNet designed by proposing an effective Adaptive Residual Block, which focuses on spatial information through the use of multi-path residual learning connections in order to improve performance with almost no increase in the amount of required computing. The results of a comprehensive study demonstrate that MPRNet performs exceptionally well.

  Publication: Mehri, Armin, Parichehr B. Ardakani, and Angel D. Sappa. "MPRNet: Multi-path residual network for lightweight image super resolution." Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2021.

- **Chapter 5: Thermal Image Super Resolution Challenges**.

  Objectives: After demonstrating that the previous chapters' contributions have surpassed the SOTA methods of solving the Single Image Super Resolution problem, we began conducting experiments to determine the most effective model and training procedures to move forward with the Thermal Image Super Resolution problem.

  Contributions: We presented a couple of models in international challenges for evaluation, and we got good ranks. We got the second position in PBVS 2020 and also attended the PBVS 2021 thermal SISR challenge.

  Publications: Rivadeneira, Rafael E., Angel D. Sappa, Boris X. Vintimilla, Sabari Nathan, Priya Kansal, Armin Mehri, Parichehr Behjati Ardakani et al. "Thermal image super-resolution challenge-PBVS 2021." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.

Rivadeneira, R. E., A. D. Sappa, B. X. Vintimilla, L. Guo, J. Hou, A. Mehri, P. B. Ardakani et al. "Thermal image superresolution challenge-PBVS 2020. In 2020 IEEE." In CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 432-439. 2020.

- **Chapter 6: SRFormer: Efficient Yet Powerful Transformer Network For Single Image Super Resolution**.

  Objectives: Transformer-based networks were introduced in many vision and NLP tasks and have shown significant performance gains when compared to CNN-based networks, but these models suffer from slower training and inference time. The aim is to propose an efficient Transformer-based network to overcome the aforementioned problems for SISR, which can also be used later for solving the cross-domain Image Super Resolution.

  Contributions: In this chapter SRFormer is being introduced, an efficient yet powerful Transformer-based SISR network that is able to generate SR images faster while requires less training time than other SOTA. To do so, a lightweight self-attention layer introduced, named Dual Attention layer (DAL). DAL generates a global attention map from two local attention weights in parallel while remaining memory efficient. Extensive experiments show that SRFormer obtains SOTA in various benchmark datasets.

  Publication: Armin Mehri, Parichehr Behjati, and Angel D.Sappa. SRFormer: Efficient super-resolution transformer-based network for single image super resolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. (*under review*)

- **Chapter 7: TnTViT: Transformer in Transformer Network for Guidance Super Resolution**.

  Objectives: The main idea is to investigate how rich texture details in visible images contribute to improve and enhance the problem of the guidance super resolution problem (cross-domain image super resolution).

  Contributions: TnTViT a novel and lightweight Transformer network based on our previously SISR model (i.e., SRFormer) proposed. The proposed model accepts two images from different domains as inputs and aggregates the features of the low-cost channel (visible image) with the corresponding features of the expensive channel (infrared image) to enhance the resolution of the infrared image to be as close as possible to human preference. Extensive experiments illustrate that the proposed method can boost the quality of infrared images by a large margin.

Publication: Armin Mehri, Parichehr Behjati, and Angel D. Sappa. TnTViT: Transformer in Transformer Network for Guidance Super Resolution. *IEEE Access*, 2022. (*under review*)

- **Chapter 8: Conclusion.** The last chapter concludes the work developed in this thesis and proposes further directions for research in cross-domain image processing problems.

2 Background

This thesis mainly focuses on addressing and solving image restoration tasks in both single- and cross-domain by proposing novel deep learning approaches. In this chapter, we first detail the problem definition of Image Colorization and Image Super Resolution. Next, we introduce some related works, including benchmark datasets, assessment methods, and optimization objectives that have been used to accomplish this research. Finally, we provide a literature review, including state-of-the-art methods related to the proposed methods.

2.1 Image Colorization

This section provides information related to Image Colorization problem.

2.1.1 Problem Definition

Image Colorization is the process of assigning an RGB color value to each pixel of a grayscale image to obtain colorized images, which is a prospective image processing technique in computer vision [67]. Image colorization is a multimodal problem in which the same target object can have different colorization schemes. For example, a pair of shoes can be white, red, yellow, or another color. In general, image colorization is a challenging and interesting research problem.

In practice, it is difficult to obtain a large number of gray image datasets to train a colorization model, so the gray image I_g is usually modeled as the output of the following equation:

$$I_g = \Phi(I_r), \tag{2.1}$$

where I_r represents the color image. Given an input grayscale image I_g with a size of $W \times H$, the input gray image I_g is mapped into a color image I_c through the image colorization model f. The equation is as follows:

$$I_c = f\left(I_g\right). \tag{2.2}$$

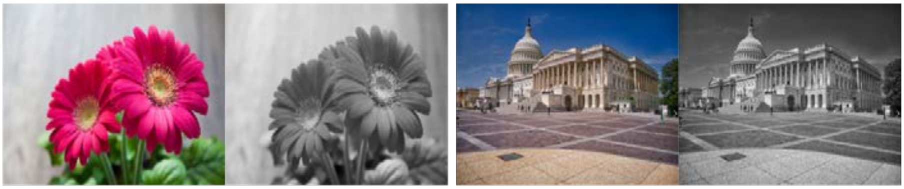


Figure 2.1: Example of grayscale Image Colorization.

For deep learning based methods, the model F is usually obtained by learning a collection of training samples. i.e., given a grayscale image collection $G = \{I_g \in \mathbb{R}^{W \times H \times 1}\}$ and corresponding real color image collection $C = \{I_c \in \mathbb{R}^{W \times H \times 3}\}$, find a model F which can minimize prediction errors L.

$$\hat{\theta} = \arg\min_{\theta} L(I_c - I_r) + \lambda \Psi(\theta). \tag{2.3}$$

here is usually certain distance measurement (such as L1 distance, L2 distance) or a combination of various distance measurement. F is the set of potential mapping functions; $\Psi(\theta)$ is the regularization term, and λ is the compromise parameter.

2.1.2 Grayscale Image Colorization

Traditional colorization methods were labor-intensive and restricted to small areas since they relied heavily on human input. The digital revolution, however, shifted the colorization process into a computer vision problem, speeding up its growth. Even though colorization research includes a wide range of topics and methods, we can typically categorize them according to the level of human intervention involved. Thus, the first category is user-guided approaches, which require user interaction, and the second category is automated data-driven methods, for which no human interaction is needed.

The second category is based on the evolution of deep learning techniques and their capacity to learn from a huge number of images. They are trained on a large collection of reference images that include images of all sorts of things. By learning the mapping function between the features of pixels in the monochromatic image and the color values of the target image, the models automatically discover the colors that naturally match real-world objects. This category eliminates the requirement for user participation during the colorization process.

Color visible spectral images are composed of luminance and chrominance components, while grayscale images are assumed to only have the luminance component. Therefore, in grayscale image colorization, the goal is to restore the

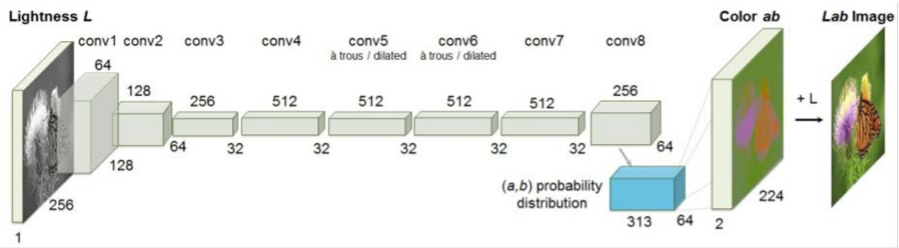


Figure 2.2: Overall network architecture of Colorful Image Colorization [187].

chrominance component in the original color image. The RGB color space is a standard linear model and is formulated by:

$$\mathbf{I}_{\text{gray}} = 0.2989 * \mathbf{I}_r + 0.5870 * \mathbf{I}_g + 0.1140 * \mathbf{I}_b. \tag{2.4}$$

Therefore, solving above equation is an inverse and ill-posed problem since two of the three image dimensions are lost. Resolving this inverse problem is a challenging problem because of the many different image conditions that need to be dealt with through a single solution.

The most simple attempt to solve the grayscale image colorization problem is a straightforward convolutional neural network with stacked layers based on a VGG network [140], such as the proposed method by Zhang et. al. in [187] (illustrated in Fig. 2.2). The model consists of multiple convolutional blocks and takes the lightness channel (L) in the CIELAB color space of the Lab transform of the image and predicts color channels (A and B). The possible color for each pixel is determined by the probabilities of belonging to one of 313 segments of the discretized and quantized ab-plane.

PSNR and SSIM are the evaluation metrics most commonly used in the colorization problem, although they do not correlate with human subjective judgment and interpretation of image quality. These evaluation metrics are described in Subsection 2.4.

2.1.3 Infrared Image Colorization

Night vision in humans is excellent yet limited, especially in inappropriate environments. Humans vision have poor vision in low light environments or no vision in a full darkness but it has the ability to see a wide range of colors perfectly when the illumination is enough. This is a biological limitation, and because of this, it has become more essential to improve night vision artificially to help the human vision

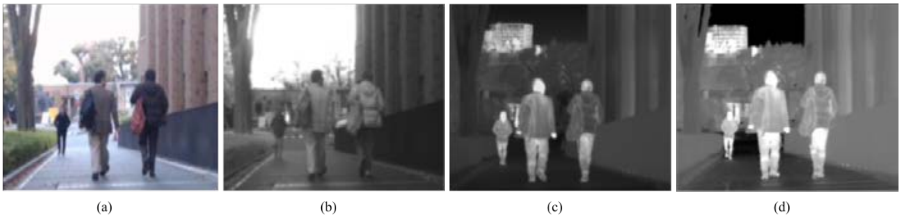


Figure 2.3: Example images showing the same scene acquired by a VIS (a), NIR (b), MWIR (c), and LWIR (d) camera [148].

system in a variety of fields, including military missions, drug studies, driving at night, and security systems.

Transforming a grayscale near-infrared (NIR) image into a multichannel RGB image is closely related to Image Colorization, where regular grayscale images are colorized, and Color Transfer, where color distributions are transferred from one RGB image to another. Both techniques, however, are not simply applicable for colorizing NIR images. They often contain multiple cues, including various optimization, feature extraction, and segmentation algorithms, and have certain prerequisites. For instance, it leverages the fact that the luminance is given by the grayscale input, and therefore the algorithms only estimate the chrominance in grayscale colorization. However, NIR colorization requires estimating both the luminance and the chrominance. On the other hand, color transfer methods are often tailored to transform multi-channel input into multi-channel output. The reduced dimensionality of single-channel NIR images renders many color transfer methods ineffective because they often require inter-color distinction to produce reasonable results.

However, it needs to be noted that NIR images and VIS images overlap in the red band in the EM spectrum; therefore, they are similar and preserve almost similar edges. For the same reasons, they tend not to work in complete darkness as they are dependent on reflected light conditions. However, they benefit from the super penetration of infrared radiation, which can overcome some visual obstacles such as clouds and fog to obtain more information (such as pedestrians, animals, road, and roadside information).

Researchers have studied how to colorize NIR images and found that they work well because there is a high correlation between NIR and RGB images. For example, Zhang et al. [190] look at the problem of recognizing faces in images taken with NIR sensors in a dark environment. They showed that the progress made in face recognition for images in the visible spectrum could not be used exactly for images in the infrared spectrum because it could not get the same results. So, the authors suggested that the NIR images be turned into color images while keeping the face identity information so that the models could recognize faces.

Moreover, the use of thermal infrared cameras has grown significantly in many areas. This is because their long wavelength lets them capture the invisible heat radiation that objects emit or reflect, no matter how bright it is. They can work around some obstacles and light changes, and they can even see things in total darkness. Thermal images are shown in shades of gray or in images that look like they have colors. But it is hard for people to understand thermal infrared images with their eyes, and they are better at seeing and understanding true colors. So, turning thermal infrared images into images in the visible spectrum is very important for making the scene easier to see and understand, especially at night.

2.2 Image Super Resolution

This section discusses the problem of Image Super Resolution by giving the problem definition and SISR frameworks in detail.

2.2.1 Problem Definition

The term "single image super resolution" refers to the process of reconstructing a high-resolution image from its lower-resolution counterpart. Let consider a Low-Resolution (LR) image is denoted by **y** and the corresponding high resolution (HR) image is denoted by **x**, then the degradation process is given as:

$$\mathbf{y} = \Phi\left(\mathbf{x};\theta_\eta\right) \tag{2.5}$$

where Φ is the degradation function, and θ_η denotes the degradation parameters (such as the scaling factor, noise, etc.). In a real-world scenario, only **y** is available while no information about the degradation process or the degradation parameters θ_η. Super-resolution seeks to nullify the degradation effect and recovers an approximation $\hat{\mathbf{x}}$ of the ground-truth image **x** as,

$$\hat{\mathbf{x}} = \Phi^{-1}\left(\mathbf{y},\theta_\varsigma\right), \tag{2.6}$$

where θ_ς are the parameters for the function Φ^{-1}. The degradation process is unknown and can be quite complex. It can be affected by several factors, such as noise (sensor and speckle), compression, blur (defocus and motion), and other artifacts. Therefore, research works prefer the following degradation model over

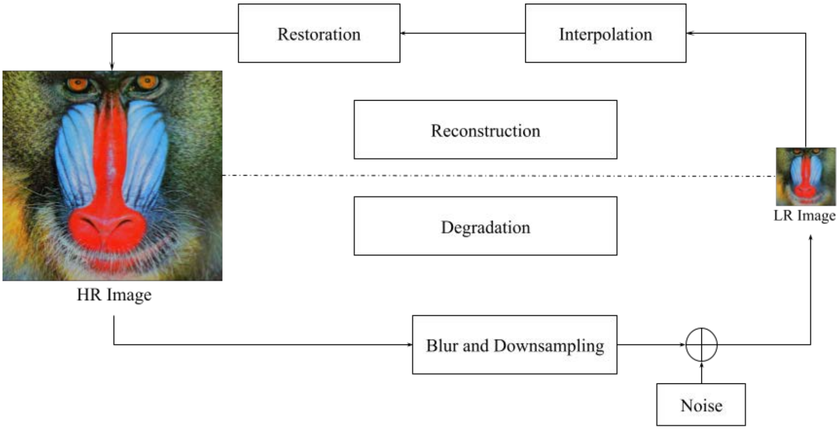


Figure 2.4: Image Super Resolution degradation and reconstruction model.

that of Equation 2.5.

$$\mathbf{y} = (\mathbf{x} \otimes \mathbf{k}) \downarrow_s + \mathbf{n}, \tag{2.7}$$

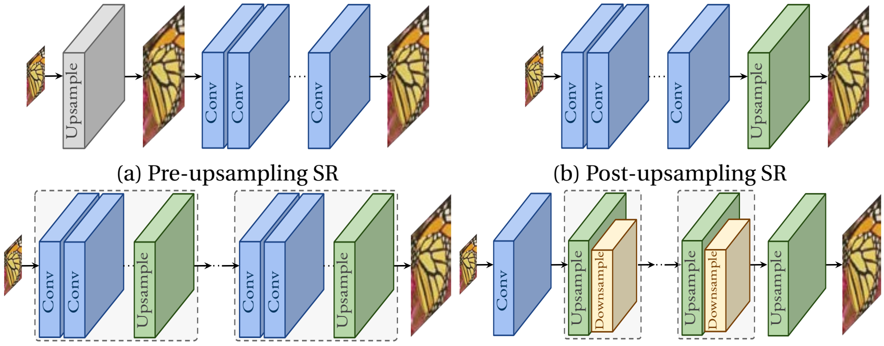
where $\mathbf{k}$ is the blurring kernel and $\mathbf{x} \otimes \mathbf{k}$ is the convolution operation between the HR image and the blur kernel, $\downarrow$ is a downsampling operation with a scaling factor s. The variable $\mathbf{n}$ denotes the additive white Gaussian noise (AWGN) with a standard deviation of σ (noise level) [5]. Figure 2.4 shows the reconstruction model of image super resolution. To this end, the objective of SR is as follows:

$$\hat{\theta} = \text{argmin}_\theta \mathscr{L}(\hat{x}_y, x_y) + \lambda \Phi(\theta), \tag{2.8}$$

where $\mathscr{L}(\hat{x}_y, x_y)$ represents the loss function between the generated HR image $\hat{x}_y$ and the ground truth image x_y, $\Phi(\theta)$ is the regularization term and λ is the trade-off parameter.

2.2.2 Super Resolution Frameworks

Due to the ill-posed nature of the SISR problem, the most important challenge is figuring out how to actually upsample the LR images in a way that the generated SR images be more sharp with fine details while minimize the artifacts. Although current models' architectures designs are widely vary, they can be classified into four model frameworks (shown in Figure 2.5) depending on the upsampling techniques and where it located in the network.



(a) Pre-upsampling SR

(b) Post-upsampling SR

(c) Progressive upsampling SR

(d) Iterative up-and-down Sampling SR

Figure 2.5: Super-resolution model frameworks based on deep learning [163].

Pre-Upsampling Super Resolution

The first category is pre-upsmapling SR, which directly learns the mapping from low-dimensional space to high-dimensional space, utilizing traditional upsampling algorithms to obtain higher-resolution images and then refining them using deep neural networks. This category is a straightforward solution, and the first work in this direction was introduced by Dong et al. [36], called SRCNN. SRCNN adopts the pre-upsampling SR framework (as Figure 2.5(a) shows) to learn an end-to-end mapping from interpolated LR images to HR images. It has gradually become one of the most popular frameworks and several approaches proposed such as [79, 139, 146, 147]. But the predefined upsampling strategy is not efficient in terms of computing because most operations are performed in a high-dimensional space.

Post-Upsampling Super Resolution

To overcome the pre-upsampling drawback, researchers suggested replacing the predefined upsampling with end-to-end learnable layers incorporated at the end of the algorithm in order to enhance computing efficiency and make full use of deep learning technologies to boost resolution automatically. Post-upsampling (see Figure 2.5(b)) is done by feeding low-resolution (LR) input into deep convolutional neural networks (CNNs) and then adding end-to-end learnable upsampling layers at the end of the network. Post-upsampling has the advantage of extracting features at a lower computational cost compared to pre-upsampling frameworks since feature extraction will happen in low-dimensional space. As a result, this framework has also become one of the most widely used [86, 95, 154]. These models differ mainly in the learnable upsampling layers, CNN structures, learning strategies, etc.

Progressive Upsampling Super Resolution

The post-upsampling SR design has allowed for significant gains in computational efficiency, although it still has several drawbacks. Training becomes significantly more difficult for larger scaling factors since there is only one stage of upsampling. Also, it needs a separate training for each scale factor. To overcome these limitations, the Laplacian pyramid SR network has been introduced to adopt a progressive upsampling structure [84](see Figure 2.5(c)).

Models built under this framework are able to deal with the multi-scale SR without suffering unnecessary spatial or temporal cost by breaking down the task into smaller pieces, which turns this strategy significantly reduces the learning difficulty, especially for deep heavy models. However, there are still issues with the models that operate within this framework. These include issues with the stability of the training and the complexity of building models for numerous phases.

Up-and-Down Sampling Super-Resolution

More recently, SISR uses an effective iterative process called back-projection [69] to better represent the interdependence of LR-HR image pairings. Iterative up-and-down sampling SR is an SR framework (depicted in Figure 2.5(d)) that repeatedly applies back-projection refinement (i.e., calculating the reconstruction error and fusing it back, to fine-tune the HR image's intensity). Several works have been proposed by using this framework such as Haris et al. [54], Li et al. [92]. This mechanism has been introduced recently into deep learning-based SR, and the framework needs further exploration.

2.2.3 Upsampling Methods

The method used to perform upsampling is just as crucial as the upsampling places in the model. While there are several traditional approaches to upsampling, CNNs are quickly becoming the preferred method for learning end-to-end upsampling. In this section, we will discuss about various classic interpolation methods and deep learning-based upsampling layers.

Interpolation-based Upsampling

Image interpolation, also known as image scaling, is commonly used by applications that work with digital images to change their size. Nearest-neighbor interpolation, bilinear interpolation, and bicubic interpolation are examples of the standard interpolation techniques. Some of these techniques are still often utilized in CNN-based SR models due to their interpretability and ease of implementation.

- **Nearest-neighbor interpolation**: The nearest-neighbor interpolation is a simple and intuitive algorithm. It selects the value of the nearest pixel for each position to be interpolated regardless of any other pixels.
- **Bilinear interpolation**: Bilinear interpolation conducts linear interpolation on one axis of the image before performing on the other one. It outperforms nearest-neighbor interpolation while remaining reasonably quick since it produces quadratic interpolation with a receptive field size of 2×2.
- **Bicubic interpolation**: Likewise, bicubic interpolation does cubic interpolation on both axes. When compared to bilinear interpolation, bicubic interpolation considers 4×4 pixels, resulting in smoother outputs with less artifacts but significantly slower performance. In reality, bicubic interpolation with anti-aliasing is the most often used approach for creating SR datasets.

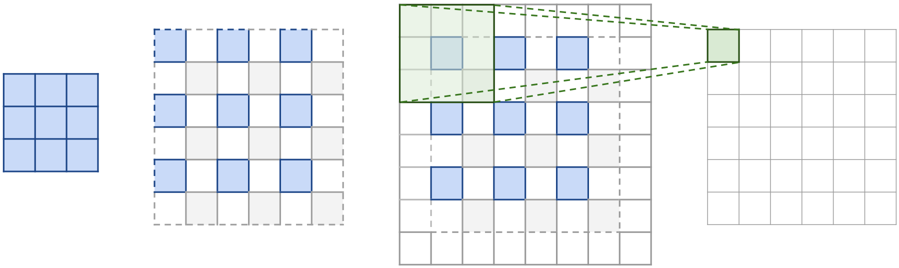
In reality, interpolation-based upsampling methods increase image resolution only by using their own image signals, with no additional data. Instead, they frequently produce unwanted side effects, like noise amplification, blurred outcomes, and computational complexity. The core concern is to replace interpolation-based approaches with learnable upsampling layers to produce sharper and well-detailed images.

Learning-based Upsampling

To address the drawbacks of interpolation-based approaches, Transposed convolution layer and sub-pixel layer are introduced to learn upsampling in an end-to-end manner.

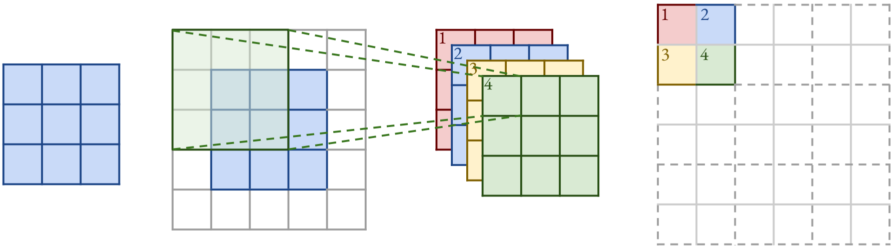
Transposed Convolutional Layer

Transposed convolution layer, a.k.a. deconvolution layer [181, 182], tries to perform transformation opposite a normal convolution, *i.e.,* predicting the possible input based on feature maps sized like convolution output. Specifically, it increases the image resolution by expanding the image by inserting zeros and performing convolution. Taking $\times 2$ SR with 3×3 kernel as an example as depicted in Figure 2.6, the input is firstly expanded twice the original size, where the added pixel values are set to 0 (Figure 2.6(b)). Then a convolution with kernel sized 3×3, stride 1, and padding 1 is applied (Figure 2.6(c)). In this way, the input is upsampled by a factor of 2, in which case the receptive field is at most 2×2. Since the transposed



(a) Starting (b) Expanding (c) Convolution

Figure 2.6: Transposed convolution layer. The blue boxes denote the input, and the green boxes indicate the kernel and the convolution output [163].



(a) Starting (b) Convolution (c) Reshaping

Figure 2.7: Sub-pixel layer. The blue boxes denote the input, and the boxes with other colors indicate different convolution operations and different output feature maps [163].

convolution enlarges the image size in an end-to-end manner while maintaining a connectivity pattern compatible with vanilla convolution, it is widely used as an upsampling layer in SR models [54, 112, 154]. However, this layer can easily cause *uneven overlapping* on each axis [123], and the multiplied results on both axes further create a checkerboard-like pattern of varying magnitudes and thus hurt the SISR performance.

Sub-Pixel Layer

The sub-pixel layer [137], another end-to-end learnable upsampling layer, performs upsampling by generating a plurality of channels by convolution and then reshaping them, as depicted in Figure 2.7. Within this layer, a convolution is firstly applied for producing outputs with s^2 times channels, where s is the scaling factor (Figure 2.7(b)). Assuming the input size is $h \times w \times c$, the output size will be $h \times w \times s^2c$. After that, the reshaping operation (a.k.a. shuffle [137]) is performed to produce outputs with size $sh \times sw \times c$ (Figure 2.7(c)). In this case, the receptive field can be up to 3×3. Due to the end-to-end upsampling manner, this layer is also widely used by SR models [1, 14, 86]. Compared with transposed convolution layer, the sub-pixel layer has a larger receptive field, which provides more contextual information to help generate more realistic details. However, since the distribution of the receptive fields is uneven and blocky regions actually share the same receptive field, it may result in some artifacts near the boundaries of different blocks. On the other hand, independently predicting adjacent pixels in a blocky region may cause unsmooth outputs.

In summary, these learning-based layers have become the most widespread upsampling techniques. In particular, in the post-upsampling framework, these layers are typically employed in the final upsampling stage for reconstructing HR images based on high-level representations extracted in low-dimensional space, thereby achieving end-to-end SR while avoiding overpowering operations in high-dimensional space.

2.3 Benchmark Datasets

Data is always necessary for data-driven models to achieve promising performance, particularly, for deep learning-based models. Recently, industry and academics have released a variety of datasets for various computer vision tasks. In this section, we are going to list the widely used datasets, which have been used in this thesis.

2.3.1 Image Colorization

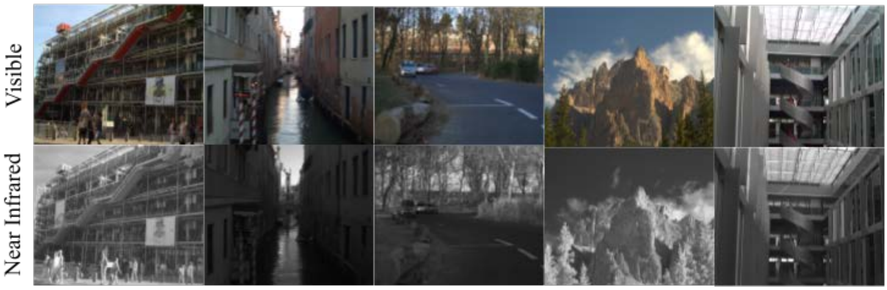


Figure 2.8: Representative test images from RGB-NIR Scene dataset [18].

RGB-NIR Scene [18] dataset consists of 477 paired images in 9 categories captured in RGB and Near-infrared (NIR). The images were captured using separate exposures from modified SLR cameras, using visible and NIR filters. This dataset consists of different scene categories such as: country, field, forest, indoor, mountain, old building, street, urban, and water. We have used RGB-NIR Scene dataset to evaluate our proposed colorization algorithms.

The RGB-NIR Scene dataset images were captured using Nikon D90 and Canon T1i cameras, using B+W 486 (visible) and 093 (NIR) filters. The cutoff between the two filters is approximately 750nm. After capture, the images were processed using dcraw. The colour capture was white balanced (dcraw -a). The NIR capture was processed using equal weights per band (dcraw -r 1 1 1 1), followed by averaging of the channels. The images were registered by extracting SIFT features at approximately 1500×2000 resolution and using RANSAC to find a similarity transform. The final transformation was recomputed via least squares from the inliers and used to resample both images in a common coordinate frame .

2.3.2 Single Image Super Resolution

There are several datasets available for image SR that differ significantly in image quantity, resolution, quality, and diversity. The widely used dataset by community for model training is DIV2K [153], which includes 800, 100, 100 training, validation, and test images respectively. Also, there are various benchmark datasets that can be used to effectively evaluate the performance of the SR methods. The representative image from all the datasets is shown in Figure 2.9.

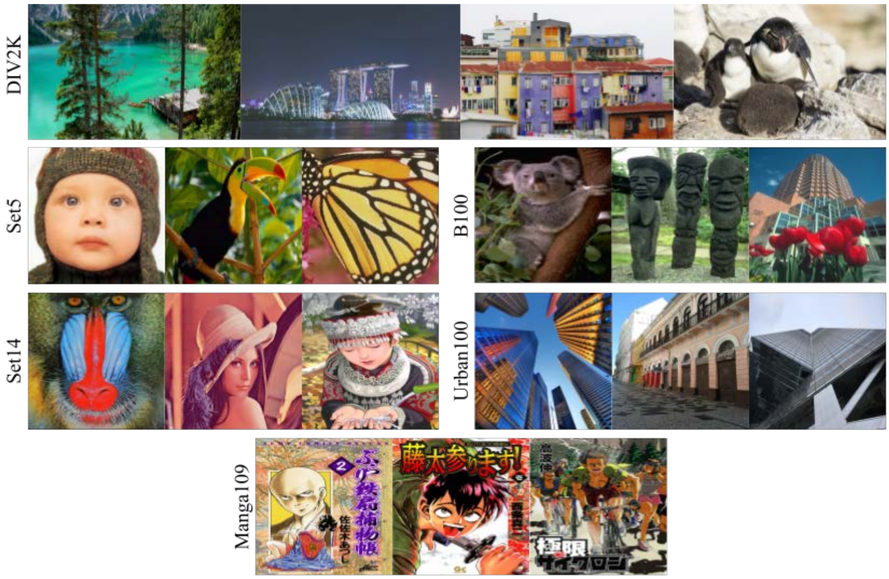


Figure 2.9: Representative test images from six most widely used super-resolution datasets used for comparing and evaluating algorithms.

- **Set5** [15] is a one of the first bechmark datasets that has been for performance evaluation of SR models. Set5 contains only 5 images, a baby, butterfly, bird, head, and a woman.

- **Set14** [183] consists of more categories as compared to Set5. However, the number of images are still low *i.e.,* 14 test images.

- **B100** [6] is another widely used benchmark datasets, which consists of 100 images of different scene, such as people, plants, food, objects and many others.

- **Urban100** [66] currently is the most challenging benchmark dataset, which includes 100 images of human-made structures *i.e.,* urban scenes.

- **Manga109** [115] is another benchmark datasets to verifying the performance of SR methods. This dataset contains of 109 drawn images by professional Japanese artists.

2.3.3 Guidance Image Super Resolution

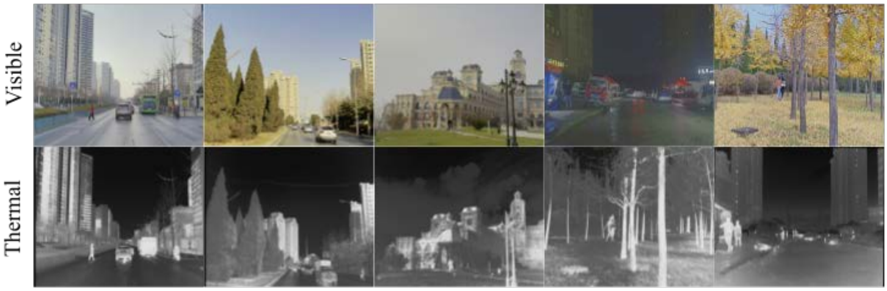


Figure 2.10: Representative some images from M3FD dataset [101].

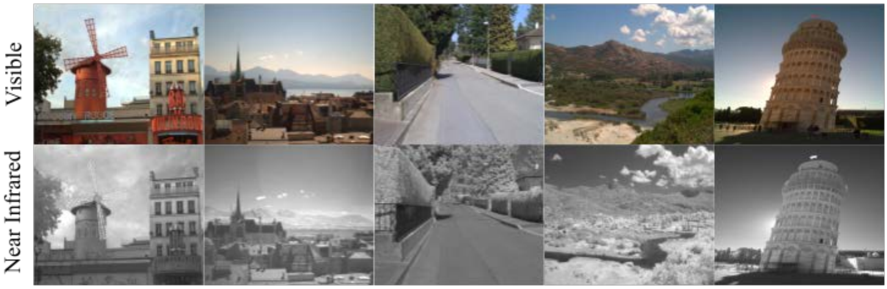


Figure 2.11: Representative some images from RGB-NIR Scene dataset [18].

The M3FD is a newly fusion dataset, which released by [101]. The M3FD dataset contains pair of visible and infrared images with resolution of 1024 × 768 and 640 × 512 respectively. The dataset built with a synchronized system of one binocular optical camera and one binocular infrared sensor to capture corresponding two modality images. We used M3FD Fusion dataset to train our GSR model which consists of 300 aligned pair images from different scenarios in Daytime, Night, and Overcast. The dataset consists of images from different scenes such as road, campus, street, forest, and many others.

The second datasets that have been used for evaluating our Guidance Super Resolution (GSR) model is RGB-NIR Scene dataset, which consists of 477 paired images taken in RGB and Near-infrared (NIR). As it mentioned previously, This dataset includes many scene types such as country, field, forest, inside, mountain, oldbuilding, street, urban, and water.

2.4 Assessment Methods

Image quality assessment (IQA) commonly falls into two categories: objective and subjective methods. Objective approaches have quickly become the standard for evaluating restoration tasks due to their transparency and consistency. However, they can only represent the recovery of image pixels from a numerical point of view, and it is difficult to precisely quantify the actual visual effect of the image. On the other hand, subjective approaches are always dependent on human subjective assessments and are primarily concerned with evaluating the perceptual quality of the image. Based on the pros and cons of the two types of methods mentioned above, several assessment methods are briefly introduced in the following SubSections.

2.4.1 Image Reconstruction Accuracy

The assessment methods applied to evaluate image reconstruction accuracy are also called *Distortion measures*, which are full-reference. Specifically, given a distorted image $\hat{x}$ and a ground-truth reference image x, full-reference distortion quantifies the quality of $\hat{x}$ by measuring its discrepancy to x using different algorithms.

Peak Signal-to-Noise Ratio

The Peak Signal-to-Noise Ratio (PSNR) [161] is the most frequently used image quality assessment (IQA) approach in the restoration tasks, and it can be calculated by the use of the mean squared error (MSE) between the ground truth image $I_y \in \mathbb{R}^{H\times W}$ and the reconstructed image $\hat{I}_y \in \mathbb{R}^{H\times W}$:

$$MSE = \frac{1}{HW}\sum_{i=0}^{H-1}\sum_{j=0}^{W-1}(I_y(i,j) - \hat{I}_y(i,j))^2, \tag{2.9}$$

$$PSNR = 10 \cdot log_{10}\left(\frac{MAX^2}{MSE}\right), \tag{2.10}$$

where MAX is the maximum possible pixel of the image. Since PSNR is highly related to MSE, a model trained with the MSE loss will be expected to have high PSNR scores. The higher the PSNR value, the smaller the difference between the reconstructed image and the original image, which means the better the image quality [90]. However, since the PSNR is based on the global statistics of image pixel values, the local visual factors of the human eye are not considered. As for human eyes, the sensitivity to different regions is different, and the perception result of a specific area is also affected by the surrounding neighboring areas, so the evaluation results of PSRN may have deviated from the perception of the human eye.

Structural Similarity Index Measure

The Structural Similarity Index Measure [161], also known as SSIM, is an another well-known evaluation method. There is a strong correlation between their pixels, which frequently contains crucial data about the object's structure due to the complex structure of natural images. While the human visual system primarily receives structural information from the visible region, it is feasible to perceive the approximate knowledge information of the image distortion by detecting the deterioration of the structural information. The measurement system of SSIM consists of three measurement modules: brightness, contrast, and structure. Thus the SSIM can be expressed as a weighted combination of three comparative measures:

$$\begin{aligned} SSIM(\hat{I}_y, I_y) &= (l(\hat{I}_y, I_y))^{\alpha} \cdot c(\hat{I}_y, I_y))^{\beta} \cdot s(\hat{I}_y, I_y))^{\gamma} \\ &= \frac{(2\mu_{\hat{I}_y}\mu_{I_y} + c_1)(2\sigma_{\hat{I}_y I_y} + c_2)}{(\mu^2_{\hat{I}_y} + \mu^2_{I_y} + c_1)(\sigma^2_{\hat{I}_y} + \sigma^2_{I_y} + c_1)}, \end{aligned} \tag{2.11}$$

where l, c, and s represents luminance, contrast, and structure between $\hat{I}_y$ and I_y, respectively, $\mu_{\hat{I}_y}$, μ_{I_y}, $\sigma^2_{\hat{I}_y}$, $\sigma^2_{I_y}$, and $\sigma_{\hat{I}_y I_y}$ are the average(μ) / variance (σ^2) / covariance (σ) of the corresponding items. A higher SSIM indicates higher similarity between two images, which has been widely used due to its convenience and stable performance on evaluating the perceptual quality.

2.4.2 Image Perceptual Quality

Because the human visual system is complicated and involves numerous variables to determine the differences between two images, such as flow and textures within the images, approaches that seek absolute similarity differences (PSNR/SSIM) will not always work well. Despite the widespread use of distortion techniques, improvements in reconstruction accuracy are not always matched by improvements in visual quality. Indeed, studies have demonstrated that distortion and perceptual quality are at conflict in some cases [16]. The image perceptual quality of an image $\hat{x}$ is defined as the degree to which it looks like a natural image, which has nothing to do with its similarity to any reference image.

Natural Image Quality Evaluator

Natural Image Quality Evaluator (NIQE) [117] is a fully blind approach for evaluating image quality. NIQE exclusively uses quantified deviations from statistical regularities observed in natural images, without the need for prior information about predicted distortions in the form of training samples and matching human assessment scores. It generates a group of local (quality-aware) image features based on a natural scene statistics (NSS) model and then fits the extracted feature vectors to a multivariate Gaussian (MVG) model. The quality of a test image is then estimated based on the distance between its MVG model and the MVG model learned from a natural image. NIQE is formulated as follows:

$$D(v_1, v_2, \sum_1, \sum_2) = \sqrt{((v_1 - v_2)^T (\frac{\sum_1 + \sum_2}{2})^{-1} (v_1 - v_2))}, \tag{2.12}$$

where v_1, v_2 and $\sum_1$, $\sum_2$ are the mean vectors and covariance matrices of the HR and SR image's MVG model. Notice that, a higher NQIE index indicates lower image perceptual quality.

Perceptual Index

In the 2018 PIRM Challenge on Perceptual Image Super-Resolution [17], perception index (PI) is first proposed to evaluate the perceptual quality. It is a combination of the no-reference image quality measures Ma and NIQE:

$$PI = \frac{1}{2}((10 - Ma) + NIQE). \tag{2.13}$$

Among them, Ma Ma et al. [110] is a non-reference quality indicator applied in the field of image super-resolution reconstruction, which does not refer to real images. It designs the types of low-level statistical features in the spatial and frequency domains to quantify super resolution artifacts, and learning a two-stage regression model to predict the quality score of SR images. A lower PI, better perceptual quality.

2.4.3 Learned Perceptual Image Patch Similarity Metric

LPIPS has been introduced by [189]; which stands for Learned Perceptual Image Patch Similarity (LPIPS). LPIPS is a metric for determining how visually similar two images are to the human eye. Using a pre-defined network like VGG, AlexNet, or SqueezeNet, LPIPS calculates the similarity between the activations of two image patches. This metric has been demonstrated to be an effective approximation of human perception. Image patches with a low LPIPS score are perceptually similar, whereas those with a high score are more dissimilar.

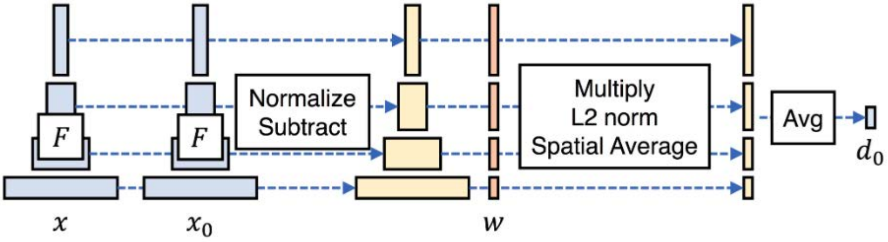


Figure 2.12: Learned Perceptual Image Patch Similarity computing distance from a network [189].

In this illustration d_0 is compute distance between two patches, x, x_0, given a network $\mathscr{F}$, first compute deep embeddings, normalize the activations in the channel dimension, scale each channel by vector w, and take the ℓ_2 distance. Then average across spatial dimension and across all layers.

2.4.4 Frechet Inception Distance

Generative Adversarial Networks(GANs) are very difficult to evaluate as compared to other networks. And, it is very important to evaluate the quality of GANs, because it can help us in choosing the right model, or when to stop the training, or how to improve the model. Out of several methods, Frechlet Inception Distance(FID) [60] is one performance metric to evaluate the quality of GANs.

FID is a performance measure that represents the difference between the feature vectors of real and fake images generated by the GAN's generator. A lower FID corresponds to images with higher quality; conversely, a higher score corresponds to images with poorer quality. [60] demonstrate that there is a correlation between lower FID scores and higher image quality when systematic distortions such as the addition of random noise and blur are performed.

The FID score is then calculated using the following equation:

$$d^2 = ||mu_1 – mu_2||^2 + Tr(C_1 + C_2 – 2 * sqrt(C_1 * C_2)) \tag{2.14}$$

where the score is referred to as d^2, showing that it is a distance and has squared units. The mu_1 and mu_2 refer to the feature-wise mean of the real and generated images. The C_1 and C_2 are the covariance matrix for the real and generated feature vectors, often referred to as sigma. The $||mu_1 – mu_2||^2$ refers to the sum squared difference between the two mean vectors. Tr refers to the trace linear algebra operation. The sqrt is the square root of the square matrix, given as the product between the two covariance matrices.

2.4.5 Cosine Similarity

It is possible to quantify the degree of two vectors, which are similar by computing their cosine similarity. To be more precise, it evaluates the degree of correspondence between the vectors' directions or orientations while disregarding any variations in their magnitude or scale. It is necessary that both vectors belong to the same inner product space for inner product multiplication to yield a scalar.

Cosine similarity is described mathematically as the division between the dot product of vectors and the product of the euclidean norms or magnitude of each vector.

$$\text{similarity } = \cos(\theta) = \frac{\mathbf{A}\cdot\mathbf{B}}{\|\mathbf{A}\|\,\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}, \tag{2.15}$$

where θ is the angle between the vectors. $A \cdot B$ is dot product between A and B and calculated as $A \cdot B = A^T B = \sum_{i=1}^{n} A_i B_i = A_1 B_1 + A_2 B_2 + \ldots + A_n B_n$. $\|A\|$ represents the L2 norm or magnitude of the vector which is calculated as $\|A\| = \sqrt{A_1^2 + A_1^2 \ldots A_1^n}$

The similarity can take values between -1 and +1. Smaller angles between vectors produce larger cosine values, indicating greater cosine similarity.

2.4.6 Reconstruction Efficiency

Although constructing deeper networks is the simplest technique to improve reconstruction performance, these models will also include additional parameters, execution time, and computing expenses. Thus, it is necessary to examine the trade-off between model performance and model complexity in expanding the practical applications. Therefore, it is critical to analyze the reconstruction efficiency using the fundamental metrics listed below.

- **Model size**: The model size is related to the storage that the devices need to store the data. A model containing more parameters is harder for the device with limited hardware to run it. Therefore, building lightweight models is conducive to the promotion and application of the algorithm. Among all the indicators, the parameter quantity of the model is the most intuitive indicator to measure the model size.

- **Execution Time**: A lightweight model typically requires a fast execution time, but the development of complicated methods has disrupted this equilibrium such as attention mechanism. i.e., when some sophisticated operations are incorporated into the model, a lightweight network may require a significant amount of time to execute. As a result, it is critical to evaluate the model's execution time.

- **Multi-Adds**: The main operations in the CNN approaches are multiplications and additions. The number of multiply-accumulate operations, is always employed to quantify model computation. The value of Multi-Adds is proportional to the time required to execute the model. To summarize, the trade-off between model performance and model complexity must still be considered.

2.5 Optimization Objective

In this section, we will introduce the necessary procedures during the model training.

2.5.1 Learning Strategy

Deep learning-based models can be mostly split into two categories, supervised learning methods and unsupervised learning methods based the learning methods.

Supervised Learning

In supervised learning restoration tasks, researchers compute the reconstruction error between the ground-truth image I_y and the reconstructed image $\hat{I}_y$:

$$\hat{\theta}_F = argmin_{\mathscr{F}} \mathscr{L}(\hat{I}_y, I_y). \tag{2.16}$$

Alternatively, researchers may sometimes search for a mapping ϕ, such as a pre-trained neural network, to transform the images or image feature maps to some other space and then compute the error:

$$\hat{\theta}_F = argmin_{\mathscr{F}} \mathscr{L}(\Phi(\hat{I}_y, \phi(I_y))). \tag{2.17}$$

Among them, $\mathscr{L}$ is the loss function which is used to minimize the gap between the reconstructed image and ground-truth image.

Unsupervised Learning

The term "unsupervised learning" refers to the process of using machine intelligence (AI) algorithms in order to recognize patterns in datasets that do not contain data points that have been categorized or labeled in any way. For example, in CinCGAN[179], a model consists of two CycleGAN [195], where parameters are upgraded through optimizing the generator-adversarial loss, the cycle consistency loss, the identity loss, and the total variation loss together in each cycle.

2.5.2 Loss Functions

The effectiveness of the deep learning models depends heavily on the loss function. Applying the appropriate loss function for a task ensures that the model learns the most relevant information for a faster, more accurate convergence. Several loss functions have been developed to penalize different aspects of the image restoration tasks to improve the quality of the restored images. Loss functions are used in deep learning models generally is a weighted sum of more than one loss function. By doing so, network can focus on different problems. In this section, we will take a closer look at the loss functions widely used in restoration tasks.

Pixel Loss

Pixel loss function is used as a metric for understanding differences between images on a pixel level and measure of how far is the target image pixels are from the generated image pixels. It mainly includes the L1 loss (*i.e.,* mean absolute error) and L2 loss (*i.e.,* mean square error):

$$\mathscr{L}_{L1}(\hat{I}_y, I_y) = \frac{1}{hwc}\sum_{i,j,k}\left|\hat{I}_y^{i,j,k} - I_y^{i,j,k}\right|, \tag{2.18}$$

$$\mathscr{L}_{L2}(\hat{I}_y, I_y) = \frac{1}{hwc}\sum_{i,j,k}\left(\hat{I}_y^{i,j,k} - I_y^{i,j,k}\right)^2, \tag{2.19}$$

where h, w and c are the height, width, and the number of channels of the image.

While L2 loss favors a high PSNR, L1 loss is believed to be more robust against outliers. Pixel Loss is one of the loss functions that has been widely used in the literature.

Content Loss

The content loss proposed by [75] to evaluate the perceptual quality of images. Specifically, it measures the semantic differences between images using a pre-trained image classification network. Denoting this network as ϕ and the extracted high-level representations on l-th layer as $\phi^{(l)}(I)$, the content loss is indicated as the Euclidean distance between high-level representations of two images, as follows:

$$L_{content} = (I_y, \hat{I}_y; \phi, l) = \frac{1}{h_l w_l c_l} \sqrt{\sum_{i,j,k} (\phi^{(l)}_{i,j,k}(\hat{I}_y) - \phi^{(l)}_{i,j,k}(I_y))^2}, \tag{2.20}$$

where h_l, w_l and c_l are the height, width and number of channels of the representations on layer l, respectively.

Essentially the content loss transfers the learned knowledge of hierarchical image features from the classification network ϕ to the different image restoration task such as SR network. In contrast to the pixel loss, the content loss encourages the output image $\hat{I}_y$ to be perceptually similar to the target image I_y instead of forcing them to match pixels exactly. Thus it produces visually more perceptible results.

Adversarial Loss

In recent years, a GANs [45] receive more attention in various vision tasks due to their powerful learning ability. To be more specific, GAN usually include of a generator and discriminator to generate and validate the image, whether each input comes from the target distribution or not. During training, through adequate iterative adversarial training, the generator try to produce outputs consistent with the distribution of real data, while the discriminator try to distinguish between the generated data and real data.

In other word, the generator tries to minimize the following function while the discriminator tries to maximize it:

$$E_x[\log(D(x))] + E_z[\log(1 - D(G(z)))] \tag{2.21}$$

where D(x) is the discriminator's estimate of the probability that real data instance x is real. E_x is the expected value over all real data instances. $G(z)$ is the generator's output when given noise z. D(G(z)) is the discriminator's estimate of the probability that a fake instance is real. E_z is the expected value over all random inputs to the generator (in effect, the expected value over all generated fake instances G(z)).

Adversarial loss has been used in several restoration tasks such as Mao et al. [113], Wang et al. [160], Yuan et al. [179].

Cycle Consistency Loss

The problem with only using adversarial loss is that the network can map the same set of input images to any random permutation of images in the target domain. Any of the learned mappings can, therefore, learn an output distribution that is similar to the target distribution. There can be many possible mapping functions between two domains. Cycle Consistency Loss has been introduced by [195] to overcome the aforementioned problem. Let say we are having two image domains Cycle Consistency Loss captures the intuition that if we translate the image from one domain to the other and back again we should arrive at where we started. Hence, it calculates the L1 loss between the original image and the final generated image, which should look same as original image. The cycle consistency loss is represented as:

$$\begin{aligned}\mathscr{L}_{\mathrm{cyc}}(G,F) &= \mathbb{E}_{x\sim p_{\mathrm{data}}(x)}[\|F(G(x))-x\|_1] \\ &+ \mathbb{E}_{y\sim p_{\mathrm{data}}(y)}\left[\|G(F(y))-y\|_1\right]\end{aligned} \tag{2.22}$$

where $||x||$ denotes the mean absolute error, or MAE, of x. Taking the mean absolute error of x and y, where x and y are both n dimensional vectors, is a method of measuring the distance between those vectors. MAE takes the absolute distance of each element, and then averages that into a single number.

Identity Loss

The identity loss also introduced by [195], which encourage the generator to preserve the color composition between input and output. This is done by providing the generator an image of its target domain as an input and calculating the L1 loss between input and the generated images. The identity loss is simple, G(y) should $\approx y$ and F(x) should $\approx x$. The identity loss can written as:

$$\begin{aligned}\mathscr{L}_{\mathrm{identity}}(G,F) &= \mathbb{E}_{y\sim p_{\mathrm{data}}(y)}\left[\|G(y)-y\|_1\right] \\ &+ \mathbb{E}_{x\sim p_{\mathrm{data}}(x)}[\|F(x)-x\|_1]\end{aligned} \tag{2.23}$$

2.5.3 Other Improvements

In addition to the learning strategies, there are other techniques further improving deep learning models such as:

- **Data augmentation**. One of the most extensively utilized approaches for improving deep learning performance is data augmentation. Scaling, rotating, cropping, and flipping are only a few examples data augmentation techniques.

- **Network interpolation**. PSNR-based models produce images closer to ground truths but introduce blurring problems, while GAN-based models bring better perceptual quality but introduce unpleasant artifacts (*e.g.,* meaningless noise-making images more realistic). In order to better balance the distortion and perception, Wang et al. [158] propose a network interpolation strategy. Specifically, they train a PSNR-based model and train a GAN-based model by fine-tuning, then interpolate all the corresponding parameters of both networks to derive intermediate models. By tuning the interpolation weights without retraining networks, they produce meaningful results with much fewer artifacts.

- **Self-ensemble**. Self-ensemble, also known as enhanced prediction [152], is an inference technique commonly used. For example, in SR task, rotations with different angles $(0°, 90°, 180°, 270°)$ and horizontal flipping are applied on the LR images to get a set of 8 images. Then these images are fed into the SR model and the corresponding inverse transformation is applied to the reconstructed HR images to get the outputs. The final prediction result is conducted by the mean or median of these output.

  Self-ensemble is a frequent inference technique to improve the prediction. For instance, in SISR rotations with different angles $(0°, 90°, 180°, 270°)$ and horizontal flips are performed to the LR images to produce a collection of 8 images. The images are then used as input into the SR model, and later the associated inverse transformation is applied to the reconstructed HR images to produce the outputs. The overall forecast result is determined by the mean or median of these outputs.

2.6 Most Related Network Frameworks

This section focuses on reviewing the most related SOTA deep learning-based approaches to this thesis.

2.6.1 Image Colorization Frameworks

Colorization problem has been studied during last decades, several techniques have been proposed to unravel this difficult task. Some of the methods proposed in the literature follow a semi-automatic approach, which means they need user interactions or to employ some user-defined search table. Other approaches, mainly learning based approaches, are based on having aligned image pairs (NIR-RGB), which in most of the cases are not available. The issues mentioned within the current work are expounded with infrared image colorization, as mentioned above, it somehow shares some common issues with monochromatic approaches to image colorization. Monochromatic image colorization algorithms vary in the ways they obtain and process data for modeling between gray-scale and RGB images.

Colorization approaches can be usually classified into two groups: parametric and non-parametric. At the training time, parametric techniques try to learn predictive functions from large color image datasets, posing the problem either as a classification of quantified color values or as a regression to continuous color space. On the other hand, in the non-parametric techniques a gray-scale image is provided as an input and then one or more color provided as source images by user or automatically; then color from reference images transferred statistics onto homogeneous regions of the input image, such as Welsh et al. [165], Gupta et al. [52], Irony et al. [70]. All the papers mentioned before are example-based approaches, which works as semi-automatic methods to transfer color statistics from reference images onto input gray-scale images. Although good results are obtained, there is a big drawback with all these techniques that is related with the requirement that input and reference images should share the same content, actually both of them should be perfectly registered, which is not the case in most of the real scenarios.

GAN networks are a kind of Convolutional Neural Network (CNN) that are able to generate samples from a given latent space, this network has been introduced by Goodfellow et al. [45]. The mentioned GAN architecture build up by a series of linear layers (fully-connected layers) and so insufficient to complex dataset. The model consists of two networks a Discriminator (D) and a Generator (G), which going to against each other. In other words, the discriminator try to distinguish the real samples from fake samples that have been generated by the generator. On the other hand, the generator job is to fool the discriminator with the generated samples (fake images) to be classified as real images. Both networks, D and G, are simultaneously optimized. As mentioned above, the main issue of the standard GAN was limited to simple datasets so shortly afterwards researcher proposed the new architecture for GAN to address this limitation, DCGAN (Deep Convolutional Generative Adversarial Network) [128] has been introduced by Radford et al. and changed the standard for most of GAN architectures. DCGAN architecture became one of the most popular and successful network design for GAN. In DCGAN instead of using series of linear layers that is only suitable for simple datasets, convolution layers without max pooling or fully connected layers are considered and furthermore convolutional stride for the down-sampling and transposed convolution for the up-sampling are used that made DCGAN architecture appropriate for complex dataset. DCGAN standard has been applied in various computer vision problems such as image colorization [144], image enhancement [86], style transfer [26], data augmentation [4] and many others.

Previous approaches are useful when paired images are provided for the training process. In the case of unpaired images, architectures such as CycleGAN [195] or Dual-GAN [175] have been proposed by learning mapping between different visual domains jointly, each as a separate generative adversarial network. Via a cycle-consistency loss ensures that applying each mapping followed by its reverse yields the identity map (i.e., *"if we translate from one domain to another and back again we must arrive where we started"*).

Regardless of the used architecture, Generative Adversarial Networks usually suffer from multiple challenges during training that needs more attention than Convolutional Neural Networks (CNNs). Such as mode collapse, convergence properties, diminished gradient and highly sensitive to the hyper-parameter selections. Arjovksy et al. [7] illustrated that the discriminator in standard GAN cannot be trained well or with a high learning rate; otherwise gradient vanish may show off and generator not able to generate samples anymore and learning will stop. They also proved that the standard GAN loss function cannot accurately deal with inappropriate distributions, for example those with disjoint supports, often found during training stage of GAN. To solve the mentioned challenges many different GANs have been proposed by using vary loss functions during training or using different *D* during the learning process such as LSGAN [113], WGAN [8] and many others. Even though the proposed approaches have been relatively successful solving these challenges (training stability, data quality, etc.), Lucic et al. [107]'s large - scale research suggests that such approaches are not improving standard GAN consistently. In addition, some of the best proposed approaches, like WGAN-GP [47], requires far more computational comparing with standard GAN. Alexi [76] illustrated that a relativistic discriminator based on integral probability metrics (IPM), is essential to make GANs similar to divergence minimization and generate reasonable forecasts on the basis of a previous knowledge. In such discriminator, half of the images in the mini-batch consider as fake. The proposed approach prove that GANs are able to generate higher quality samples, less computational and more stable than the previous approaches.

2.6.2 Single Image Super Resolution Frameworks

In this section, recent state-of-the-art SR deep learning approaches are detailed. In below subsections, SR lightweight models, which focus on compressing the number of parameters and operations, attention mechanism, and vision transformers are reviewed.

Deep Learning Based Single Image Super-Resolution

Dong et al. [35] present one of the first work using CNN to tackle the SR task (i.e., SRCNN). The SRCNN receives an upsampled image as an input that cost extra computation. Later on, to address this drawback, FSRCNN [37] and ESPCN [137] have been proposed to reduce the large computational and run time cost by upsampling the features near to the output of the network. This tactic leads results in efficient approaches with low memory compared to SRCNN. However, the entire performance could be reduced if there are not enough layers after the upsampling process. In addition, they cannot manage multi-scale training, as the size of the input image differs for each upsampling scale.

Even though the strength of deep learning shows up from deep layers, the above-mentioned methods are referred to as shallow network due to the training difficulties. Therefore, Kim et al. [78] use residual learning to ease the training challenges and increase the depth of their network by adding 20 convolutional layers. Then, [147] has proposed memory block in MemNet for deeper networks and solve the problem of long-term dependency with 84 layers. Thus, CNN-based SR approaches demonstrate that deeper networks with various types of skip connections show better performance. Thereby, Lim et al. [95] introduce EDSR by expanding the network size and enhancing the residual block by omitting the batch normalization from residual block. Zhang et al. [192] propose RDN with residual and dense skip connections to fully use hierarchical features. Li et al. [89] propose a network with more than 160 layers plus improved residual units. Despite of the fact that they achieve higher PSNR values, the number of parameters and operations are increased, which leads to high risk of overfitting and limits for real-world applications.

Deep Learning Lightweights Single Image Super Resolution

In recent years the interest of building lightweight and efficient models has been increased in SISR to reduce the computational cost. Several lightweight networks have been introduced, such as SRCNN [35], FSRCNN[37], ESPCN[137], which were the first attempts, but they could not perform well. Later, Ahn et al. [1] design a network that is suitable in the mobile scenario by implementing a cascade mechanism beyond a residual network (CARN), in order to obtain lightweight and improve reconstruction but it is at the cost of reduction of PSNR. Then, a neural architecture search (NAS)-based strategy has been also proposed in SISR to construct efficient networks—MoreMNA-S [29] and FALSR [28]. Howerver, due to limitation in strategy, the performance of these models are limited. Later, [120] introduces MAFFSRN by proposing multi-attention blocks to improve the performance. Recently, LatticeNet [100] introduces an economical structure to adaptively combine Residual Blocks, which achieve good results. All these works suggest that the lightweight SR networks can keep a good trade-off between PSNR and parameters.

Attention Mechanism

Attention can be described as a guide to bias the allocation of available computer resources to the most important informative elements of an input. Recently, some works have focused on attention mechanism for deep neural networks. Hu et al. [63] introduce squeeze-and-excitation (SE) block, a compact module to leverage the relationship between channels. Also, Woo et al. [167] propose a Convolutional Block Attention Module (CBAM) to exploit the inner-spatial and inner-channel relationship of features to achieve a performance improvement in image classification.

Recently, RCAN [191] designs a very deep network with a channel attention mechanism to enhance the reconstruction results by only considering inner-channel information, which call first-order statistics. In contrast, Dai et al. [31] introduce the second-order attention network in order to explore more powerful feature expression. More recently, Li et al., [100] propose enhanced spatial attention (ESA) to make the residual features to be more focused on critical spatial contents.

Vision Transformer

Transformer networks shows breakthrough performance in the Natural Language Process (NLP). In contrast to ConvNets, Transformer networks have advantage of capturing long-range dependency in the input with the global self-attention. The core idea of Transformer is "self-attention" module, which is capable of capturing long-term information between sequence elements.

The impressive performance of Transformer-based networks in the NLP domain, inspires the computer vision community to adapt the Transformer for vision tasks. The first work in this direction has been done by Alex et al. who proposes ViT [38] as a Vision Transformer, which replace the standard CNN with Transformer and directly train on the medium size flattened patches with large-scale data pretraining.

Since introducing the first work, many Transformer-based architectures have been proposed for the vision tasks in image recognition [178], object detection [21, 104], segmentation [157, 169], and action recognition [126, 136]. In addition, Transformer based models have been studied for the low-level vision problems such as super resolution [93, 106], image colorization [82], denoising [164], and image restoration [180]. For instance, DETR [21] is a transformer network designed for object detection, which can predict a set of objects and model their relationships. SwinIR introduced by Jingyun et. al. [93] for low-level vision tasks by using Swin Transformer [104], which applying self-attention within local image regions to solve the low-level vision problems.

Although the Transformer based networks achieve excellent performance in low-level Vision tasks, these methods are still depends on providing heavy GPU resources to train the model, which is not feasible or available to most of the researchers. Also, the computational complexity of self-attention in Transformers can increase quadratically with the number of token to mix (i.e., image patches), thereby prohibiting its application to high-resolution images.

2.6.3 Guidance Super Resolution Frameworks

Guidance Super Resolution (GSR) techniques have been used to upsample images from a different domains to generate more accurate SR images by using the information of other domain images (i.e, visible images) while having such a high resolution infrared images are expensive. Traditional GSR approaches, such as joint bilateral upsampling [80] and rapid bilateral filtering [11] are already studied for this task, however these methods frequently over-smooth the reconstructed image. Recently, by advancing deep learning methods several approaches have been introduced to boost performance of GSR task. GSR techniques have been studied in different super resolution domains such as depth-map SR, infrared SR, thermal SR, hyperspectral SR and some others. MSG-Net [68], employ CNNs to accomplish guidance super resolution, which was the first CNN model that attempts to upsample depth images under multi-scale guidance from the corresponding HR visible images.

Most of GSR methods are based on the Siamese algorithm, which let the network to accept two inputs and perform simultaneous feature extraction from other spectral images and visible images. These images are then fused in different level of the network and upsampled to provide a high-resolution images. Furthermore, GSR approaches with similar structure used in guidance hyperspectral SR methods include [83, 138]. Also, some models proposed for guidance infrared SR such as [24, 141]. Feras et al. [3] propose a multimodal sensor fusion model to enhance the thermal images with help of RGB images. Also, some approaches for cross-modal guidance super resolution extract edges from the visible images in order to obtain high-frequency features. The use of edge-based guiding facilitates the reconstruction of higher-frequency features such as [170, 193]. Despite that the aforementioned approaches achieve reasonable performance, these method are limited to a fix scale factor and not ideal for real world application due to number of network parameters and their performance.

2.7 Summary

This chapter begins with a discussion of the problem definitions. Then, we discuss benchmark datasets, evaluation methodologies, frameworks, and optimization targets. Finally, a summary of relevant works is given.

3 Colorizing Near Infrared Images through a Cyclic Adversarial Approach of Unpaired Samples

This chapter presents the article published at:

Armin Mehri, and Angel D. Sappa. "Colorizing near infrared images through a cyclic adversarial approach of unpaired samples." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2019.

This paper presents a novel approach for colorizing near infrared (NIR) images. The approach is based on image-to-image translation using a Cycle-Consistent adversarial network for learning the color channels on unpaired dataset. This architecture is able to handle unpaired datasets. The approach uses as generators tailored networks that require less computation times, converge faster and generate high quality samples. The obtained results have been quantitatively—using standard evaluation metrics—and qualitatively evaluated showing considerable improvements with respect to the state of the art.

3.1 Motivation

In recent years, image acquisition devices have expanded significantly due to the increase in computational power and the reduction in electronics prices. Improving the sensor technology has lead to a large family of cameras capable of capturing information from various spectral bands or additional information (3D, 4D); hence nowadays we can have: panoramic 3D images; multispectral images; HD 2D images; video sequences at a high frame rate; and many others. Regardless of the large number of possibilities, the fact that the human visual system is sensitive to (400-700 nm) the classical RGB representation is preferred if the information needs to be provided to the final user. Therefore, for better user understanding, representing the information in the range of 400-700nm is preferred [144]. Out of this spectral range the NIR band is one of the most widely used band.

The NIR spectral band is the nearest band to the human eye perception system, so NIR images share various attributes with visible spectrum images. The use of NIR images is concerned with their ability to segment images according to the material of the object. For example, most coloring matter utilized for colorization of materials are slightly transparent to NIR. In other word, the distinction within the NIR intensities is not solely due to specific color of the material, however conjointly to the absorption and coefficient of reflection of the materials of a given object. The aforementioned attributes (absorption and reflectance) are interesting for applications such as video surveillance, detection and remote sensing for crop stress. In these two contexts (i.e., video surveillance and remote sensing), it is quite troublesome to orient once near infrared images are provided to the final users, as color discrimination is lacking or incorrect color deployment. Hence, obtaining realistic RGB image representations from NIR images is a needed in most of these applications.

NIR image colorization shares some similarities with those approaches proposed in the literature for gray scale image colorization or color transfer functions (e.g., [124], [25], [187]). In spite of the similarity with these approaches, due to the nature of NIR images, their colorization is more challenging. In recent years several approaches for NIR image colorization have been proposed (e.g., [144], [96], [145]). Most of them are learning based approaches where couple of registered NIR and RGB images are provided during the training stage. The limitation with all these approaches is related with the need of these couple of paired images (NIR-RGB). In general, although there are Single Sensor Cameras (e.g., [142]) where RGB and NIR information is acquired at the same time, NIR images are taken by one camera while the corresponding RGB image by another camera. This means that there are shifts between the acquired images, or in some cases even worse since just the NIR images are provided.

In the last few years, Generative Adversarial Networks (GANs) have drawn attention in many field of computer vision to help researchers to build powerful models, where there were difficulty by using only simple Convolutional Neural Networks (CNNs). Nevertheless, most of the GAN techniques [71] have focused on supervised context. The unpaired (NIR-RGB) problem mentioned above, can be tackled by a GAN architecture in the unsupervised context under a cyclic structure (CycleGAN) [195]. CycleGAN learns to map images from one domain (source domain) onto another domain (target domain) when paired images are unavailable. This functionality makes models appropriate for image to image translation/colorization in the context of unsupervised learning. In the current paper a novel CycleGAN architecture is proposed for colorizing unpaired NIR-RGB images; the main contributions of our proposed model, compared with baseline [195], are as follows: *i*) it utilizes tailored generators, which can work better in colorization context and

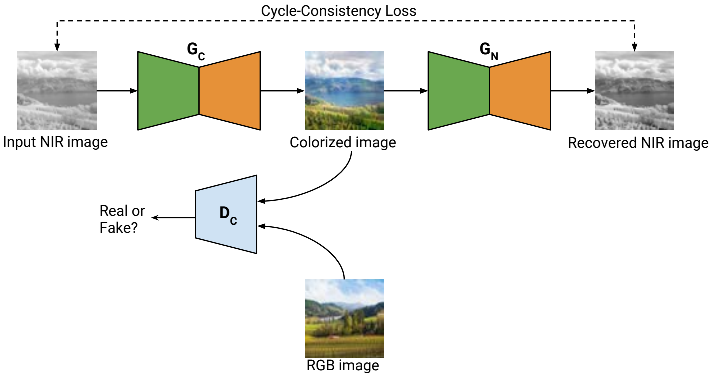


Figure 3.1: Illustration of the CycleGAN architecture used for NIR image colorization with unpaired NIR-RGB datasets.

have less computation time and less parameters size; ii) it converges faster than the baseline approach [195]; finally, iii) it produces higher quality images.

3.2 Proposed Approach

This section explains in details the approach proposed for colorizing NIR images. As mentioned previously most of recent work on colorization have proposed the usage of a deep convolutional generative adversarial network on aligned paires of images, which in most of the cases do not represent a real scenario. In the current work the usage of a CycleGAN to colorize NIR images to a RGB representation is proposed (see Figure 3.1), when an aligned paired dataset does not exist.

In order to handle inputs and outputs in both generators, a model that is feed with three channels is proposed. This model will receive as an input three channels, which could correspond to: i) a given NIR image three times (this is in the G_C case); or ii) a RGB image (this is the G_N case). A loss function different to the one proposed in [195] is used to minimize the overall classification error in the training process, which improves the generalization capability of the model.

The proposed architecture built up by a series of convolutional and transposed convolutional layers; relu and leaky relu as non-linear activation functions; for generators and discriminators respectively. Moreover, every layer of D uses the spectral normalization and instance normalization in G. Also, it is worth to mention pooling layer have not been used in the networks, instead strided convolutions

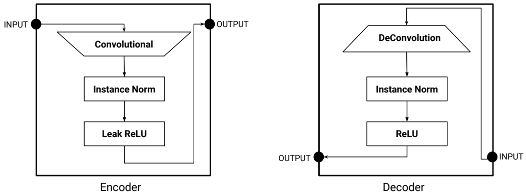


Figure 3.2: Illustration of the Encoder and Decoder structures of the proposed approach.

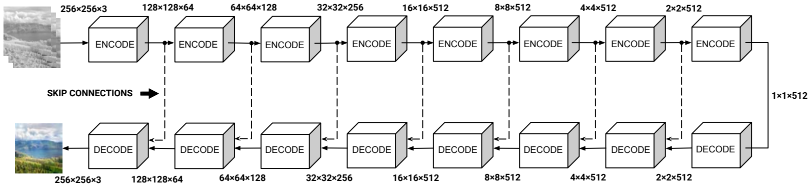


Figure 3.3: Illustration of the structure of generators for NIR image colorization.

used in order to keep as much features as possible, since a pooling layer is downsampling the feature depends on stride number, which leads to data in features map to loss. Dropout layers are used in the terms of noise to few layers of generators in order to prevent overfitting and modal collapse. Added noise in few layers of *G* leads network to generate the necessary variability of the training set, to be able to generalize the learning of the colorization process.

Both networks (*G*, *D*) are based on feed-forward deep neural networks, which play a min-max game against each other. The near infrared image given to *G* as an input data with the image size 256×256 pixels, and networks try to transforms the given sample (NIR image) onto the interested form of the data we concerned, a RGB representation. On the other hand, *D* takes a set of data, either a real sample or a produced sample, and produces a probability of that data being real. The network *D* is optimized in order to increase the likelihood of giving a high probability to the real data and a low probability to the generated data (i.e., if the probability is near to 1 it means the NIR image is correctly colored, while if probability is near to 0 it means that NIR image is wrongly colored).

3.2.1 U-net as Generators

ResNet architecture [55] has been used as a part of the generators in the CycleGAN [195], showing that it is quite powerful in transfiguring one image to another image and has been achieved the reasonable results in style transfer, photo enhancement, season transfer and other several applications. Unfortunately, it could not achieve acceptable results in learning the color between domains and transferring the learned color without affecting the samples' shapes, since the network after some number of epochs starts to transfigure between domains; hence the net will stop to learn the color and also networks need plenty of data in RGB domain. The first contribution of current work with respect to baseline model [195] is to select the generators, which are able to generate better samples and learn accurately the colors of the different objects in images and works better when not enough data are available in both domains. The architecture proposed in this paper (U-net [133]) also leads to have less computational time in training process. U-net based architecture [133], proposed as generators of the models, has showed efficiency on a wide range of approaches, especially in the colorization problems (e.g., [71],[188]).

The Unet architecture build up based on three components: 1) encoder where the input passes through a series of down-sampling layers (i.e., convolutional layers to extract the feature samples); 2) bottleneck layer, which helps the model to share all information pass through all the layers so low-level information will be available directly among the net. To give the generator a means to circumvent the bottleneck for information like this, in the pix2pix model [71] the skip connections, following the general shape of a "U-Net", which simply concatenates all channels at layer i with those at layer $n-i$; 3) decoder, the last component of U-net, which do the reverse process of the encoder (i.e., back to the normal image from the extracted feature by pass through the series of transposed convolutional layers (up-sampling)). Also, U-net shows that is quite powerful in the case of understanding the color from one domain and transferring onto another domain (NIR images). U-net architecture with skip connections is illustrated in Figure 3.3.

3.2.2 Loss Functions

In the proposed model a multi-term loss function ($\mathscr{L}_{final}$) has been used by combination of RaLSGAN loss, Cycle Consistency loss, Structural Similarity loss (SSIM) and Identity loss. The combination of these loss functions leads to achieve better image quality for human perceptual criteria as presented in the experimental result section.

The RaLSGAN loss function [76] is applied to both generators G_C, G_N and their discriminators D_C, D_N of the model respectively:

$$\mathscr{L}_{RaLSGAN}^{G_i} = \mathbb{E}_{x_f \sim \mathbb{P}}[(C(x_f) - \mathbb{E}_{x_r \sim \mathbb{P}} C(x_r) - 1)^2] + \mathbb{E}_{x_r \sim \mathbb{P}}[(C(x_r) - \mathbb{E}_{x_f \sim \mathbb{Q}} C(x_f) + 1)^2] \tag{3.1}$$

$$\mathscr{L}_{RaLSGAN}^{D_i} = \mathbb{E}_{x_r \sim \mathbb{P}}[(C(x_r) - \mathbb{E}_{x_f \sim \mathbb{Q}} C(x_f) - 1)^2] + \mathbb{E}_{x_f \sim \mathbb{Q}}[(C(x_f) - \mathbb{E}_{x_r \sim \mathbb{P}} C(x_r) + 1)^2] \tag{3.2}$$

where $\mathbb{P}$ and $\mathbb{Q}$ are the distributions of real and generated data respectively; $C(x_r)$ and $C(x_f)$ are the probability of D for real and fake data.

The Cycle Consistency loss function is defined as follow:

$$\mathscr{L}_{cyc}(G_C, G_N) = \mathbb{E}_{n \sim p_{data}(n)}[||G_N(G_C(n)) - n||] + \mathbb{E}_{c \sim p_{data}(c)}[||G_C(G_N(c)) - c||] \tag{3.3}$$

where n and c correspond to domain images (n for NIR images and c for color images).

The Structural Similarity Index (SSIM) [161] has been used during training process, where the aim of using such loss function is to help the learning model to generate a visually improved image. The structural loss function defined as below:

$$\mathscr{L}_{SSIM} = \frac{1}{NM} \sum_{p=1}^{P} 1 - SSIM(p). \tag{3.4}$$

The Identity loss function employed to regularize the generator. The aim of using such loss function is if something already looks like from the target domain, should not map it into a different image.

$$\mathscr{L}_{identity}(G_C, G_N) = \mathbb{E}_{c \sim P_{data}(c)}[||G_C(c) - c||] + \mathbb{E}_{n \sim P_{data}(n)}[||G_N(n) - n||]. \tag{3.5}$$

The final objective $\mathscr{L}_{final}$ is obtained as below:

$$\mathscr{L}_{final} = \mathscr{L}_{RaLSGAN} + \lambda \mathscr{L}_{Cycle} + \mathscr{L}_{SSIM} + \gamma \mathscr{L}_{Identity} \tag{3.6}$$

where λ, γ are the weights to Cycle Consistency and Identity loss function, which play as regularization terms impacting on the optimization of the model. Assigning a bigger weights lead the model to have better reconstruction loss and model will make smaller changes. On the other hand, a smaller weights increase the risk of artifacts and lead the model to bring more dramatic changes with respect to input images.

3.2.3 Spectral Normalization

The performance control of the discriminator is an ongoing challenge in training Generative Adversarial Networks. The density ratio estimates by the discriminator

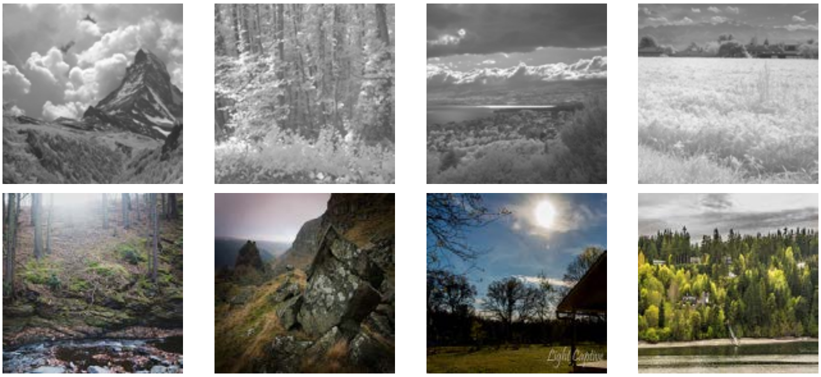


Figure 3.4: Unpaired set of images (256×256 pixels each) used for training the proposed approach and CycleGAN [195]; $(top-row)$ NIR images from [19]; $(bottom-row)$ RGB images collected from internet.

in high-dimensional spaces is often imprecise and unstable during learning phase, so generators do not learn the multimodel structure of the target. To solve the mentioned issue [118] proposed the normalization method. It helps to stabilize the training of discriminators by applying spectral normalization. Hence, discriminator becomes more stable and the network converge faster in less number of epochs. In the current work spectral normalization has been used so that the network learns the structure of images much better and generates better image quality comparing to baseline model [195].

3.2.4 Better Cycle Consistency

Cycle Consistency loss function is one of the main features of CycleGAN, which simply motivates generators to prevent needless changes and generates images that share structural similarity with inputs. Also, the Cycle Consistency helps a lot to make training phase stable in the early stages, but becomes a problem in later stages to generate realistic images. Since Cycle Consistency is a form of regularization we propose to progressively decrease the weight of cycle loss after half way of training process. Nevertheless, λ (in eq. (3.6)) needs to be checked to not become 0 in order to prevent the generators become unstable and unconstrained.

3.2.5 Two Time-Scale Update Rule

Training GANs, unlike of CNNs, needs more attention since mode collapse may occur in the learning process, when the generator generates a restricted variety of samples, or even the same sample, regardless of the input and prevent GAN to learn the target distribution. In [60] the authors propose the two time-scale update rule (TTUR), which improves the general performance, convergence speed and helps to prevent the mode collapse of GANs. TTUR has been applied to the proposed approach with ADAM stochastic optimizer to risk reduction of mode collaps and also to make sure that the discriminators converge in the training process.

3.3 Experimental Results

The proposed approach has been evaluated by using NIR images and their corresponding RGB, which were used as ground truth. The data set has been obtained from [19]; it contains pairs of NIR-RGB images of 1024×680 pixels each from different categories. It should be mentioned that dataset images are correctly registered and a pixel-to-pixel correspondence is guaranteed for quantitative and qualitative evaluation. Only categories with similar scenarios have been chosen for training the proposed model. The selected categories are as follow: country (50 pairs of images), field (51 pairs of images), forest (52 pairs of images) and mountain (50 pairs of images). The objective is to train the network in scenarios that contain similar objects.

The NIR images from the mentioned categories have been used during training while the corresponding RGB images (ground truth) have not been used neither during the training nor testing phases; they are only used for quantitative and qualitative evaluations. The RGB images used during the training process have been collected from internet (700 images); all the collected images correspond to scenarios similar to those from the aforementioned categories. Each pair of the original NIR and RGB images (from [19]) has been split up into two smaller images of 680×680 pixels each, resulting in a total of 406 pairs. From this set 68 pairs of images have been randomly selected and keep aside for evaluating the performance of the proposed approach. The rest of NIR images have been resized to 256×256 pixels, which was the size used to feed the network. All the RGB images collected from internet have been also resized up to 256×256 pixels each. In order to increase the number of images for training a data augmentation process has been applied (horizontal flipping and random crop). Figure 3.4 shows just four pairs of these unpaired (NIR-RGB) images used for training.

Results obtained with the proposed approach have been compared with results obtained using the baseline model presented in [195]. Quantitative and qualitative results from this NIR image colorization are presented in next sub sections. The proposed network has been trained using a 3.2 eight core processor with 62GB of memory with a NVIDIA GeForce GTX TITAN X GPU; on average the training process took near to 13 hours to complete 200 epochs. The model has been trained by using ADAM stochastic optimizer due to several advantages, slight memory requirements, it is computationally effective, also leads network to converge faster compared with the other stochastic optimizer and it prevents from overfitting. Dataset has been normalized from range of (0, 255) to (-1, 1); normal weights initialized with mean 0 and standard derivation 0.2 used in the proposed approach. The hyper-parameters were tuned during training stage as follows: learning rate 0.0003 and 0.0009 for generators and discriminators respectively; weight decay 1e-8 for generators, exponential decay rate 0.50, 0.999 for the first and second momentum (beta1, beta2); leak relu 0.2; cycle consistency weight 100; dropout with 0.5 probability.

3.3.1 Evaluation Metrics

In order to assess the performance of the proposed approach average Angular Error (AE) is considered. It is a widely used evaluation measure in color constancy research. AE is defined as the average angular distance between every obtained RGB pixel ($RGBo_{i,j}$) with the corresponding ground truth ($RGBg_{i,j}$). AE is used as an evaluation metrics since this measure is quite similar to the human spectator. AE is defined as:

$$AE = cos^{-1}\left(\frac{dot(RGB_o, RGB_g)}{norm(RGB_o) \times norm(RGB_g)}\right) \tag{3.7}$$

Additionally, Fréchet Inception Distance (FID) [60] has been used for comparing the similarity between obtained images and ground truth images in an embedded space. The FID is computed by using the Inception model up to a specific layer. Hence, in the case of two sets of multivariate Gaussians the FID between two distributions is obtained by calculating their means and covariances:

$$FID(X_g, X_o) = ||\mu_g - \mu_o||_2^2 + Tr(\Sigma_g + \Sigma_o - 2(\Sigma_g \Sigma_o)^{\frac{1}{2}}) \tag{3.8}$$

where the X_g is the set of real images (ground truth) and X_o is the set of obtained images. Lower FID results show the obtained images are more similar to ground truth images.

3.3.2 Quantitative Results

v Table 3.1 presents the quantitative results based on average AE and FID with the scenarios mentioned above (set of 68 pairs used as ground truth) for both models (proposed approach and baseline model [195]). According to the obtained average AE, the proposed approach improves CycleGAN in about 40%. In the case of FID metrics, the proposed approach gets an improvement of almost 39% with respect to CycleGAN. It can been seen that the proposed approach has smaller errors than the baseline model [195]. These results show that Relativistic loss and SSIM loss functions help to enhance the performance of the original CycleGAN [195]. Figure 3.5 depicts the box plot for the average AE of both approaches.

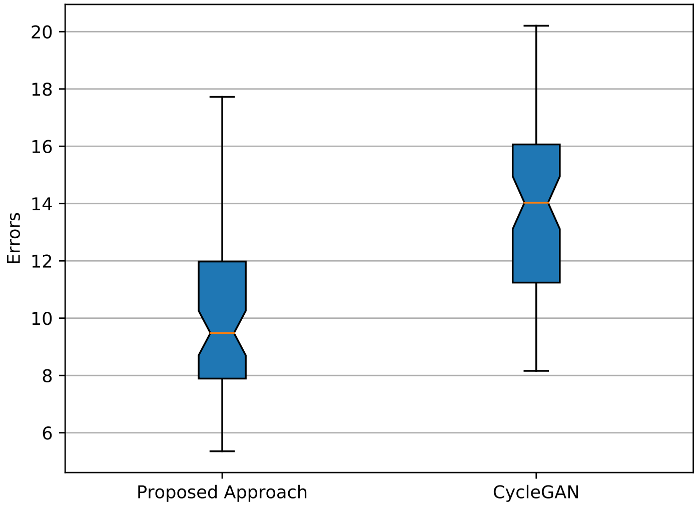


Figure 3.5: Avarage AE distribution for both approaches.

3.3.3 Qualitative Results

Figure 3.6 depicts some illustrative results for comparisons, both with respect to CycleGAN [195] and the corresponding ground truth. These images correspond to the set of 68 pairs of images mentioned above that have not been used neither dur-

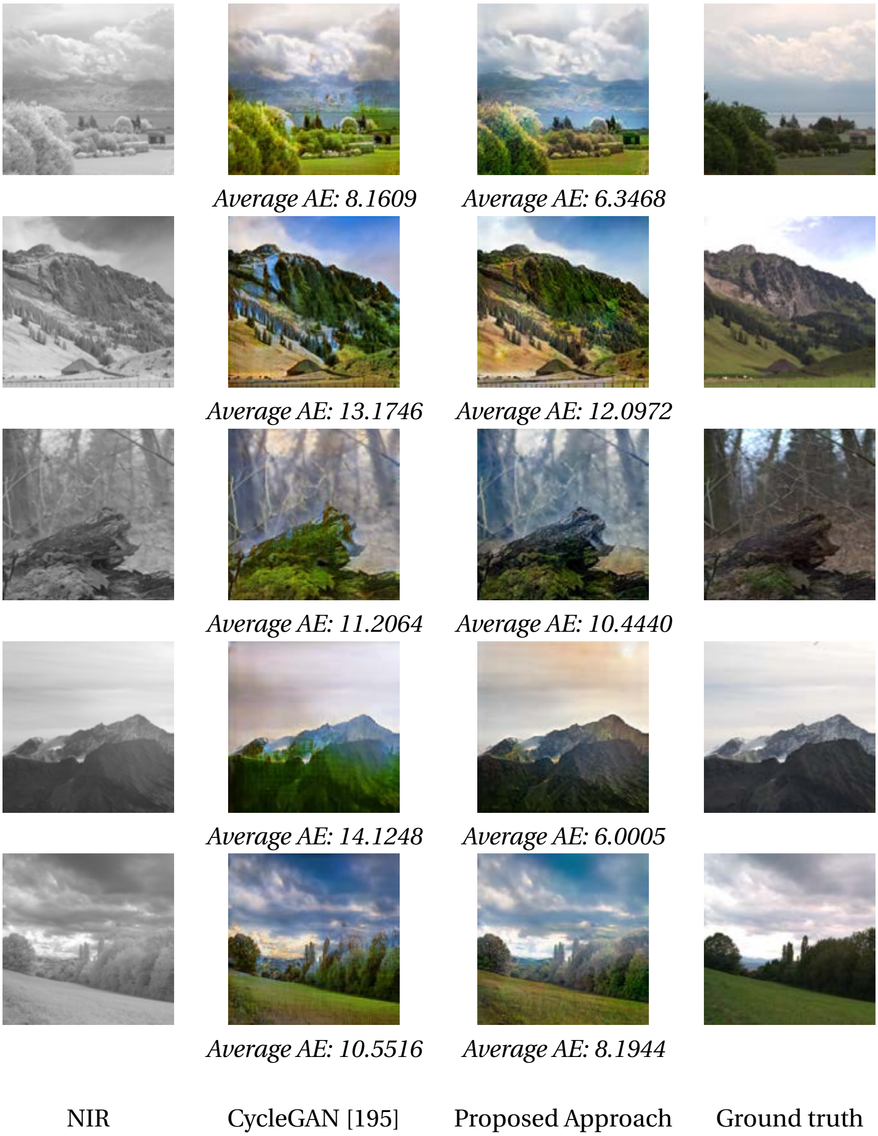


Figure 3.6: Colorized NIR images obtained with CycleGAN and with the proposed approach. RGB images (ground truth) are provided for qualitative evaluation. The numbers below the images show the average AE between the obtained colorized image and the ground truth.

	AE	FID
CycleGAN [195]	13.87	146.77
Prop. App.	**10.04**	**105.21**

Table 3.1: Comparative results between proposed approach and CycleGAN using evaluation metrics from Sec. 3.3.1.

ing the training nor during the validation stages. Each column shows the given NIR images, colorized with baseline model [195], colorized with the proposed approach and the ground truth respectively. It should be mentioned that all categories are trained simultaneously and also our colorized NIR images look quite better than the baseline model when compared with the ground truth.[1]

3.4 Summary

This paper proposes a novel architecture by using a Cycle-Consistent Adversarial Network in the context of colorization. The proposed approach address the challenging problem of colorizing NIR images when the ground truth is not available during the learning phase (i.e., in the unsupervised learning context) by using the appropriate generators and loss functions. Experimental results have shown that the NIR images colorized with proposed approach are visually better than those obtained with the CycleGAN baseline model as well as lower quantitative values are obtained.

[1] Additional results are provided at http://bit.ly/2VQG4B0

3.5 Supplementary Material

3.5.1 Additional Qualitative Results

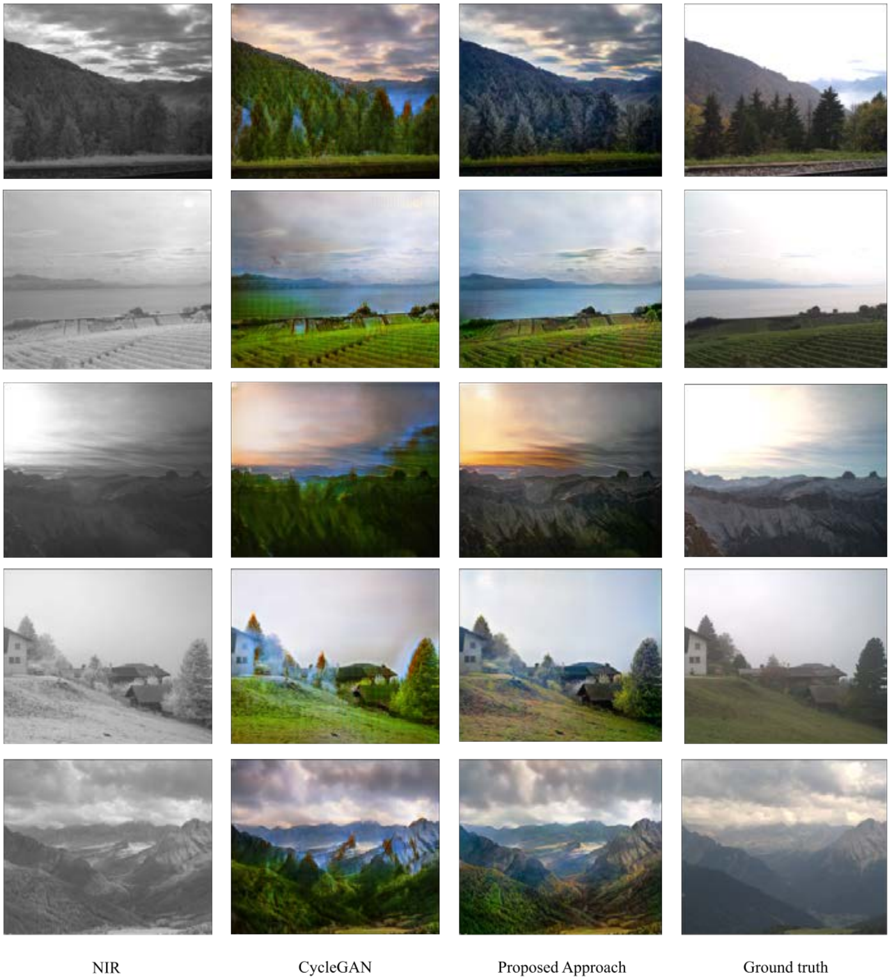


Figure 3.7: Additional visual results.

4 MPRNet: Multi-Path Residual Network for Lightweight Image Super Resolution

This chapter presents the article published at:

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, pp. 2704-2713. 2021.

Lightweight super resolution networks have extremely importance for real-world applications. In recent years several SR deep learning approaches with outstanding achievement have been introduced by sacrificing memory and computational cost. To overcome this problem, a novel lightweight super resolution network is proposed, which improves the SOTA performance in lightweight SR and performs roughly similar to computationally expensive networks. Multi-Path Residual Network designs with a set of Residual concatenation Blocks stacked with Adaptive Residual Blocks: (i) to adaptively extract informative features and learn more expressive spatial context information; (ii) to better leverage multi-level representations before up-sampling stage; and (iii) to allow an efficient information and gradient flow within the network. The proposed architecture also contains a new attention mechanism, Two-Fold Attention Module, to maximize the representation ability of the model. Extensive experiments show the superiority of our model against other SOTA SR approaches.

4.1 Motivation

Single Image Super Resolution (SISR) targets to recover a high-resolution (HR) image from its degraded low-resolution (LR) one with a high visual quality and enhanced details. SISR is still an active yet challenging topic to research due to its complex nature and high practical values in improving image details and textures. SR is also critical for many devices such as HD TVs, computer displays and portable

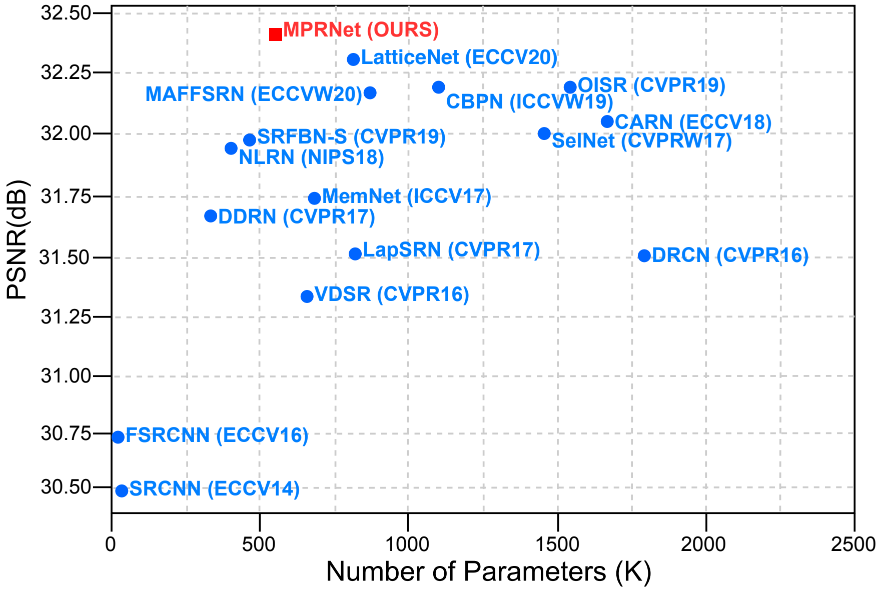


Figure 4.1: PSNR *vs*. Parameters trade-off on Set5 (×4). MPRNet achieves superior performance among all lightweight models.

devices like cameras, smartphones, tablets, just to mention a few. Moreover, it leads to improvements in various computer vision tasks, such as object detection [43], medical imaging [46], security and surveillance imaging [196], face recognition [119], astronomical images [105] and many other domains [98, 162, 176]. Image super-resolution is challenging due to the following reasons: *i*) SR is an ill-posed inverse problem, since instead of a single unique solution, there exist multiple solutions for the same low-resolution image; and *ii*) as the up-scaling factor increases, the complexity of the problem increases [35]. The retrieval of missing scene details becomes even more complicated with greater factors, which often leads to the reproduction of incorrect information.

Due to the rapid development of deep learning methods, recent years have witnessed an explosive spread of CNN models to perform SISR. The obtained performance has been consistently improved by designing new architectures or introducing new loss functions. Though significant advances have been made, most of the works in SR were dedicated to achieve higher PSNR with the design of a very deep network, which causes the increase in the numbers of computational operations. Besides that, most of the existing SISR methods are trained and evaluated on simulated datasets that assume simple and bicubic degradation. Unfortunately, SISR models trained on such simple datasets are hard to generalize for practical applications since degradations in the real-world are unknown.

In this paper, to design a practical network for real-world applications and tackle with mentioned downsides, a novel lightweight architecture is introduced, referred to as Multi-Path Residual Network (MPRNet), to adaptively learn most valuable features and construct the network to focus on learning high-frequency information. Additionally, to seek a better trade-off between performance and applicability, we introduce a novel module, referred to as Residual Module (RM), which contains Residual Concatenation Blocks that are connected to each other with a Global Residual connection; build with a set of Adaptive Residual Blocks (ARB) with a Local Residual Connection (LRC). Each ARB is defined as a divers residual pathways learning to make use of all kind of information form LR image space, which the main parts of the network can access to more rich information. So, our MPRNet design has the benefits of multi-level learning connections and also takes advantage of propagating information throughout the network. As a result, each block has access to information of the precedent block via local and global residual connections and passes on information that needs to be preserved. By concatenating different blocks followed by 1 × 1 convolutional layer the network can reach to both intermediate and high-frequency information, resulting in a better image reconstruction. Finally, in order to enhance the representation of the model and even make it robust against challenging datasets and noise, we propose a lightweight and efficient attention mechanism, Two-Fold Attention Mechanism (TFAM). TFAM is working by considering both the inner channel and spatial information to highlight the important information. This TFAM helps to adaptively preserve essential information and overpower the useless ones. The proposed model is illustrated in Figure 6.2. In brief, the main contributions are in three-fold:

- An efficient Adaptive Residual Block (ARB) is proposed by well-focusing on spatial information via a multi-path residual learning to enhance the performance at a negligible computational cost. Comprehensive study shows the excellent performance of ARB.
- A new attention mechanism (TFAM) is proposed to adaptively re-scale feature maps in order to maximize the representation power of the network. Since its low-cost, it can be easily applied to other networks, and has the better performance than other Attention Mechanisms.
- A lightweight network (MPRNet) is proposed to effectively enhance the performance via multi-level representation and multiple learning connections. The

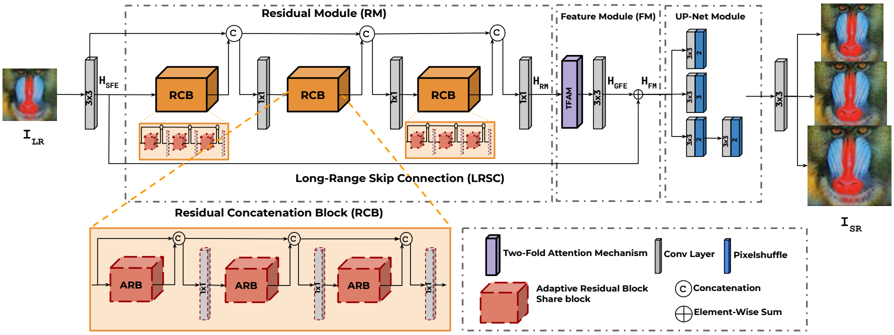


Figure 4.2: The overall network architecture of the proposed Multi-Path Residual Network (MPRNet).

MPRNet is built by fusing the proposed ARB with the robust TFAM to generate more accurate SR image. MPRNet achieves the excellent performance among all the lightweight state-of-the-art approaches with lower model size and computational cost (Figure 6.1).

4.2 Multi-Path Residual Network

4.2.1 Network Structure

The proposed model (MPRNet – Figure 6.2) consists of four different modules, namely, Shallow Feature Extraction (SFE); Residual Module that contains Residual Concatenation Blocks (RCBs); Feature Module that includes a Two-Fold Attention Mechanism (TFAM) and a Global Feature Extractor with a Long-Range Skip Connection; and the multi-scale UP-Net module at the end of network. Let's consider $\{I_{LR}, I_{SR}\}$ as the input and output of the network respectively. The SFE is a Conv layer with a kernel size of 3×3, which can be formulated as follow:

$$\boldsymbol{H}_{SFE} = f_{SFE}(\boldsymbol{I}_{LR}; W_c), \tag{4.1}$$

where $f_{SFE}(\cdot)$ and W_c indicates Conv operation and parameters applied on I_{LR}. $\boldsymbol{H}_{SFE}$ denotes the output of SFE, which later is used as the input to Residual Module. Lets $\boldsymbol{H}_{RM}^{i,j}$ be the output from the i-th Residual Concatenation Block (RCB) that has j-th inner Adaptive Residual Blocks (ARBs). The Residual Module can be defined as:

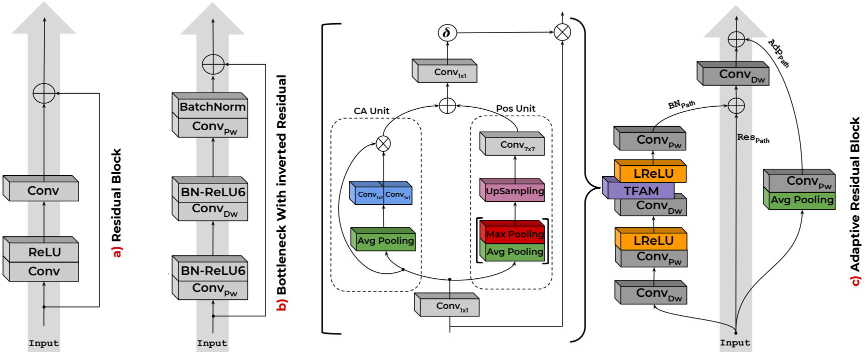


Figure 4.3: Illustrations of different structure of residual blocks: a) Residual block in EDSR [95]; b) Bottleneck with inverted residual from [62]; c) Proposed Adaptive Residual Block and Two-Fold Attention Module.

$$\boldsymbol{H}_{RM} = f([\boldsymbol{H}_{SFE}, ..., \boldsymbol{H}^{i-1}_{RCB}(\boldsymbol{H}^{j-1,R}_{ARB}; W^{j}_{c}), \boldsymbol{H}^{i}_{RCB}]; W^{i}_{c}), \tag{4.2}$$

where $\boldsymbol{H}_{RM}$ is the output of the Residual Module. Note that our RM contains multi-level learning connections followed by a 1 × 1 Conv layer to control the output after each block, which helps our model to quickly propagate information all over the network (lower to higher layers and vice-versa in term of back propagation) and also let the network to learn multi-level representations. So, i-th RCB can be defined as:

$$\boldsymbol{H}^{i}_{RCB} = f([\boldsymbol{H}^{j,R}_{ARB}, ..., \boldsymbol{H}^{j-1,R}_{ARB}(\boldsymbol{H}^{i-1}; W^{i}_{c})]; W^{j}_{c}). \tag{4.3}$$

Then, the output of RM feed to the Feature Module by firstly refining the feature maps (i.e., re-calibrate) throughout the TFAM and then extracting more abstract features. Later, accumulate with LRSC to efficiently alleviate the gradient vanishing/exploding problems and make sure that network has access to unmodified information before UP-Net:

$$\boldsymbol{H}_{FM} = f_{GFE}(\boldsymbol{H}_{TFAM}(\boldsymbol{H}_{RM}; W_c); W_c) + \boldsymbol{H}_{LRC}, \tag{4.4}$$

where $\boldsymbol{H}_{TFAM}$ denotes our TFAM and $\boldsymbol{H}_{LRC}$ is Long-Range Residual Connection. The last stage is the Multi-Scale Up-Net Module to reconstruct the image from obtained feature-maps. The upsampling module is inspired by [1] and followed by a Conv layer:

$$\boldsymbol{H}_{UP} = f_{pix}^{\uparrow}(\boldsymbol{H}_{FM}), \tag{4.5}$$

where $f_{pix}^{\uparrow}(\cdot)$ indicates the Up-net module function and $\boldsymbol{H}_{FM}$ is the output of FM. The upsampled features are reconstructed with a Conv layer:

$$\boldsymbol{I}_{SR} = f_{REC}(\boldsymbol{H}_{up}) = \boldsymbol{H}_{MPRNet}(\boldsymbol{I}_{LR}), \tag{4.6}$$

where $f_{REC}(\cdot)$ and $\boldsymbol{H}_{MPRNet}(\cdot)$ denote the reconstruction layer and function of our MPRNet. In the next subsections, more details about the Adaptive Residual Block and Two Fold Attention Mechanism are given.

4.2.2 Adaptive Residual Block

This research focuses on designing a efficient and effective Residual Block based on Depthwise (Dw) and Pointwise (Pw) Convolutions for SISR. [134] introduced linear bottleneck with an inverted residual structure. However, this structure deliver chances of losing information and weaken the propagation capability of gradients across layers, due to gradient confusion arising from the narrowed feature space [32, 88]. Thus, we propose a novel Residual Block that mitigates the aforementioned issues; it is well-optimized especially for the SR tasks, called Adaptive Residual Block (ARB). Unlike [134], ARB introduces new features and operations by proposing a multi learning pathways with a completely new structure. Each learning path is responsible to extract different kind of information before aggregation. So, the main part of network can have access to more rich information and performs notably well in noisy LR and generates more accurate SR image. The ARB consists of three different learning pathways that are detailed below. Figure 4.3 shows each of the ARB components.

Bottleneck Path: We design our Bottleneck path (BN) based on the following insights: i) Extract richer spatial information since spatial information is key importance in SR tasks; ii) prevent very wide feature maps in the middle of the building block, which unavoidably growing the computational load of relevant layers; iii) preserve the BN path low-cost and efficient. Thus, Dw Convolutions with small kernel size (3×3) are chosen since they are lightweight and they can learn expressive features when conducted to the high dimensional space. So, we initiate the BN path by using a Dw convolution with kernel size 3×3 towards the high dimensional features space to richer spatial information to be encoded and generate meaningful representations. Also, a Pw convolution is used after each Dw convolution in our design to produce new features by encoding the inter-channel information and reduce the computational cost. We shared the same number of channels and resolution along the BN path to prevent of sudden rise of computational burden in middle of the path. Furthermore, we conjunct our TFAM into the BN path after the second Dw convolution to spotlight the informative features along the channel and spatial axes. By doing so, the BN path is working with high dimensional features space, which makes the pathway efficient, low-cost, and well-focused on spatial context information compared to [134].

Adaptive Path: It is proposed by taking the advantages of global average pooling accompanied by a 1×1 Pw convolution. Average Pooling layers have been employed to take the average value of the features from the feature space to smooth and eliminate the noise from the LR image and reduce the dimensionality of each feature map but retains the important information to help the network to generate robust feature maps in challenging situations—noisy LR image. So, the network can generate a sharper and well-detailed SR image.

Residual Path: Unlike [134] that puts the residual path between narrowed feature space that cause gradient confusion, in our ARB, we place the residual path on the high dimensional representations to transfer more information from the bottom- to top-layers. Such structure facilitate the gradient propagation between multiple layers and help the network to optimize better during training.

Thus, the information from BN- and Res-paths aggregate together, followed by another Dw convolution. We found out adding the Dw convolution before final aggregation with Adp path is essential for performance improvement since Dw encourage the network to learn more meaningful spatial information. Extensive experiments show that, our ARB is more beneficial than the existed ones for SISR tasks and improved the results with a large margin.

4.2.3 Two-Fold Attention Module

A novel Attention Mechanism (TFAM) has been proposed to boost the performance of our Adaptive Residual Block and refine the high-level information in the Feature Module (FM) by focusing on both channel and spatial information. The best way to amplify efficiency of ARB is through the union of the channel and spatial attention mechanism, since the residual features need to be well-focused on both information. In detail, TFAM is designed to focus on the important features on the channel information via CA unit and spotlight on the region of interest via Pos unit. Thus, each unit can learn 'what' and 'where' to attend in the channel and spatial axes respectively to recover edges and textures more accurately. As a result, TFAM works better than other attention mechanism [63, 64, 100, 167] by emphasising informative features and reducing worthless ones.

Channel Unit. CA unit starts with an average pooling to exploit first-order

Table 4.1: Comparison with lightweight SOTA methods on the Bicubic (**BI**) degradation for scale factors [×2, ×3, ×4]. **Red** is the Best and **Blue** is the second best performance. We assume that the generated SR image is 720P to calculate Multi-Adds (MAC).

	Methods	**VDSR** [78]	**LapSRN**[84]	**MemNet**[147]	**NLRN**[99]	**SRFBN_S**[92]	**CARN**[1]	**CBPN** [194]	**OISR_RK2_s**[57]	**MAFFSRN-L**[120]	**LatticeNet**[108]	**MPRNet [Ours]**
Params-MAC	×4	$655K-612.6G$	$813K-149.6G$	$677K-2662.4G$	$350K-32.5$	$483K-119G$	$1592K-90.9G$	$1197K-97.9G$	$1540K-114.2G$	$830K-38.6G$	777K-43.6G	**538K-31.3G**
Dataset	Scale	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM
Set5	×2	37.53/0.9587	37.52/0.9590	37.87/0.9597	38.00/0.9603	37.78/0.9597	37.76/0.9590	37.90/0.9590	37.90/0.9600	38.07/0.9607	**38.15/0.9610**	38.08/0.9608
	×3	33.66/0.9213	———	34.09/0.9248	34.27/0.9266	34.20/0.9255	34.29/0.9255	———	34.39/0.9273	34.45/0.9277	34.53/0.9281	**34.57/0.9285**
	×4	31.35/0.8838	31.54/0.8850	31.74/0.8893	31.92/0.8916	31.98/0.9594	32.13/0.8937	32.21/0.8944	32.21/0.8903	32.20/0.8953	32.30/0.8962	**32.38/0.8969**
Set14	×2	33.03/09124	33.08/0.9130	33.28/0.9142	33.46/0.9159	33.35/0.9156	33.52/0.9166	33.60/0.9171	33.58/0.9172	33.59/0.9177	33.78/0.9193	**33.79/0.9196**
	×3	29.77/0.8314	———	30.00/0.8350	30.16/0.8374	30.10/0.8350	30.29/0.8407	———	30.33/0.8420	30.40/0.8432	30.39/0.8424	**30.42/0.8441**
	×4	28.01/0.7674	28.19/0.7720	28.26/0.7723	28.36/0.7745	28.45/0.7779	28.60/0.7806	28.63/0.7813	28.63/0.7822	28.62/0.7822	28.68/0.7830	**28.69/0.7841**
B100	×2	31.90/0.8960	31.80/0.8950	32.08/0.8978	32.19/0.8992	32.00/0.8970	32.09/0.8978	32.17/0.8989	32.18/0.8996	32.23/**0.9005**	32.25/**0.9005**	**32.25**/0.9004
	×3	28.82/0.7976	———	38.96/0.8001	29.06/0.8026	28.96/0.8010	29.06/0.8434	———	29.10/0.8083	29.13/0.8061	29.15/0.8059	**29.17/0.8073**
	×4	27.29/0.7251	27.32/0.7280	27.40/0.7281	27.48/0.7306	27.44/0.7313	27.58/0.7349	27.58/0.7356	27.58/0.7364	27.59/0.7370	27.62/0.7367	**27.63/0.7385**
Urban100	×2	30.76/0.9140	30.41/0.9100	31.31/0.9195	31.81/0.9249	31.41/0.9207	31.92/0.9256	32.14/0.9279	32.21/0.8950	32.38/0.9308	32.43/0.9302	**32.52/0.9317**
	×3	27.14/0.8279	———	27.56/0.8376	27.93/0.8453	27.66/0.8415	28.06/0.8493	———	28.03/0.8544	28.26/0.8552	28.33/0.8538	**28.42/0.8578**
	×4	25.18/0.7524	25.21/0.7560	25.50/0.7630	25.79/0.7729	25.71/0.7719	26.07/0.7837	26.14/0.7869	26.14/0.7874	26.16/0.7887	26.25/0.7873	**26.31/0.7921**

Table 4.2: Comparison with SOTA methods on challenging datasets ("**BD**" and "**DN**") for scale factor ×3. **Red** is the Best and **Blue** is the second best performance.

Dataset	**Methods**	**Bicubic**	**SPMSR**[125]	**SRCNN**[35]	**FSRCNN**[37]	**VDSR**[78]	**IRCNN_G**[185]	**IRCNN_C**[185]	**SRMD(NF)**[154]	**RDN**[192]	**MPRNet [Ours]**
	Degradation	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM
SPSet5	**BD**	28.34/0.8161	32.21/0.9001	31.75/0.8899	26.58/0.8224	33.29/0.9139	33.38/0.9182	29.55/0.8246	34.09/0.9242	**34.57/0.9280**	**34.57**/0.9278
	DN	24.14/0.5445	—-	27.04/0.7638	24.28/0.7124	27.42/0.7372	24.85/0.7205	26.18/0.7430	27.74/0.8026	28.46/0.8151	**28.54/0.8175**
Set14	**BD**	26.12/0.7106	28.89/0.8105	28.64/0.7997	24.86/0.7246	29.58/0.8259	29.73/0.8292	27.33/0.7135	30.11/0.8364	**30.53/0.8447**	30.47/0.8427
	DN	23.14/0.4828	—-	25.56/0.6592	23.25/0.5956	25.60/0.6706	23.84/0.6091	24.68/0.6300	26.13/0.6974	**26.60/0.7101**	26.25/0.6954
B100	**BD**	26.02/0.6733	28.13/0.7740	27.33/0.7500	24.15/0.6728	28.61/0.7900	28.65/0.7922	26.46/0.6572	28.98/0.8009	**29.23/0.8079**	29.19/0.8062
	DN	22.94/0.4461	—-	25.45/0.6198	23.95/0.5695	25.22/0.6271	23.89/0.5688	24.52/0.5850	25.64/0.6495	25.93/0.6573	**25.95/0.6616**
Urban100	**BD**	23.20/0.6661	25.84/0.7856	25.19/0.7591	22.95/0.6836	26.68/0.8019	26.77/0.8154	24.89/0.7172	27.50/0.8370	**28.46/0.8581**	28.31/0.8538
	DN	21.63/0.4701	—-	23.59/0.6580	21.74/0.5724	23.33/0.6579	21.96/0.6018	22.63/0.6205	24.28/0.7092	24.92/0.7362	**25.00/0.7406**

statistics of features followed by two Conv layer, which they work side by side, each seeing half of the input channels, and producing half the output channels, and both subsequently concatenated to even have more low-cost unit. Thus, CA unit modulates features globally, where the summary statistics per channel are computed. Then, used to emphasize meaningful feature maps while redundant useless features are diminished. Especially, CA unit focuses on 'what' is meaningful given an input image.

Positional Unit. Pos unit designed as a complementary unit to our CA unit. The feature map information is varied over spatial positions therefore, Pos unit concerns about the position of the informative part of the image and focuses on that region. Pos unit requires a large receptive field to work perfectly in SR tasks unlike the classification task. Thus, Average- and Max pooling operations with a large kernel size have been employed and then concatenated them to generate an efficient feature descriptor. afterward, an UpSampling layer is used to retrieve the spatial dimensions, which is followed by a Conv layer to generate a spatial attention map.

Finally, highlighted information from both units aggregated together followed a 1×1 Conv layer and a sigmoid operation to firstly, recover the channel dimensions and then generate the final attention mask. Also, a residual connection used to transfer HR features to the end of module.

4.3 Experimental Results

4.3.1 Setting

Datasets & Evaluation Protocol. Following previous works [31, 100], we use $DIV2K$ [153] dataset to train (800 images) and validate (100 images) our model. The proposed model is evaluated with the standard benchmark datasets, namely, $Set5$ [15], $Set14$ [183], $B100$ [114], and $Urban100$ [66]. Two widely used quantitative metrics have been considered to measure its performance: PSNR and SSIM [161], computed between the obtained images and the corresponding ground truths. Both metrics are computed on the Y channel in the $YCbCr$ space.

Degradation Models. Following the work of [192], three different degradation models created to simulate LR images and make fair comparisons with available methods. Firstly, a bicubic (BI) down-sampling dataset with scaling factors [×2, ×3, ×4] has been created. Blur-Down-sampled (BD) is the second one to blur and down-sample HR images with a Gaussian kernel 7×7, and $\sigma = 1.6$. Then, images are down-sampled with scaling factor ×3. Aside from the BD, a more challenging model has been created, referred to as (DN). DN degradation model is down-sampling HR images with bicubic followed by adding 30% Gaussian noise.

Training Details. In the training stage, RGB input patches are used with size of 64×64 from each of the randomly selected 64 LR training images. Patches are augmented by random horizontally flips and 90 degree rotation. AdamP [59] optimizer has been employed. The initial learning rate set to 10^{-3} and its halved every 4×10^{5} steps. $L1$ is used as loss function to optimize the model. The PyTorch framework is used.

4.3.2 Comparison with state-of-the-art Methods

Results with BI Degradation

Table 6.3 presents comparisons between the proposed MPRNet and 10 most recent lightweight SOTA models on BI degradation model for scale factor [×2, ×3, and ×4] to verify the effectiveness of our MPRNet (we exclude some lightweight methods [35, 37, 79, 137, 146, 162] from table 6.3 since their results are worse than MemNet). Table 6.3 also contains the number of parameters and operations to show the model complexity. In almost all the cases, our MPRNet achieves superior results among all the aforementioned approaches. MPRNet performs especially well on Urban100. This is particularly because the Urban100 includes rich structured contents and our model can consistently accumulate these hierarchical features to form of more representative features and well-focused on spatial context information. This characteristic can be confirmed by our MPRNet SSIM scores, which focuses on the visible structures in the image. In Figure 4.4 a couple of qualitative results on scale

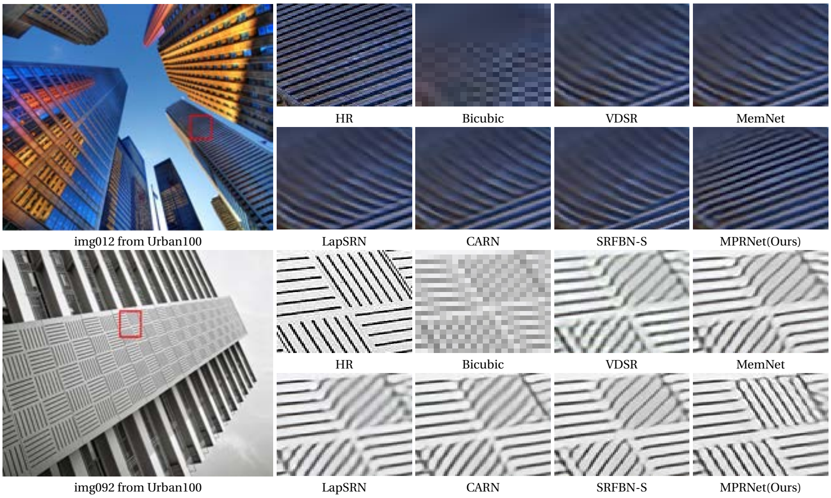


Figure 4.4: Qualitative results on **BI** degradation dataset with scale factor ×4.

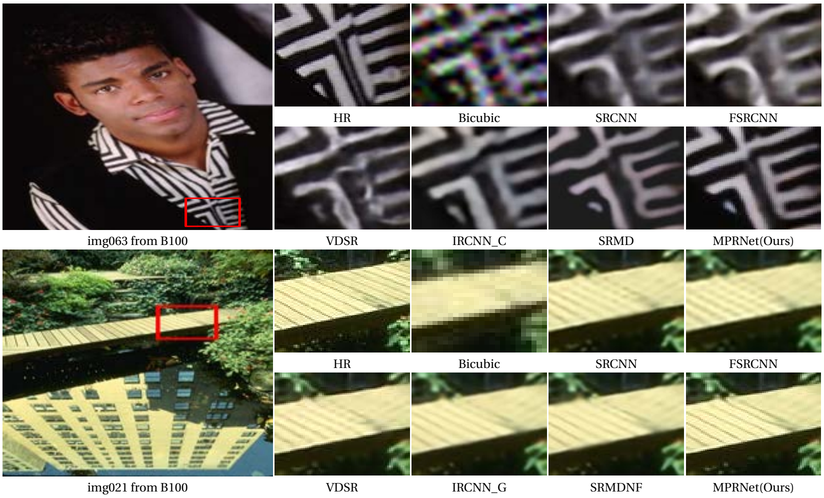


Figure 4.5: Qualitative results on **DN** and **BD** degradation datasets with a scale factor ×3.

Table 4.3: Effect of Attention Mechanisms and proposed Adaptive Residual Block on SOTA models. The best **PSNR** (dB) are highlighted.

Name	EDSR					RCAN					MSRN				
	Baseline	**Attention Modules**			**ResBlock**	**Baseline**	**Attention Modules**			**ResBlock**	**Baseline**	**Attention Modules**			**ResBlock**
Channel and spatial attention residual[64]		✓					✓					✓			
Enhanced Spatial Attention[100]			✓					✓					✓		
Two-Fold Attention Module[Ours]				✓	✓				✓	✓				✓	✓
Adaptive Residual Block[Ours]					✓					✓					✓
PSNR on Set5 (×4)	32.46	32.48	32.51	32.54	**32.65**	32.63	32.64	32.67	32.70	**32.78**	32.25	32.27	32.30	32.34	**32.39**
PSNR on Urban100 (×4)	26.64	26.66	26.69	26.71	**26.79**	26.82	39.84	26.86	26.89	**26.96**	26.22	26.25	26.29	26.32	**26.41**

factor ×4 are depicted. The proposed MPRNet can generally yield to more precise details. In both images in Figure 4.4, the texture direction of the reconstructed images from all compared methods is completely wrong. However, results from the proposed MPRNet makes full use of the abstract features and recover images accurately similar to ground truth texture.

Results with BD and DN Degradation Models

In Table 4.2, the performance of MPRNet on BD and DN benchmark datasets, together with SOTA methods, are presented. Due to degradation mismatch, SRCNN, FSRCNN, and VDSR for both BD and DN have been re-trained. As can be appreciated, MPRNet achieves remarkable results over all the lightweight SOTA models on challenging benchmark datasets. RDN [192] also listed as a high-capability model to show the superior performance of MPRNet compared to very costly model in the BD and DN datasets. RDN performs sightly better in some BD datasets but not in DN datasets. Obviously, this result was expected since RDN is very expensive compared to low-cost MPRNet (it is almost ×44 more costly). Figure 4.5 depicts some visual results on both challenging BD and DN benchmark datasets. As can be appreciated the MPRNet with the help of the proposed TFAM performs better in comparison with SOTA methods in terms of producing more convincing results by cleaning off noise and blurred regions from SR images, which results in a sharper SR image with fine details.[1]

4.3.3 Ablation Study

To further investigate the performance of the proposed model, a deep analysis on the Two-Fold Attention Module, the Adaptive Residual Block, and Residual Learning Connections is performed via an extensive ablation study.

[1] Additional analyses (such as Inference time, Memory consumption, and etc.) and more visual results can be found in supplementary material.

Table 4.4: Impact of different Attention Mechanisms on MPRNet.

Dataset	**Baseline**	**SE**	**CBAM**	**CSAR**	**ESA**	**TFAM**
Set14 (×4)	28.57	28.59	28.54	28.61	28.64	**28.67**
Urban100 (×4)	26.19	26.21	26.18	26.23	26.25	**26.29**

Table 4.5: Effect of different configs of Residual Block and each learning pathway of the Adaptive Residual Block

Configs	**MobileNet BnBlock**	**EDSR ResBlock**	**RCAN ResBlock**	**ARB$_B$**	**ARB$_{BA}$**	**ARB$_R$**	**ARB**
BN$_p$				✓	✓	✓	✓
Adp$_p$					✓		✓
Res$_p$						✓	✓
B100 (×4)	27.24	27.44	27.52	27.46	27.58	27.55	**27.63**
Urban100 (×4)	25.79	25.96	26.08	26.05	26.15	26.11	**26.31**

Two-Fold Attention Module. In this section, Deep investigation of the impacts of our proposed TFAM on SOTA SR models are provided. The performance of image SR has improved greatly with the application of Attention Mechanism (AM). Table 4.3 shows the performance of applying recent AMs including Channel and spatial attention residual (CSAR) [64], Enhanced Spatial Attention (ESA)[100], and our Two-Fold Attention Module (TFAM) on EDSR, RCAN, and MSRN. For a fair comparison, all the models were re-trained with their default setting and AMs are added to the end of their Block, and replaced in the same place as RCAN's Channel

Table 4.6: Study on combining different Residual Connections.

Options		**Baseline**	**1st**	**2nd**	**3rd**	**4th**	**5th**
Residual Learning Connections	**LRC**	✗	✓			✓	✓
	GRC	✗		✓		✓	✓
	LRSC	✗			✓		✓
PSNR on Set5 (×3)		34.42	34.40	34.47	34.45	34.52	**34.57**
PSNR on Urban100 (×3)		28.30	28.29	28.35	28.33	28.38	**28.42**

Attention placed. As can be seen, by using the aforementioned attention module, the performance of the baseline models are increased that shows the importance of AM in SR tasks. By applying the CSAR to the mentioned approaches, PSNR improves in EDSR and MSRN but does not show enough improvement in RCAN. In contrast, ESA is enhanced version of CASR, which combine both the channel and spatial information, improves all the baseline models. However ESA cannot completely boost the power of the networks due to lack of highlighting informative feature in spatial information. For this propose, we introduce Two-Fold Attention Module, which consider both channel and spatial information and maximize the performance of the networks. TFAM extracts the channel and spatial statistic among channels and spatial axis to further enhance the discriminative ability of the network. As a results, TFAM shows better performance than all the aforementioned ones and boosted the baseline SOTA.

Furthermore, Table 4.4 contains the study on impact of recent AMs on our MPRNet. Namely, SE[63], CBAM [167], CSAR [64], ESA [100], and TFAM. We apply all the aforementioned AM to our ARB blocks and Feature Module, and provide the performance. The proposed MPRNet with CBAM, could not achieve better results than baseline or SE due to losing channel information and applying the Max-pooling in CA unit which shows harm the performance. Unlike, MPRNet with CASR achieves better results than CBAM and SE because of considering both channel and spatial information but not better than ESA. However, our TFAM performs better among all the AMs by calculating the first order statistics on CA unit and applying Avg- and Max-pooling operations along the channel axis, which is effective in highlighting informative regions and extracts the most important features like edges.

Table 4.3 also shows the efficiency of our ARB with conjunction of TFAM when it is applied to other SOTA models. As indicated, ARB with TFAM together can improve the PSNR of SOTA models with a large margin.

Adaptive Residual Block. Table 4.5 presents the impact of different Residual Blocks and the proposed Adaptive Residual Block (ARB) on our MPRNet. In this work, three different structures of residual blocks from SOTA models are considered to compare with our proposed ARB, namely, MobileNet-BottleneckBlock, EDSR-ResBlock, RCAN-ResidualChannelBlock. All the models were trained with the same settings. As can be seen, MobileNet-BottleneckBlock could not perform well in SR tasks due to difficulty of extracting high-frequency information and gradient confusion. EDSR-ResBlock is the ResNet without batch normalization layer, but still could not achieve good results due to the lack of extracting rich feature maps and eliminating noises from LR feature space. RCAN-ResidualChannelBlock performs better than aforementioned ResBlock due to channel attention in their structure. However RCAN-ResidualChannelBlock did not show better results than our proposed ARB since our ARB can learn more expressive spatial information, have access to high-dimensional information and also with the help of TFAM can maximize the whole performance of block.

Additionally, effect of each learning pathways of ARB on the performance is provided. ARB_B, ARB_{BA} and ARB_R are Adaptive Residual Block with bottleneck path; ARB with bottleneck and adaptive paths; ARB with bottleneck and residual path respectively. As shown in Table 4.5, MPRNet with all learning pathways (ARB) achieves the best performance among all the mentioned ResBlock and combinations of different ARB learning pathways. This is caused by allowing the main parts of network to focus on more informative components of the LR features and force the network to focus more on abstract features, which are important in SR tasks. Furthermore, the proposed pathways helps the model to converge better and performs better than all the baseline models. In a nutshell, information propagates locally via residual path, adaptively extract the informative features via adaptive path, and learn more meaningful spatial information by Bottleneck path. By doing so, information is transmitted by multiple pathways inside of ARB and main parts of network access to more expressive and richer feature maps, resulting in superior PSNR.

Effect of Residual Learning Connections. Table 4.6 shows the extensive study of the impact of Residual Learning Connections on our design of MPRNet, i.e. Local Residual Connection (LRC), Global Residual Connection (GRC), and Long Range Skip Connection (LRSC). In this work, residual connections except LRSC comprise concatenation followed by a 1×1 Conv layer. As we can see, MPRNet without any residual connection performs relatively low (i.e. baseline). However, MPRNet with only GRC in Residual Module shows better performance than baseline since GRC transports the information from mid- to high-layers and helps the model to better leverage multi-level representations by collecting all information before the next module.

On the contrary, MPRNet with only LRC inside Residual Concatenation Block could not perform better than the MPRNet with GRC. This behavior was expected as mentioned in [56] that 1×1 Conv layer on the residual connection can confuse optimization and prevent information propagation due to multiplicative manipulations. However, MPRNet can show better performance by using both connections ($4th$ col.). This is due to GRC eases the information propagation issue that LRC suffers from.

To end this, LRSC also added to the MPRNet to carry the shallow information to high-level layers. Thus, information is transferred by multiple connections, which mitigates the vanishing gradient problem and network has access to multi-level representation. As a results, MPRNet with all connections ($5th$ col.) can performs greatly better.

Model Complexity Analysis. Figure 6.1 indicates the comparison regard to the model size and PSNR with 15 recent state-of-the-art SR models. Our MPRNet achieves the best performance among all the lightweight SR approaches with much fewer parameters and achieves better or comparable results when compared with computationally expansive models. This shows that our MPRNet is well-balanced in terms of model size and reconstruction results.

4.4 Summary

This paper proposes a novel lightweight network (MPRNet) that achieves the best performance against all existing lightweight SOTA approaches. The main idea behind of this work is to design an advanced lightweight network to deliver almost similar results to heavy computational networks. A novel Residual Module is proposed to let abundant low-level information to be avoided through multiple connections. In addition, an efficient Adaptive Residual Block is proposed to allows MPRNet achieves more rich feature-maps through the multi-path learning. Furthermore, to maximize the power of the network a Two-Fold Attention Module is proposed, which refine the extracted information along channel and spatial axes to further enhance the discriminative ability of the network. Extensive evaluations and comparisons are provided.

4.5 Supplementary Material

The following items are contained in the supplementary material:

1) Memory Complexity Analysis
2) Inference Time and Memory Consumption
3) Width Multiplier
4) Additional Qualitative Results

4.5.1 Memory Complexity Analysis

In this section, we compare the proposed MPRNet with the most recent lightweight and expensive networks: LapSRN, VDSR, DRCN, SelNet, DRRN, MemNet, SRFBN, CARN, MSRN, OISR, CBPN, MAFFSRN, and LatticeNet in term of number of MAC operations (Multi-Adds) and reconstruction results (PSNR) to show the efficiency of the purposed MPRNet. In Figure 4.6, reconstruction results (PSNR) and MAC (G), which shows the number of multiply-accumulate operations, are illustrated. As we can see, our MPRNet can achieve better results with a large gap among all the recent networks with less needed MAC operations; and even perform better than

MSRN, which has more than 160 layers by only 13% of the total number of MSRN multiply-accumulate operations (1365.4G).

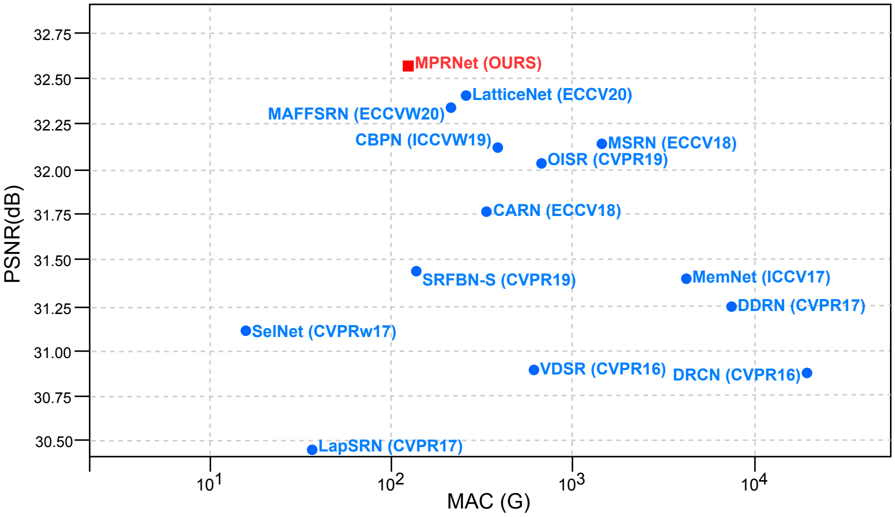


Figure 4.6: PSNR $vs.$ MAC on Urban100 for scale factor ×2.

4.5.2 Inference Time and Memory Consumption

Table 4.7 illustrates the superiority of the proposed MPRNet in terms of Inference Time (s) and Memory Consumption (MB) when it compares with the recent light- and heavy-weight state of the art approaches on Urban100 for scale factor ×4, namely MemNet, SRFBN, CARN, RCAN, RDN, EDSR. We consider the pyTorch version of MemNet instead of Caffe version due to large memory consumption in Caffe. The inference time and memory consumption of each approach is evaluated using their official code on the same environment. The MPRNet has the fastest inference time while using less memory compared to other approaches, which reflect the efficiency of the proposed method.

4.5.3 Depth Multiplier

In Table 4.8, the effect of depth multiplier on model size and reconstruction results are illustrated. Similar to MobileNetV2, we employed depth multiplier ($alpha$) to make our MPRNet even more light cost with small reduction in performance. Depth

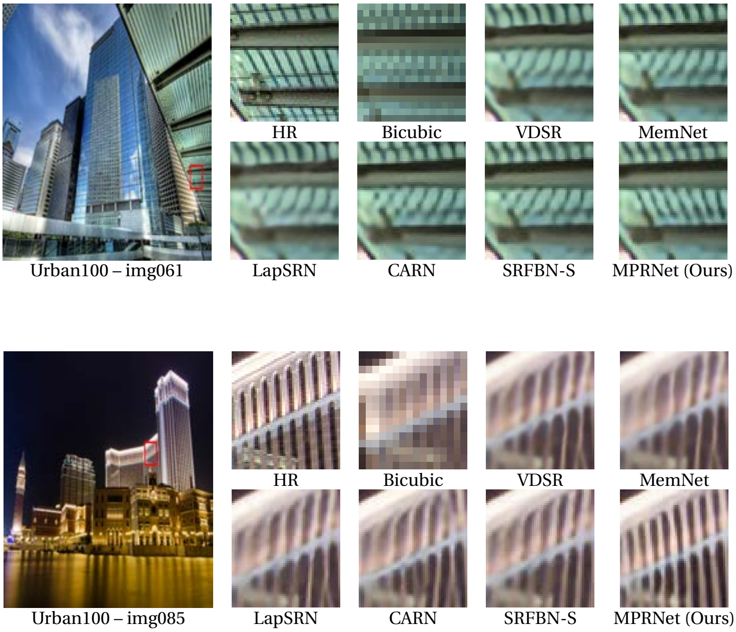


Figure 4.7: Qualitative results on **BI** degradation model with a scale factor ×4 on $Urban100$ dataset.

Table 4.7: Average Inference Time (s) and Memory Consumption (MB) comparisons with other SOTA models on Urban100 for scale factor ×4.

Model	Params.	Time	Memory	PSNR
MemNet	$667K$	0.543	$3,170$	25.54
SRFBN-S	$483K$	0.0069	$2,960$	25.71
CARN	$1592K$	0.0047	$3,015$	26.07
RCAN	$16000K$	0.5927	$2,731$	26.82
RDN	$22000K$	0.0294	$3,835$	26.61
EDSR	$43000K$	0.0841	$8,263$	26.64
MPRNet [Ours]	**$538K$**	**0.0095**	**$2,154$**	**26.31**

multiplier is a float number between 0 and 1 that controls the depth of input layer. $\alpha = 1$ is the baseline model. By decreasing α, model size and computational cost are reduced. As can be seen, the proposed MPRNet with $372.7K$ ($\alpha = 0.25$), can achieve a good performance among the lightweight SOTA methods.

Table 4.8: Impact of Depth Multiplier on MPRNet

Depth Multiplier	1.0	0.75	0.5	0.25
# Parameters	$538.2K$	$470.4K$	$416.2K$	$372.7K$
Set114 (×4)	32.38	32.23	32.01	31.84
Urban100 (×4)	26.31	26.16	25.99	25.83

As we can see, by analyzing the number of parameters and MAC operations vs PSNR, inference time, memory consumption, and reconstruction result, the proposed MPRNet can prove that it is well-balanced in terms of speed, accuracy and computation cost.

4.5.4 Additional Qualitative Results

In this section, additional results are provided showing the superiority of the SR images obtained with the proposed model. Qualitative results with all degradation models (i.e., **BI**, **BD**, and **DN**) are presented below.

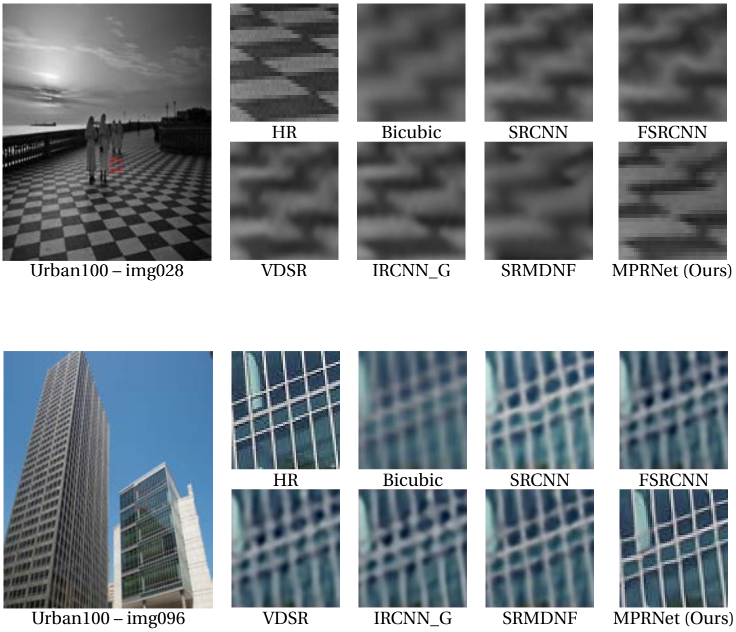


Figure 4.8: Qualitative results on **BN** degradation model with a scale factor ×3.

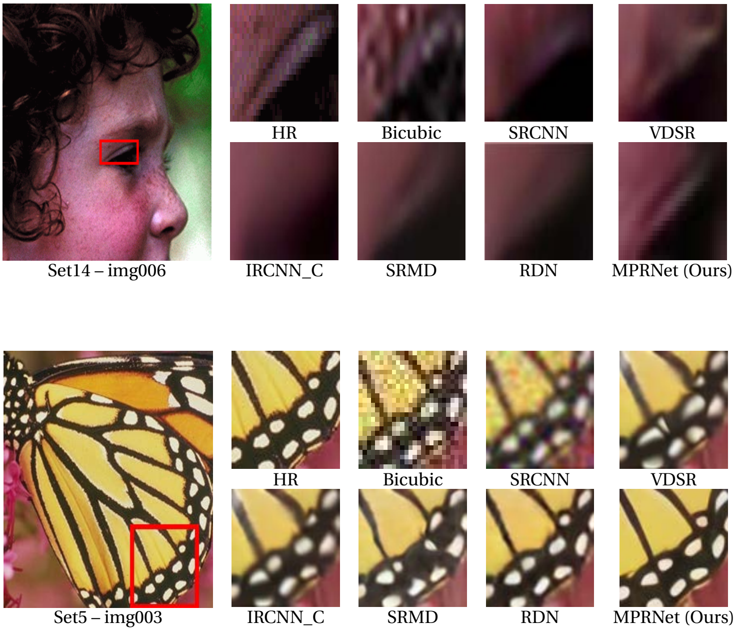


Figure 4.9: Qualitative results on **DN** degradation model with a scale factor ×3.

5 Thermal Image Super-Resolution Challenge

This chapter presents the article published at:

Rivadeneira, Rafael E., Angel D. Sappa, Armin Mehri, Parichehr Behjati Ardakani et al. "Thermal image super-resolution challenge-pbvs 2021." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.

Abstract: This paper summarizes the top contributions to the first challenge on thermal image super-resolution (TISR), which was organized as part of the Perception Beyond the Visible Spectrum (PBVS) 2020 workshop. In this challenge, a novel thermal image dataset is considered together with state-of-the-art approaches evaluated under a common framework. The dataset used in the challenge consists of 1021 thermal images, obtained from three distinct thermal cameras at different resolutions (low-resolution, mid-resolution, and high-resolution), resulting in a total of 3063 thermal images. From each resolution, 951 images are used for training and 50 for testing while the 20 remaining images are used for two proposed evaluations. The first evaluation consists of downsampling the low-resolution, mid-resolution, and high-resolution thermal images by ×2, ×3 and ×4 respectively, and comparing their super-resolution results with the corresponding ground truth images. The second evaluation is comprised of obtaining the ×2 super-resolution from a given mid-resolution thermal image and comparing it with the corresponding semi-registered high-resolution thermal image. Out of 51 registered participants, 6 teams reached the final validation phase.

5.1 Motivation

Single image super-resolution (SR) is a challenging, ill-posed problem, that is still solved using conventional methods. In recent years, deep learning techniques have shown better results. Most of these methods have been largely used in the visible spectral domain. In contrast to visible spectrum images, thermal images tend to have poor resolution, which could be improved by using learning-based traditional SR methods. These methods work by down-sampling and adding noise and blur to the given image. The poor quality noisy and blurred images, together with the given ground truth images, are used in the learning process.

The approach mentioned above has been frequently used to tackle the SR problem, however there are few contributions where the learning process is based on the usage of a pair of images (low and high-resolution images) obtained from different cameras. A novel thermal image dataset has been created containing images with three different resolutions (low-resolution (LR), mid-resolution (MR), high-resolution (HR)) obtained with three distinct thermal cameras.

The TISR Challenge[1] consists of creating a solution capable of generating a SR thermal image in ×2, ×3, and ×4 scales from cameras with different resolutions, in the conventional way by downsampling, and adding noise to the given ground truth image. Additionally, a ×2 SR image must be generated from the image obtained with a MR camera. This ×2 SR image is evaluated with respect to the corresponding image obtained from a HR camera.

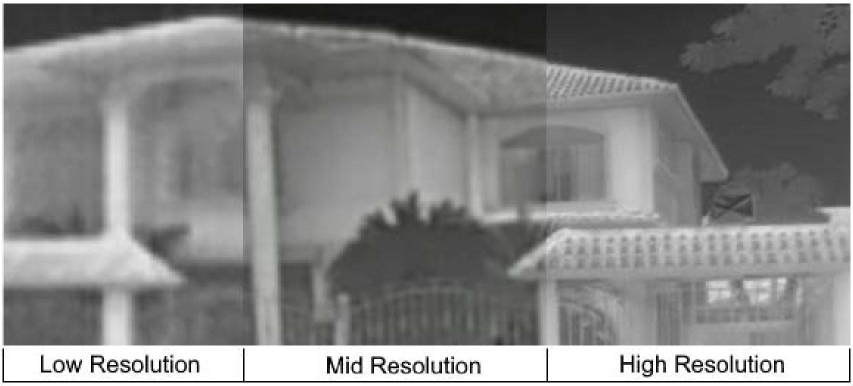


Figure 5.1: A mosaic with three different resolution thermal images from each camera for visual comparison: ($left$) crop from a LR image; ($middle$) crop from a MR image; ($right$) crop from a HR image [131].

The remainder of this paper is organized as follows. Section 5.2 introduces the objectives of the challenge, and presents the dataset and evaluation methodology. Section 5.3 summarizes the results obtained by the different teams. In Section 5.4, a short description of each teams' approach is provided. Finally, the paper is concluded in Section 5.5.

[1] http://vcipl-okstate.org/pbvs/20/challenge.html

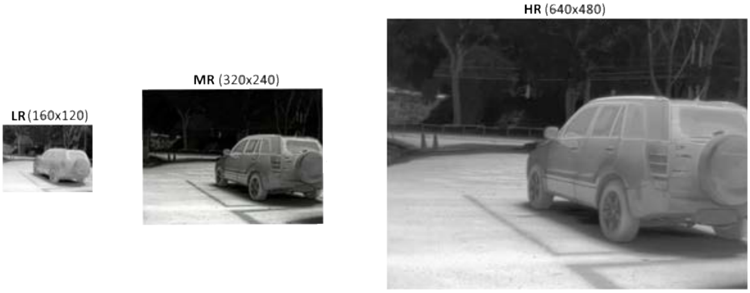


Figure 5.2: An example of thermal images acquired by each camera. $(left)$ LR image with 160×120 native resolution from an Axis Domo P1290. $(middle)$ MR image with 320×240 native resolution from an Axis Q2901-E. $(right)$ HR image with 640×480 resolution from an FC-6320 FLIR (native resolution is 640×512) [131].

5.2 TISR Challenge

The objectives of the TISR challenge are the following: (i) promote state-of-the-art approaches for the SR problem in the thermal image domain; (ii) evaluate and compare the different solutions; and (iii) promote a novel thermal image dataset to be used as a benchmark by the community working on the thermal image SR problem.

Table 5.1: Thermal Camera Specifications (Note: HR images have been cropped to 640×480) [131].

Image Description	Camera Brand	FOV	Focal Length	Native Resolution	Total # of Images
Low (LR)	Axis Domo P1290	35.4	4mm	160×120	1021
Mid (MR)	Axis Q2901-E	35	9mm	320×240	1021
High (HR)	FC-632O FLIR	32	19mm	640×512*	1021

5.2.1 Thermal Image Dataset

The dataset used in this challenge was recently presented in [131]. It consists of a set of 1021 thermal images acquired by using three thermal cameras with different resolutions. The dataset contains images from indoor and outdoor scenarios under various lighting conditions (e.g., morning, afternoon, and night) and objects (e.g., buildings, cars, people, vegetation). The cameras were mounted in a rig that minimizes the baseline distance between the optical axis such that the acquired images are almost registered. Figure 5.1 presents a mosaic obtained with images from each camera (i.e., LR, MR, and HR). The camera parameters are given in Table 5.1 and illustrations from each camera depicted in Figure 5.2.

5.2.2 Evaluation Methodology

Peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) measures are computed over a small region of the images in order to evaluate the performance of the proposed solution. In this challenge two kinds of evaluations are performed. For the first evaluation, a set of 10 down-sampled and noisy images from each resolution (LR, MR, and HR) are considered. Downsampling scale factors of ×2, ×3, and ×4 are performed and Gaussian noise of 10% is added. Figure 5.3 presents an illustration of this first evaluation process.

The second evaluation consists of computing the PSNR and SSIM of the obtained SR images with respect to the corresponding ground truth images. The ground truth images are of the same resolution as the computed SR, but are acquired with a different higher resolution camera. For the second evaluation, a set of 10 MR images is considered. The obtained SR images, from the MR set, are compared with the corresponding HR images, which have been acquired with another camera. SIFT, SURF, and ORB descriptors are used to acquire characteristic keypoints between the SR images, from the MR set, and the corresponding HR thermal images. With these feature points the mapping parameters needed to overlap the two images are computed (see [131] for more details). The evaluation measures (PSNR and SSIM) are performed over a central cropped region of the image. Figure 5.4 illustrates the second evaluation process.

5.3 Challenge Results

From 51 participants registered in the challenge, 6 teams made it to the final phase and submitted results together with their corresponding extended abstracts. Table 5.2 shows the average results (PSNR and SSIM) for each team in the two evaluations. A brief description of the thermal SR approach proposed by each team is presented in Section 5.4. Information about the team members and their affiliations is provided in Appendix A. According to the figures presented in Table 5.2, the winner of the TISR Challenge - PBVS 2020 is the MLCV-Lab_SVNIT_NTNU team, who achieved the top results in most of the evaluation tasks. The COUGER AI team achieved the best results for the second evaluation. Teams that did not reach the baseline results (i.e., bicubic interpolation) were not considered in this report.

Team	Evaluation 1						Evaluation 2	
	×2		×3		×4		×2 **(MR to HR)**	
	PSNR	**SSIM**	**PSNR**	**SSIM**	**PSNR**	**SSIM**	**PSNR**	**SSIM**
HPZ-OSU	**26,06**	0,8686	26,11	0,8373	27,32	0,8589	19,98	0,7416
CVC-UAB	26,04	0,8651	25,97	0,8326	27,12	0,8555	19,93	0,7419
MLCV-Lab_SVNIT_NTNU	25,81	**0,8858**	**26,35**	**0,8531**	**27,72**	**0,8758**	20,02	0,7452
LISA-ULB	25,57	0,8401	25,17	0,7583	26,31	0,7824	20,09	0,7385
COUGER AI	25,45	0,8529	25,96	0,8271	27,31	0,8498	**20,36**	**0,7595**
RVL-UTA	24,72	0,8325	25,37	0,8211	26,11	0,8415	19,90	0,7391
Bicubic	24,47	0,8511	25,37	0,8172	26,74	0,8421	20,24	0,7515

Table 5.2: TISR Challenge: the average results from the evaluations detailed in Section 5.2.2. The bold and underline values correspond to the first and second best results, respectively.

5.4 Proposed Approaches and Teams

This section briefly presents the approaches proposed by the different teams.

5.4.1 HPZ-OSU

HPZ-OSU team follows the s-LWSR super-resolution framework [87**? ?**], where the images are processed in small patches through residual networks. To deal with the added noise and the noise from thermal images' nature, a noise reduction process, referred to as PZ, is proposed. This approach has been originally designed for color noise, but it also works well with other kind of noise, such as white noise or Gaussian noise. The proposed network has been trained for the three super-resolution tasks of Evaluation 1.

The proposed architecture is presented in Figure 5.5; it first converts the image into the YUV space, then applies two-time-filtering along the horizontal and vertical direction. For each target pixel, the filter kernel is defined by the YUV values of the local pixels together with the relative distances to the target pixel.

All experiments were performed on a workstation with an 8GB NVIDIA 1070 GPU, using the Python programming language with PyTorch as a platform. The given dataset has been split up into 800 images for training and 151 images for testing. For the first evaluation, per each epoch, different noise values have been added, the model with best PSNR score has been selected. For the second evaluation, the same network is trained using HR (640x480) images downsampled by 2.

5.4.2 CVC-UAB

CVC-UAB team proposes a Lightweight Multi-Path Residual Network (LMPRNet) intended for thermal image super-resolution. This architecture makes the network pay attention to learning more abstract features by letting abundant low-frequency features to be avoided via multiple connections. Additionally, to seek a better trade-off between performance and applicability, a novel module is introduced, referred to as Residual Module (RM), which contains Residual Concatenation Blocks that connected to each other with global skip-connection; build with a set of Adaptive Residual Blocks (ARB) with local skip-connection, see Figure 5.6. Each ARB is defined as a divers residual path-ways learning to make use of all kinds of information form LR space. The LMPRNet design has the benefits of a multi-level learning connection and also takes advantage of propagating information throughout the network. As a result, each block has access to information of the preceding block via local and global skip-connections and passes on information that needs to be preserved. By concatenating different blocks followed by 1 × 1 convolutional layer the network can reach both intermediate and high-frequency information, resulting in better image reconstruction. Finally, a new practical attention mechanism (TFAM) is proposed by focusing on both channel and spatial information. The main objective of TFAM is to enhance the representation power of the model by emphasizing informative features and reduce worthless ones. In contrast to [167] that applies sequentially two modules by two joint sigmoid operations, which is not practical for lightweight models and edge devices. The proposed TFAM applies channel and positional units simultaneously with a different set of operations and a single hard sigmoid function.

In the training stage, input patches with a size of 60 × 60 from each of the randomly selected 64 training images were used. The number of patches is augmented by random horizontally flips and 90-degree rotation. The Adam optimizer with default setting has been employed. The initial learning rate set to 10^{-3} and its halved every 4×10^5 steps. $L1$ is used as a loss function to optimize the model. The proposed LMPRNet model is implemented in the PyTorch framework. The proposed network has been trained using a 3.2 eight-core processor with 32GB of memory with a NVIDIA GeForce GTX TITAN X GPU.

5.4.3 MLCV-Lab_SVNIT_NTNU

Figure 5.7 depicts the framework proposed by MLCV-Lab_SVNIT_NTNU team for thermal image super-resolution. A new ResBlock module, inspired from Inception network [27], is designed (see Figure 5.8). A channel attention (CA) module [**?**] is also adopted to adaptive re-scale the channel-wise features by considering interdependencies between channels. Furthermore, the local skip connection is utilized in each ResBlock in which the higher layer gradients are bypassed to the lower layer. In addition, long skip connections after six number of ResBlocks, which bypasses the higher layer gradients directly to the first convolution layer, are used. These skip connections help to solve the problem of exploding or vanishing gradient. In the proposed method, the parametric exponential linear unit (i.e., PeLU) activation function [155] is utilized. The feature maps are up-scaled to the desired resolution level by using the sub-pixel convolution layer.

All experiments have been performed on a workstation with the following specifications: Intel Core $i7-7700K$ processor, 32 GB RAM with NVIDIA GeForce GTX 1070 8GB GPU. The code is implemented using Tensorflow library. The proposed network is trained using l_1 loss function with a learning rate of 10^{-4} and the same is optimized using Adam optimizer with $\beta = 0.5$. The proposed model was trained up to 50,000 iterations with a batch size of 4. The given training images are augmented using flipping and rotating operations. In the up-sample block, the value of f is set to 63 for ×3 and 64 for ×2 and ×4.

5.4.4 LISA-ULB

The LISA-ULB team introduces a model referred to as VCycles BackProjection (VCBP); it is designed to be scalable to meet the requirements of upscaling the image by (×2, ×3, ×4) factors while maintaining the performance with a small number of parameters. The main contributions are: (i) an iterative module of shared parameters and Backprojection procedures between cycles; (ii) a new training strategy by constructing the model backwardly.

The model shown in Figure 5.9 consists of four modules: Encoder (E), Decoder (D), Upsampler and Downsampler. D and E are one convolutional layer map of the image to and from its multidimensional features. The downsampler is a one convolutional layer with stride=2 to downscale the features from the high dimension to the low dimension space. The upsampler is one dense network with 4 layers and one deconvolution layer at the end when the module is last in the sequence. All upsampler modules at the same level (L1, L2, L3, L4) share the same parameters.

The model first upsamples the input image to the target size using bicubic interpolation and uses this image with the low-resolution (LR) image as inputs for the model. The model is responsible to generate the residual high-frequency (HF) information and add them to the encoded features before decoding them back into the image space. In each VCycle the model downscales the accumulated features from the previous and the current cycle to be the input for the next cycle. This procedure enforces the model to generate features in the high-resolution (HR) space while maintaining the similarity in its LR space. All downsampler modules share the same parameters to ensure that all earlier down-sampled features are similar to the encoded features of the LR input image. Only the last down-sampled features are used for the backprojection loss.

$\mathscr{L}^{L1}(x, y) = \mathbb{E}[|x - y|]$ is used for the content loss and the Backprojection loss. The total loss function is:

$$\mathscr{L}^{L1}(SR, HR) + 0.01 * \mathscr{L}^{L1}(BP, a),$$

(BP) and (a) are the encoded features in the LR space of the down-sampled super-resolved features and the input image features respectively.

The VCBP network has been implemented in Pytorch and performed on a NVIDIA TITAN XP. The proposed model was trained using the AdamW optimizer followed by SGDM. Instead of building the model forwardly, a new building and training procedure is proposed, which add models backwardly for each new upscaling factor as shown in Figure 5.10. The last model in the sequence responsible for generating all the natural super-resolved images, while added earlier stages responsible for producing intermediate super-resolved images. This allows to train the model on larger image sizes. The model has 883K parameters and each upscaling factor module was trained on the three training sets.

5.4.5 COUGER AI

The COUGER AI team proposes an architecture for the task of generating ×2, ×3 and ×4 resolution images, which are acquired at three different resolutions. The proposed approach is based on a neural network that utilizes the coordinate convolutional layer [102] and residual units [**?**], along with the multi-level supervision and attention unit to map the information between LR images to MR and HR images.

As shown in Figure 5.11, firstly, the bicubic interpolated input image is mapped in Cartesian space using the coordinate convolutional layer [102]. In each base block, two residual units along with the one convolutional layer are used. The output of each base block is up-sampled according to the output resolution requirement. All the up-sampled outputs are then fused to the Convolutional Block Attention Module (CBAM) [167]. Also, to improve the pixel-wise resolution, a multi-level supervision is applied on each up-sampled layer, inspired on [77] [122].

To supervise the model output, a combination of three losses are used: mean squared error (MSE), SSIM, and Sobel, i.e.,

$$Total_{Loss} = MSE + SSIM_{Loss} + SOBEL_{Loss}.$$

In total, three $(\times 2, \times 3, \times 4)$ networks are trained for generating the high-resolution images from the low-resolution input images in Keras 2.2.4. Input images are normalized between 0 to 1 and introduced with Gaussian noise (mean = 0 and sigma = 10). The dataset was trained using a NVIDIA 1080 GTX GPU.

5.4.6 RVL-UTA

The RVL-UTA team presents a novel network for thermal image super-resolution (SR) called the Multiscale Residual Channel Attention Network (MSRCAN). The architecture is inspired by state-of-the-art methods to recover details from low-resolution (LR) RGB images such as: very deep residual channel attention networks (RCAN) [**?**], learning a mixture of deep networks for single image SR (MSCN) [97], and multiscale convolutional neural networks (CNNs) (MSSR) [73]. RCAN allows deeper CNN models, which result in more feature representation. MSCN uses multiple parallel inference modules with sequentially increased dilation factors and an adaptive weight (AW) module allowing for multiscale SR outputs and pixel-wise AW summation. MSSR provides a model for parallel CNN paths with different depths corresponding to multiscale SR image outputs. The proposed MSRCAN implements a combination of all these networks to produce higher PSNR and SSIM scores, as well as sharper SR output images.

The inputs to MSRCAN are bicubicly up-sampled LR images. These LR images pass through parallel RCAN SR inference modules, which produce high-resolution (HR) estimates that are aggregated using AW modules at the pixel level. The receptive field of the convolutions within each of the SR modules linearly increases against the number of modules. This is done by increasing the dilation factor by two for each of the parallel modules. Different receptive fields of each SR inference module allow the network to produce an HR estimate for varying scales as was done with MSSR. Each SR inference module is pixel-wise multiplied with its corresponding AW module according to the architecture of MSCN. The sum of these pixel-wise products produces the HR estimate. An overview of the network architecture is shown in Figure 5.12.

MSRCAN was trained on a workstation with a NVIDIA Quadro P4000 GPU, Intel Core i7-8700 CPU, and 32GB of RAM. It was written with the Python 3.7.6 programming language and the Tensorflow v1 library. In addition, the following modules were used: PIL, Pyelastix, OpenCV, ImageIO, TQDM.

5.5 Summary

This paper summarizes the best contributions to the Thermal Image Super-Resolution Challenge - PBVS 2020, where 51 teams from 17 different countries have participated and 6 teams reached the final validation phase. This was the first time this challenge has been proposed and a wide interest from the research community has been observed. Undoubtedly, the results from this year will be used as the benchmark for next year's challenge. This challenge has also been an opportunity to promote the evaluation dataset used by the participating teams.

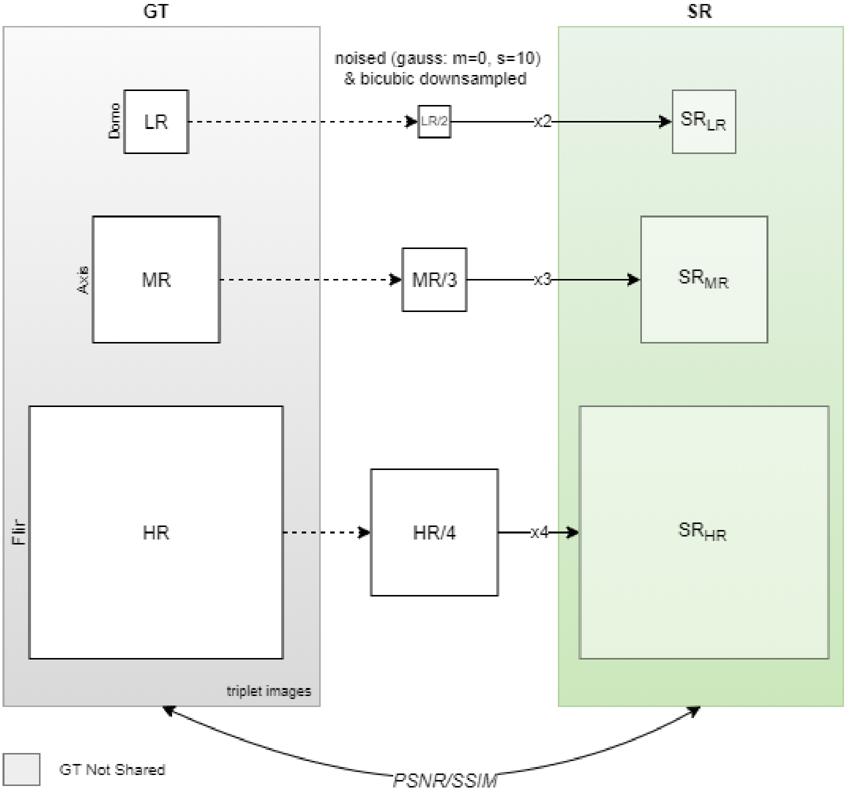


Figure 5.3: An illustration of the first evaluation process (×2 for low, ×3 for mid, ×4 for high).

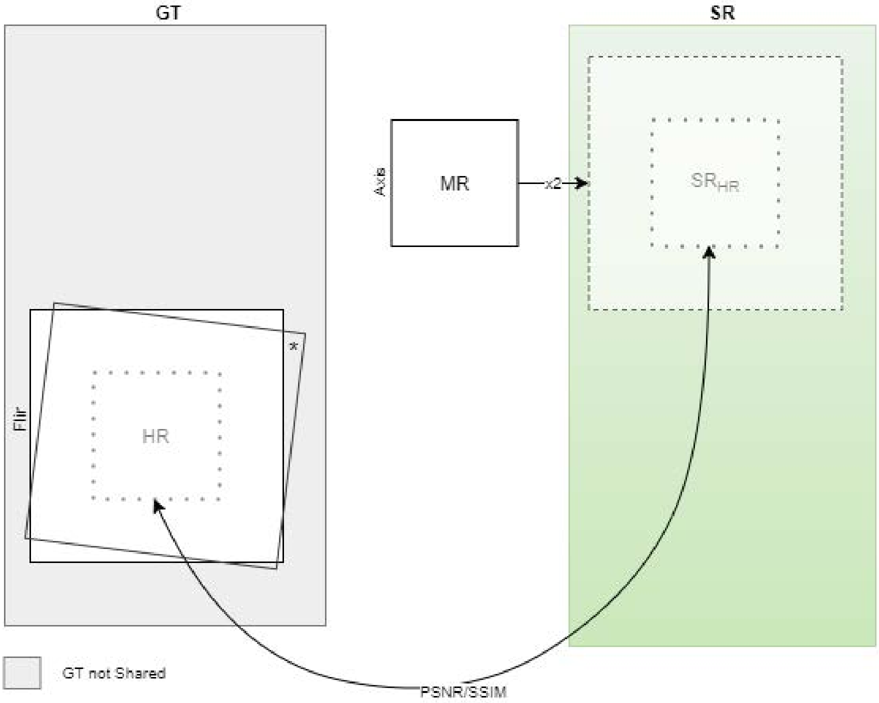


Figure 5.4: A illustration of the second evaluation process. Note that this evaluation is applied over a set of 10 MR images.

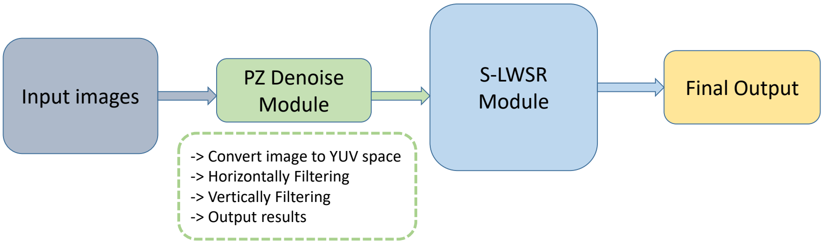


Figure 5.5: HPZ-OSU proposed architecture with color noise reduction process, where the $s - LWSR$ module uses the framework in [87].

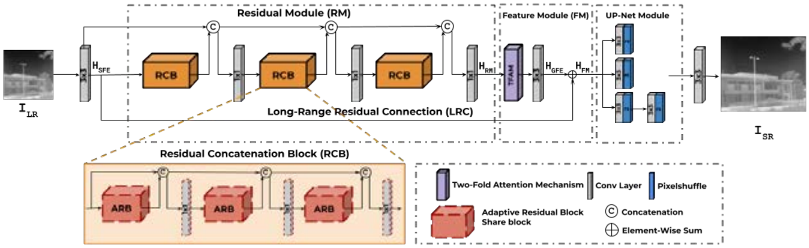


Figure 5.6: CVC-UAB architecture of the proposed Lightweight Multi-Path Residual Network (LMPRNet).

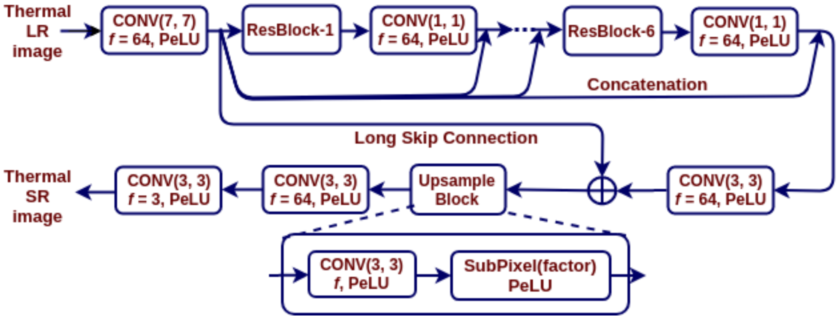


Figure 5.7: MLCV-Lab_SVNIT_NTNU proposed architecture.

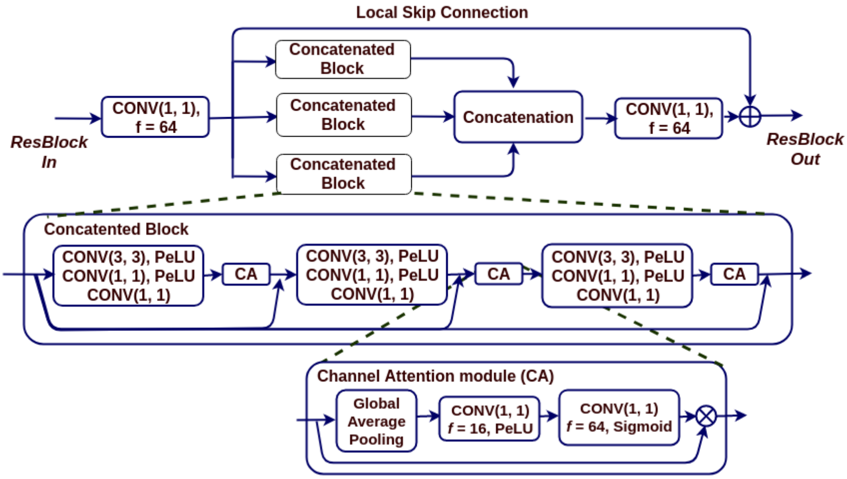


Figure 5.8: ResBlock design for the MLCV-Lab_SVNIT_NTNU architecture.

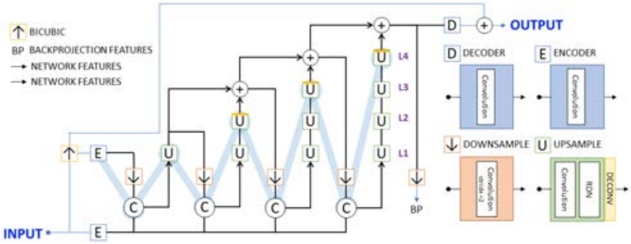


Figure 5.9: LISA-ULB proposed architecture.

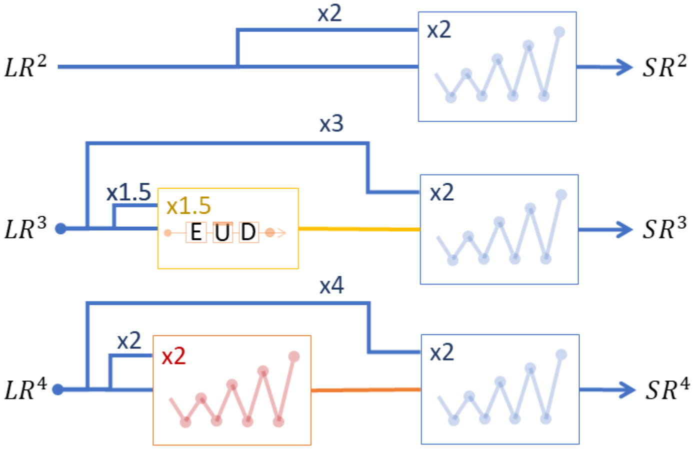


Figure 5.10: LISA-ULB proposed building model.

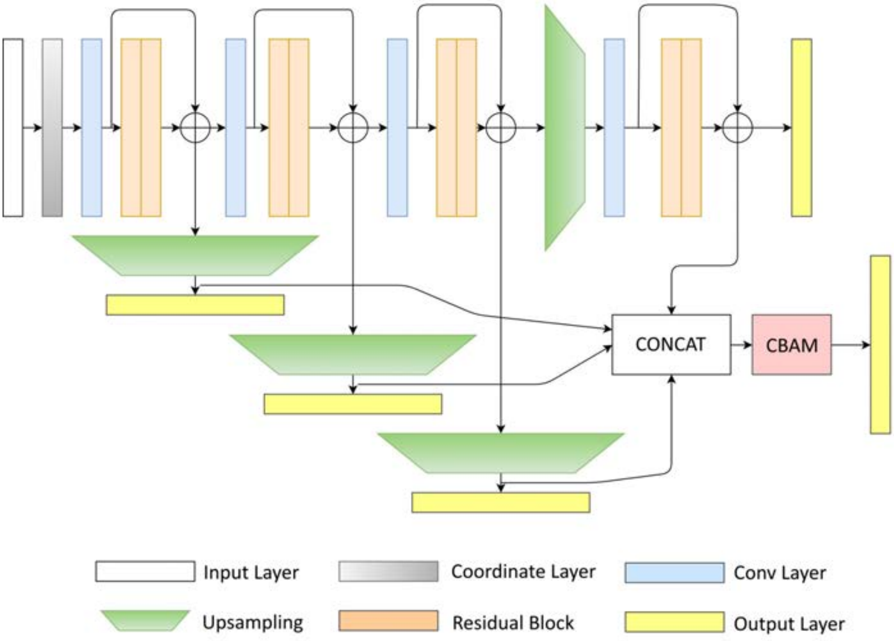


Figure 5.11: COUGER AI proposed architecture: A Multi-Level Supervision Model

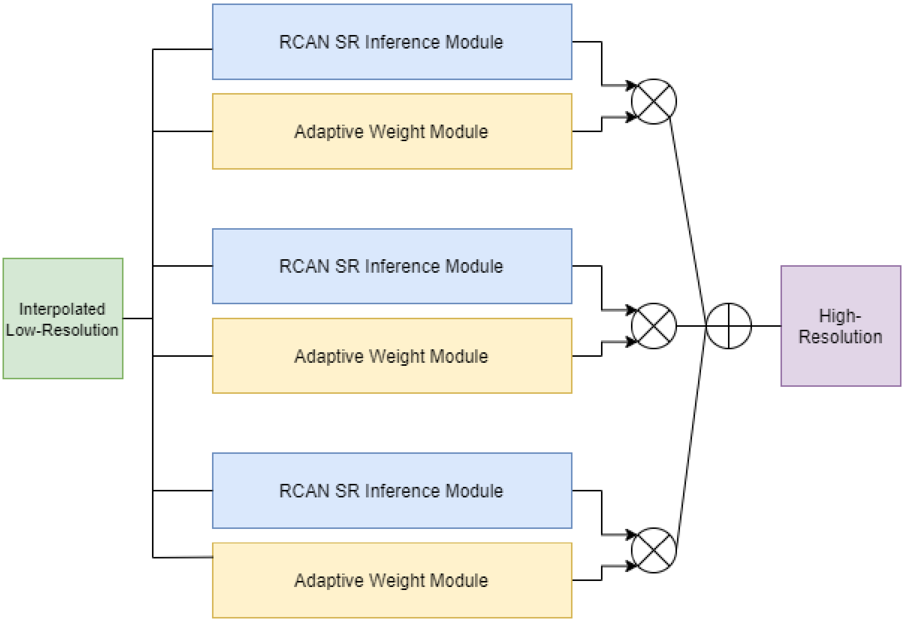


Figure 5.12: The RVL-UTA proposed MSRCAN architecture.

6 SRFormer: Efficient Yet Powerful Transformer Network For Single Image Super Resolution

This chapter presents the article submitted at:

Armin Mehri, Parichehr Behjati, and Angel D. Sappa. SRFormer: Efficient Yet Powerful Transformer Network For Single Image Super Resolution.IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022. (under review)

Recent breakthroughs in single image super resolution have investigated the potential of deep Convolutional Neural Networks (CNNs) to improve the performance. However CNNs based models suffers from limited respective field and in adaptability to the input content. Recently, Transformer based models were presented, which demonstrated major performance gains in Natural Language Processing and Vision tasks while mitigating the drawbacks of CNNs. Nevertheless, Transformer computational complexity can increase quadratically for high-resolution images and the fact that it ignores the original structures of the image by converting them to 1D structure can be problematic to capture the local context information and adapting them for real-time applications. In this paper, we present SRFormer, an efficient yet powerful Transformer based architecture by making several key designs in the building Transformer blocks and Transformer layer such that allows to consider the original structure of the image (i.e., 2D structure) while capturing both local and global dependencies without raise of computational demands and memory consumption. We also present Gated MLP Feature Fusion module to aggregate the features of different stages of Transformer blocks by focusing on inter-spatial relationship, while adding minor computational cost to the network. Extensive experiments are carried out on a number of super resolution benchmark datasets to evaluate the proposed SRFormer approach. SRFormer delivers superior performance when compared to the state-of-the-art algorithms both Transformer and Convolutional based networks.

6.1 Motivation

Super Resolution has been studied since 1974, when Gerchberg [12] introduced the notion of Super Resolution (SR) to improve optical system resolution over and above diffraction, since then the idea of super resolution has been defined as a way for obtaining high resolution (HR) images from its degraded low resolution (LR) image with a high visual quality, more realistic textures and enhanced in details of the given low-resolution input image.

Although super resolution being explored for decades, single image super resolution is still an active yet challenging topic in Computer Vision due to its complex nature and high practical values in improving image details and textures. The recent success of image super resolution has the potential to significantly improve the quality of media content, resulting in better user experiences. For example, the digital zoom algorithm used in mobile cameras and the image enhancement technology used in digital devices. Furthermore, this core technology can be applied to a wide range of Computer Vision tasks, which leads to improvements in various Vision tasks, such as object detection [43], medical imaging [46], security and surveillance imaging [196], face recognition [119], astronomical images [105] and many other domains [98, 176].

There are several reasons that makes image super resolution remains challenging: i) Super Resolution is fundamentally an ill-posed inverse problem. There are multiple solutions for the same low quality image instead of an unique single solution. ii) The complexity of problem increases, as the up-scale factor increases. The retrieval of missing scene details becomes even more complicated with greater factors, which often leads to the reproduction of incorrect information; and iii) there are fundamental uncertainties among the LR and HR data since the down-sampling of different HR images may lead to a similar LR image [35].

Formerly, different methods were utilized to tackle the super resolution problems, such as statistical methods, prediction-based methods, patching methods, edge-based methods and sparse representation method. However, researchers have lately been using Deep Learning (DL) approaches to solve the problems of image super resolution due to advance progress in computers computational power.

Deep learning ConvNet based approaches have consistently demonstrated significant improvement to the classical methods over the last decade. Numerous deep convolutional neural networks introduced [1, 28, 92, 100] as well as many lightweight networks and techniques to reduce the computational complexity of the networks, such as using filter pruning [50, 61], knowledge distillation [41, 171] to minimize computing time by narrowing the network and, quantization techniques [111] to lowering the computational complexity while keeping the original network's

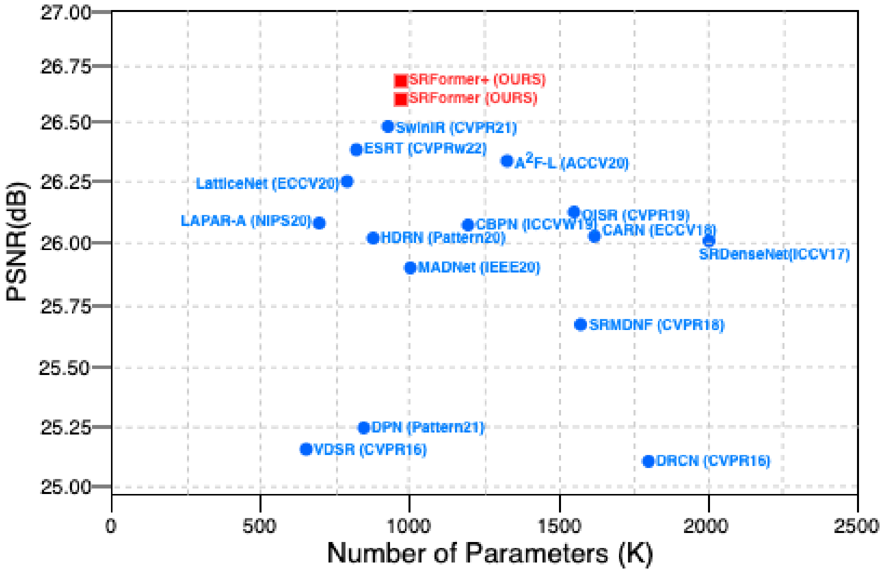


Figure 6.1: PSNR *vs.* Model size trade-off on Urban100 (×4). SRformer achieves superior performance among all the CNN and Transformer based network.

architecture. However, these techniques often leads to poor performance due to several reasons such as lower network capacity, long inference time and a large number of operations due to several iterate through the forward process.

In addiction, ConvNet based approaches suffers from two main issues that come from the fundamental of the convolution layer. First, there is no content dependency in the interactions between images and convolution kernels. The same convolution kernel uses to restore various images regions, which it is not the ideal solution. Second, convolution is effective to capture local context information but ineffective for capturing long-range dependency [93].

Transformer [156] introduced to tackle the aforementioned problems of convolution layer, by designing a self-attention mechanism to capture global interactions between contexts, has shown promising performance in several Vision and NLP tasks [21, 38, 104]. However, self-attention mechanism computational cost increases quadratic when dealing with spatial resolution and also ignore the local 2D structure information of the image by processing images as a 1D structure [49]. Furthermore, these methods usually needs to occupy heavy GPU memory, which greatly limits their flexibility and application scenarios for low-capacity devices.

In this paper, we propose a novel lightweight approach for single image super resolution task, namely SRFormer by bringing the strengths of both convolution layer and Transformer layer together to address the aforementioned problems. By advancing both Convolution and Transformer together, SRFormer is able to capture both local context information and global interactions between contexts, while computationally stay efficient. The combination of both CNN and Transformer together with the precise design of our SRFormer architecture, allow our model to perform exceptionally well on benchmark datasets with faster training and inference time compare to other Transformer based network. It is worth to mention that, SRFormer trained with only a single GPU for 3 days while SwinIR trained on 8 GPUs for almost 2 days to achieve their results. Also, SRFormer has a advantage of multi-scale training, which can generate SR images with different scale factors [×2, ×3, ×4] in one training phase while other methods needs to train separately for each scale factors. As illustrated in Fig. 6.1, The proposed SRFormer yields to **26%** improvement on average of all benchmark datasets for scale factor 4 when compared to SwinIR [93]– SOTA Transformer based model which shows the efficiency of the propose model.

The main contributions of our work can be summarized as follow:

- We present SRFormer, an efficient yet powerful Transformer based network for single image super resolution task, which is faster in training and inference while generate more accurate SR images.

- We present a lightweight Dual Attention layer, which significantly improves the reconstruction quality by generating global attention map from two local attention weights, which obtain individually by two branches in parallel while its not memory hunger.

- We present a low cost Gated MLP Feature Fusion module that yields to a powerful representation by aggregating multi-stage feature representation from Transformer blocks with minor computation complexity.

- Extensive experiments show that SRFormer achieves state-of-the-arts on various benchmark datasets for SISR task compared to CNN/Transformer based networks.

The rest of the paper is organized as follows: Section 6.2 describes the proposed SRFormer and its core components in detail. Experimental comparisons against several state of-the-art methods are presented in Section 6.3. Model investigation presents in section 6.4. Section 6.5 concludes the paper.

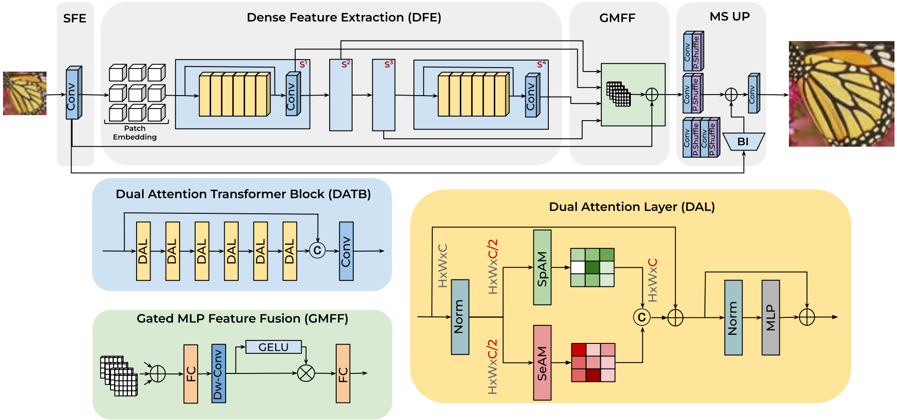


Figure 6.2: The overall network architecture of the proposed SRFormer.

6.2 Proposed Method

In this section, the overall network architecture of proposed SRFormer is described. Later, the detailed information of Dual Attention layer is provided.

6.2.1 Overall Pipeline

The primary goal is to design an efficient Transformer based architecture, which is able to generate a well detailed high quality images while remaining computationally efficient. Thus, we utilize the basic Transformer structure but specially designed for efficient network structure with significant performance gains compared to existed CNN and Transformer networks. The overall architecture of the SRFormer is illustrated in Fig.2. In particular, the proposed SRFormer consists of four modules: Shallow Feature Extraction (SFE), Dense Feature Extraction (DFE), Gated MLP Feature Fusion (GMFF), and Multi-Scale Up-Sampling (MS-UP) modules. We defined I_{LR} and I_{SR} as the low-quality input and high-quality output of our network, respectively.

6.2.2 Shallow Feature Extraction

The convolution layer prove that can performs well at early visual processing, which leading to improve the performance of the network [168]. Therefore, a single 3×3 convolutional layer is applied on the given low-quality input image I_{LR} to extract the initial features and map the input image space to a higher dimensional feature space to generate better SR image. Therefore, we extract the shallow features F_0 as:

$$\boldsymbol{F}_0 = Conv_{3\times3}(\boldsymbol{I}_{LR}), \tag{6.1}$$

6.2.3 Dense Feature Extraction

Next, the extracted shallow feature pass through the Dense Feature Extraction F_{DFE} as an input. DFE built up with the set of Transformer blocks. The input first processed by input embedding such as patch embedding for Vision Transformers (ViTs):

$$\boldsymbol{I}_{EMB} = InputEmb(F_0), \tag{6.2}$$

where I_{EMB} denotes the embedding tokens with length of N sequence and C embedding dimension. Our Dense Feature Extraction module takes embedding tokens as a input to our Transformer blocks. Specifically, Dense Feature Extraction contains of several Transformer blocks, which include i^{th} Transformer layers and a 1 × 1 Conv layer at the end of each block with benefit of waterfall residual connection to transfer the information from previous stage to current stage. The shallow features from SFE process through different Transformer stages to extract more abstract features and spotlights the high-level information (further details provided in section 7.2.7). Thus, we extract the feature as follow:

$$\boldsymbol{F}_{DFE} = H_{DFE}(I_{EMB}), \tag{6.3}$$

where $H_{DFE}(.)$ is Dense Feature Extraction module with several Transformer blocks, which can be seen as

$$\boldsymbol{F}_i = Conv_{1\times1}(C[H_{DATB}(F_{i-1}), X_{i-1}], i = 1, 2, ..., K, \tag{6.4}$$

where $H_{DATB}(.)$ denotes the i_{th} Transformer blocks. C denotes the concatenation operation between the input feature of each $DATB$ block and the output. By concatenating a convolutional layer within each stage of the Transformer block, help to transfer inductive bias from the convolution operation into the Transformer-based network and provide a more solid foundation for the later aggregation of shallow and deep features together.

6.2.4 Gated MLP Feature Fusion

The aim of Gated MLP Feature Fusion (GMFF) design is to highlight the location information in the stacked feature map of different stages of Transformer blocks.

GMFF consists of N stacked residual $DATB$ as shown in Fig 6.2. GMFF first, accumulates the multi-stage features form different Transformer stages to create a multi-stage representations of the input image. Then, passes the features through the lightweight MLP network. However, in contrast to standard MLP network, we propose a novel MLP module by using a 3 × 3 Depthwise Conv layer inside the module to leak the spatial information in order to boost the network performance since highlighting such features are important in super resolution task to achieve high performance. Also, gating mechanism used by formulating the element wise product of two parallel routes of linear transformation layer that one is activated with the GELU [58]. Thus, Gated MLP Feature Fusion can be formulated as follow:

$$\boldsymbol{F}_{GMFF} = MLP(GELU(Conv_{3\times3}(MLP(F_i)))) + F_0, \tag{6.5}$$

where F_{GMFF} denotes the output of our feature aggregation of multi-stage Transformer block with the initial features, which later used by Multi Scale Up-Sampling module. In ablation study, we will show the effectiveness of our proposed Gated MLP Feature Fusion compared to the standard MLP network.

6.2.5 Multi Scale Up-Sampling

Given the feature from previous modules, which contains an aggregation of low- and high-level information, our model generate a high-quality image I_{SR}. Multi Scale Up-Sampling (MSUP) module takes the features directly from GMFF module to be able to reconstruct the high quality output. MSUP consists of several convolutional and pixel-shuffle layers to upsample the features to the corresponding sizes in one training phase instead of training for each interested scale factor separately. Furthermore, we incorporate a global connection path H_{UP} with only a bicubic interpolation to grant access to the original LR information and facilitate the back-propagation of the gradients. The Multi Scale Up-Sample module can be formulated as:

$$\boldsymbol{I}_{SR} = H^{\uparrow}_{Rec}(F_0 + F_{GMFF} + H_{UP}(I_{LR})), \tag{6.6}$$

where $H_{Rec}(\cdot)$ and I_{SR} denotes the up-sampling module and high quality reconstructed image respectively:

6.2.6 Loss Function

To keep the consistency with previous works, we use L_1 loss as a cost function during training to optimize the parameters of proposed SRFormer.

$$L_1(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N} \| I_{SR} - I_{HR} \|_1 , \tag{6.7}$$

where I_{SR} obtained by taking low-quality image as the input of our model and I_{HR} is the corresponding ground-truth.

In the next subsections, more details about our Transformer layer are given.

6.2.7 Dual Attention Layer

This section presents the proposed Dual Attention layer with completely revising the token mixer (i.e., self-attention). As well known, self-attention is playing important role to achieve high performance in Natural Language Processing (NLP) and Computer Vision Transformer based networks. However, self-attention can be problematic due to several reasons, especially when it comes to work with spatial resolution, which involves high-resolution images. The computational complexity of self-attention increases quadratically to the number of tokens to mix. Besides that, self-attention treats images as flatten sequence, which neglects the original structure of images therefore it ignores the adaptability in channel dimensions, which proven important for visual task. Also, self-attention does not take into account the local contextual information due to nature of self-attention. Thus, we introduce Dual Attention layer to overcome the aforementioned shortages by generating a global attention map with less computational cost compared to existing token mixer. Dual Attention generates a global attention map by aggregating two local attention maps, which are separately obtained by using two different branches, CNN based Attention Module and Transformer self-attention in parallel. By doing so, unlike to the previous token mixer, the Dual Attention can also consider both long range dependency and local contextual information with less computationally complexity.

As shown in Fig 6.2, we design our Dual Attention in a way that it splits the channel features equally for both attention module branches (SpAM and SeAM). From Norm layer tensor X, both of our branches receive half of the input tensor to create the local attention maps individually. SeAM is a self-attention Transformer, which first generates the query (Q), key (K), and value (V) projections enriched with local context. We apply SeAM only across the channels rather than spatial dimensions. Our SeAM uses only depth-wise convolutions to emphasize the channel-wise spatial context before computing feature covariance to produce the attention map. Thus, Q, K, V computed as:

$$Q = W_d^Q Y, K = W_d^K Y, V = W_d^V Y \tag{6.8}$$

where $W_d^{(\cdot)}$ is the 3×3 bias-free depth-wise convolution. Next, query and key projections reshape in a way that their dot-product interaction generates a transposed-attention map. Thus, attention map generate as follow:

$$Attention(Q,K,V) = W_d(V.Softmax(K.Q/\alpha)) + X \tag{6.9}$$

where X is the input feature map and α is a learnable scaling parameter that is used to regulate the magnitude of the dot product of K and Q before applying the Softmax function. Similar to previous works [93, 156, 177] we perform the attention function for h times to learn separate attention maps in parallel in our SeAM module.

Second branch of Dual Attention layer is Spatial Attention Module (SpAM), which is an almost parameter free attention mechanism. SpAM receive the other half of the input tensor to generate the local attention map. The goal of SpAM module is to encode the spatial information, which represents the importance of each pixel in the input feature with negotiable cost. Given half of the input tensor information, the channels of the input tensor are reduced by mean and max operations, of which the shape is $1 \times H \times W$. The obtained features concatenated, then passed through a convolution layer with kernel size of 7×7. After, a sigmoid activation layer apply to the output feature to generate the attention weights of shape $1 \times H \times W$ which are later multiply with the input tensor to refined tensors of shape $C \times H \times W$. Thus, the SpAM can be formulated as follow:

$$X = Sigmoid(Conv_{7\times7}[F_{Mean}(X), F_{Max}(X)]) * X \tag{6.10}$$

where $F_{Mean}(\cdot)$ and $F_{Max}(\cdot)$ denotes for mean and max operations. Later, generated local attention maps from SpAM and SeAM are concatenated together to obtain a unify global attention map with less computational cost. Thus, the generated attention map contains both long range dependency and local context information with enrich of spatial features.

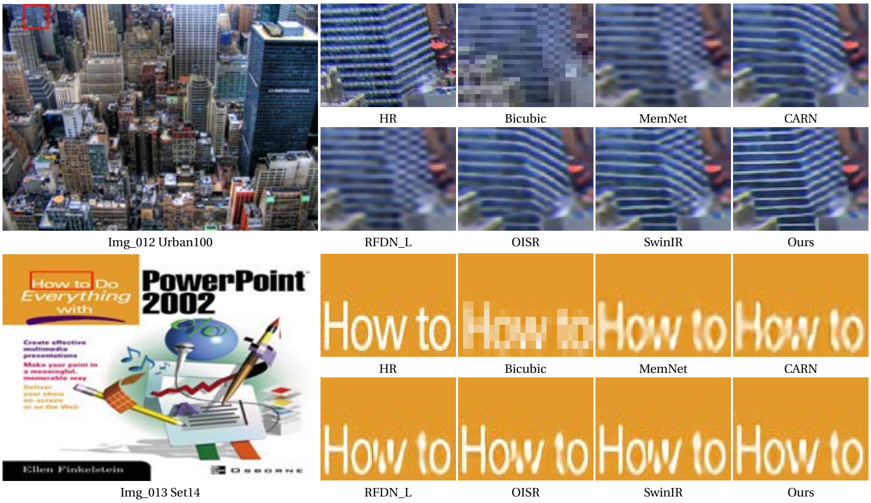
Following that, a multi-layer perceptron (MLP) with two fully connected layers and a GELU non-linearity activation function between them is employed for further feature modifications. The norm layer is also added before MLP, and both modules contains the residual connection between them. Thus, the entire procedure inside of our Dual Attention is as follow:

$$\begin{aligned} X &= (Norm(SpAM(X/2) + SeAM(X/2))) + X \\ Y &= MLP(Norm(X)) + X \end{aligned} \tag{6.11}$$

where $Norm(\cdot)$ stands for normalization layer and Y for the output feature map.

Table 6.1: Average PSNR/SSIM comparison with state-of-the-art CNN- and Transformer-based methods with the same range of network parameters on the Bicubic (**BI**) degradation for scale factors [×2, ×3, ×4] (Transformer based methods separated with horizontal line). **Red** is the Best and **Blue** is the second best performance. We assume that the generated SR image is 720P to calculate Multi-Adds (MAC). SRFormer with self-ensemble results are **Highlighted**.

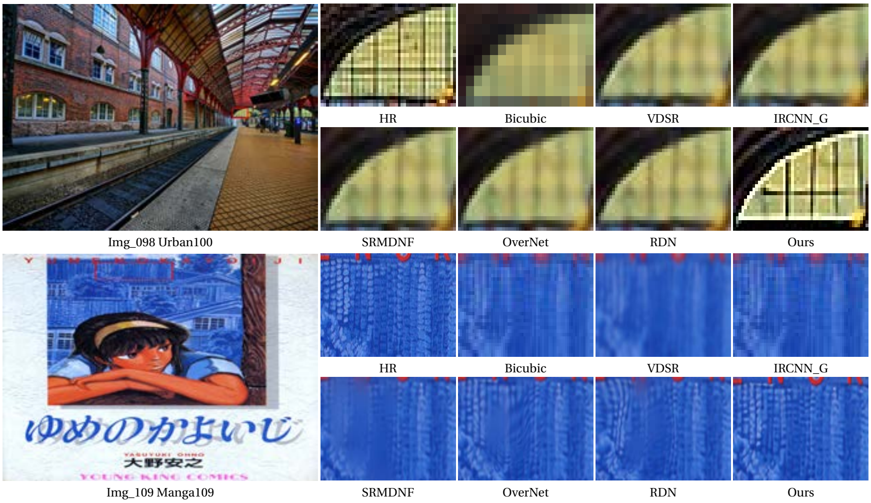
Scale	Method	Params	FLOPs	Set5		Set14		B100		Urban100		Manga109	
				PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
×2	VDSR [78]	665K	613G	37.53	0.9587	33.03	0.9124	31.90	0.8960	30.76	0.9140	37.22	0.9750
	DRCN [79]	1,774K	17,974G	37.53	0.9587	33.03	0.9124	31.90	0.8960	30.76	0.9140	37.22	0.9750
	CARN [1]	1,592K	223G	37.76	0.9590	33.52	0.9166	32.09	0.8978	31.92	0.9256	38.36	0.9765
	CBPN [194]	1,036K	240.7G	37.90	0.9590	33.60	0.9171	32.17	0.8989	32.14	0.9279	–	–
	FALSR-A[30]	1,021K	234.7G	37.82	0.9595	33.55	0.9168	32.12	0.8987	31.93	0.9256	–	–
	SRMDNF[186]	1,513K	348G	37.79	0.9600	33.32	0.9150	32.05	0.8980	31.33	0.9200	–	–
	LAPAR-A[91]	548K	171G	38.01	0.9605	33.62	0.9183	32.19	0.8999	32.10	0.9283	38.67	0.9772
	OISR-LF-s[57]	1,370K	316.2G	38.02	0.9605	33.69	0.9178	32.20	0.9000	32.21	0.9290	–	–
	LatticeNet [109]	756K	169.5G	38.15	0.9610	33.78	0.9193	32.25	0.9005	32.24	0.9302	–	–
	MADNet [85]	878K	187.1G	37.94	0.9604	33.46	0.9167	32.10	0.8988	31.74	0.9246	–	–
	HDRN [74]	878K	316.2G	37.75	0.9590	33.49	0.9150	32.03	0.8980	31.87	0.9250	38.07	0.9770
	DPN [94]	832K	140G	37.52	0.9586	33.08	0.9129	31.89	0.8958	30.82	0.9144	–	–
	A^{2}F-L [159]	1,363K	306.1G	38.09	0.9607	33.78	0.9192	32.23	0.9002	32.46	0.9313	38.95	0.9772
	ESRT [106]	677K	–	38.03	0.9600	33.75	0.9184	32.25	0.9001	32.58	0.9318	39.12	0.9774
	SwinIR [93]	878K	195.6G	38.14	0.9611	33.86	0.9206	32.31	0.9012	32.76	0.9340	39.12	0.9783
	SRFormer (Ours)	958K	183.8G	38.11	0.9611	33.92	0.9221	32.35	0.9023	32.82	0.9398	39.23	0.9801
	SRFormer+ (Ours)	958K	–	**38.18**	**0.9621**	**33.98**	**0.9232**	**32.41**	**0.9036**	**32.88**	**0.9409**	**39.29**	**0.9821**
×3	VDSR [78]	665K	613G	33.66	0.9213	29.77	0.8314	28.82	0.7976	27.14	0.8279	37.22	0.9750
	DRCN [79]	1,774K	17,974G	33.82	0.9226	29.76	0.8311	28.80	0.7963	27.15	0.8276	32.24	0.9343
	CARN [1]	1,592K	119G	34.29	0.9255	30.29	0.8407	29.06	0.8034	28.06	0.8493	33.50	0.9440
	SRMDNF [186]	1,530K	156G	34.12	0.9250	30.04	0.8370	28.97	0.8030	27.57	0.8400	–	–
	LAPAR-A [91]	544K	114G	34.36	0.9267	30.34	0.8421	29.11	0.8054	28.15	0.8523	33.51	0.9441
	OISR-LF-s[57]	1,550K	160.1G	34.39	0.9272	30.35	0.8426	29.11	0.8053	28.24	0.8544	–	–
	LatticeNet [109]	765K	76.3G	34.53	0.9281	30.39	0.8424	29.15	0.8059	28.33	0.8538	–	–
	MADNet [85]	930K	88.4G	34.26	0.9262	30.29	0.8410	29.04	0.8033	27.91	0.8464	–	–
	HDRN [74]	878K	187.1G	34.24	0.9240	30.23	0.8400	28.96	0.8040	27.93	0.8490	33.17	0.9420
	DPN [94]	832K	114.2G	33.71	0.9222	29.80	0.8320	28.84	0.7981	27.17	0.8282	–	–
	A^{2}F-L [159]	1,367K	136.1G	34.54	0.9283	30.41	0.8436	29.14	0.8062	28.40	0.8574	33.83	0.9463
	ESRT [106]	770K	–	34.42	0.9268	30.43	0.8433	29.15	0.8063	28.46	0.8574	33.95	0.9455
	SwinIR [93]	886K	87.2G	34.62	0.9289	30.54	0.8463	29.20	0.8082	28.66	0.8624	33.98	0.9478
	SRFormer (Ours)	958K	81.6G	34.67	0.9301	30.59	0.8470	29.26	0.8095	28.72	0.8652	34.06	0.9488
	SRFormer+ (Ours)	958K	–	**34.72**	**0.9313**	**30.66**	**0.8484**	**29.32**	**0.8105**	**28.79**	**0.8686**	**34.11**	**0.9502**
×4	VDSR [78]	665K	613G	31.35	0.8838	28.01	0.7674	27.29	0.7251	25.18	0.7524	28.83	0.8809
	DRCN [79]	1,774K	17,974G	31.54	0.8850	29.19	0.7720	27.32	0.7280	25.12	0.7560	29.09	0.8845
	SRDenseNet [154]	2,015K	390G	32.00	0.8931	28.50	0.7782	27.53	0.7337	26.05	0.7819	30.41	0.9071
	CARN [1]	1,592K	91G	32.13	0.8937	28.60	0.7806	27.58	0.7349	26.07	0.7837	30.47	0.9084
	CBPN [194]	1,197K	97.9G	32.21	0.8944	28.63	0.7813	27.58	0.7356	26.14	0.7869	–	–
	SRMDNF [186]	1,555K	89G	31.96	0.8930	28.35	0.7770	27.49	0.7340	25.68	0.7730	–	–
	LAPAR-A [91]	659K	94G	32.15	0.8944	28.61	0.7818	27.61	0.7366	26.14	0.7871	30.42	0.9074
	OISR-LF-s[57]	1,520K	114.2G	32.14	0.8947	28.63	0.7819	27.60	0.7369	26.17	0.7888	–	–
	LatticeNet [109]	777K	43.6G	32.30	0.8962	28.68	0.7830	27.62	0.7367	26.25	0.7873	–	–
	MADNet [85]	1,002K	54.1G	32.11	0.8939	28.52	0.7799	27.52	0.7340	25.89	0.7782	–	–
	HDRN [74]	867K	316.2G	32.23	0.8960	28.58	0.7810	27.53	0.7370	26.09	0.7870	30.43	0.9080
	DPN [94]	832K	140G	31.42	0.8849	28.07	0.7688	27.30	0.7256	25.25	0.7546	–	–
	A^{2}F-L [159]	1,374K	77.2G	32.32	0.8964	28.67	0.7839	27.62	0.7379	26.32	0.7931	30.72	0.9115
	ESRT [106]	751K	–	32.19	0.8947	28.69	0.7833	27.69	0.7379	26.39	0.7962	30.75	0.9100
	SwinIR [93]	897K	49.6G	32.44	0.8976	28.77	0.7858	27.69	0.7406	26.47	0.7980	30.92	0.9151
	SRFormer (Ours)	958K	41.3G	32.56	0.9018	28.86	0.7884	27.73	0.7429	26.61	0.8013	31.01	0.9168
	SRFormer+ (Ours)	958K	–	**32.62**	**0.9037**	**28.91**	**0.7904**	**27.82**	**0.7441**	**26.68**	**0.8025**	**31.10**	**0.9184**



Img_012 Urban100 HR Bicubic MemNet CARN RFDN_L OISR SwinIR Ours

Img_013 Set14 HR Bicubic MemNet CARN RFDN_L OISR SwinIR Ours

Figure 6.3: Visual results of **BI** degradation model for ×4 scale factor.



Img_098 Urban100 HR Bicubic VDSR IRCNN_G SRMDNF OverNet RDN Ours

Img_109 Manga109 HR Bicubic VDSR IRCNN_G SRMDNF OverNet RDN Ours

Figure 6.4: Visual results of **BD** degradation model for ×4 scale factor.

Table 6.2: Quantitative results with **BD** degradation model. Performance is shown for scale factor ×3. The best and second best results are highlighted in red and blue respectively. SRFormer with self-ensemble results are **Highlighted**.

Methods	Degrad.	Set5		Set14		B100		Urban100		Manga109	
		PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
SRCNN [36]	BD	32.05	0.8944	28.80	0.8074	28.13	0.7736	25.70	0.7770	29.47	0.8924
VDSR [78]	BD	33.25	0.9150	29.46	0.8244	28.57	0.7893	26.61	0.8136	31.06	0.9234
IRCNN_G [185]	BD	33.38	0.9182	29.63	0.8281	28.65	0.7922	26.77	0.8154	31.15	0.9245
IRCNN_C [185]	BD	29.55	0.8246	27.33	0.7135	26.46	0.6572	24.89	0.7172	28.68	0.7701
SRMDNF [186]	BD	34.09	0.9242	30.11	0.8364	28.98	0.8009	27.50	0.8370	32.97	0.9391
RDN [192]	BD	34.57	0.9280	30.53	0.8447	29.23	0.8079	28.46	0.8581	33.97	0.9465
OverNet[14]	BD	34.59	0.9287	30.46	0.8310	29.13	0.8060	28.24	0.8485	–	–
CASGCN [174]	BD	34.62	0.9283	30.60	0.8458	29.30	0.8196	28.68	0.8611	34.27	0.9476
SRFormer (Ours)	BD	34.78	0.9306	30.76	0.8487	29.45	0.8215	28.79	0.8635	34.41	0.9505
SRFormer+ (Ours)	BD	**34.82**	**0.9316**	**31.83**	**0.8498**	**29.52**	**0.8238**	**28.84**	**0.8683**	**34.46**	**0.9517**

Table 6.3: Quantitative results with **DN** degradation models. Performance is shown for scale factor ×3. The best and second best results are highlighted in red and blue respectively.SRFormer with self-ensemble results are **Highlighted**.

Methods	Degrad.	Set5		Set14		B100		Urban100		Manga109	
		PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
SRCNN [36]	DN	25.01	0.6950	23.78	0.5898	23.76	0.5538	21.19	0.5737	23.75	0.7148
VDSR [78]	DN	25.20	0.7183	24.00	0.6112	24.00	0.5749	22.22	0.6096	24.20	0.7525
IRCNN_G [185]	DN	25.70	0.7379	24.45	0.6305	24.28	0.5900	22.90	0.6429	24.88	0.7765
IRCNN_C [185]	DN	26.18	0.7430	24.68	0.6300	24.52	0.5850	22.63	0.6205	24.74	0.7701
SRMDNF [186]	DN	27.74	0.8026	26.13	0.6924	25.64	0.6495	24.28	0.7092	26.72	0.8590
RDN [192]	DN	28.46	0.8151	26.60	0.7101	25.96	0.6573	24.92	0.7362	28.00	0.8590
OverNet[14]	DN	28.49	0.8200	26.62	0.7116	25.95	0.6602	24.93	0.7365	–	–
SRFormer (Ours)	DN	28.62	0.8225	26.78	0.7129	26.12	0.6621	25.11	0.7384	28.17	0.8616
SRFormer+ (Ours)	DN	**28.66**	**0.8233**	**26.84**	**0.7137**	**26.20**	**0.6632**	**25.18**	**0.7391**	**28.23**	**0.8621**

6.3 Experimental Results

6.3.1 Setting

Datasets

following prior works [31, 100], $DIV2K$ dataset has been used for training and validating the network. $DIV2K$ splits to 800 high-quality images for training phase, 100 validation images and 100 test images. SRFormer train with all training images and use validation images in the training phase. To evaluate the proposed method, five standard benchmark datasets have been used, namely, $Set5$ [15], $Set14$ [183], $B100$ [114], $Urban100$ [66], $Manga109$ [115].

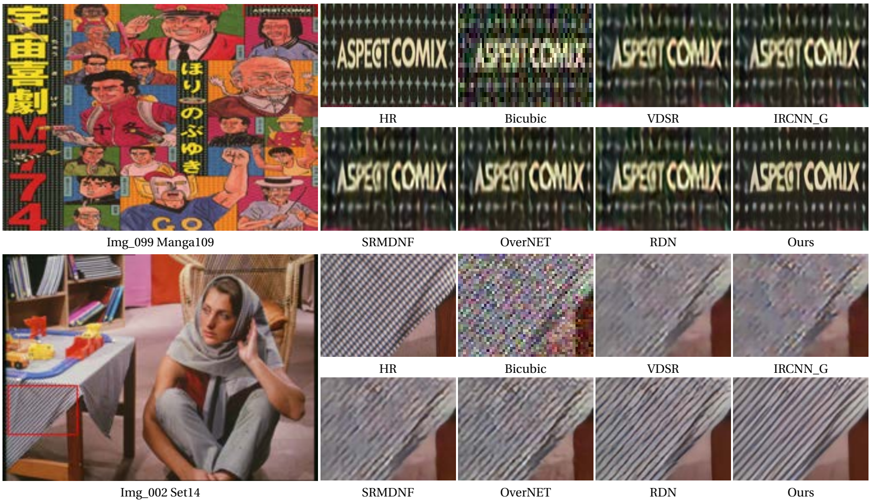


Figure 6.5: Visual results of **DN** degradation model for ×4 scale factor.

Evaluation Protocol

Two widely used quantitative metrics have been considered to measure the performance of our SRFormer in order to maintain consistency with previous works. Peak Signal-to-Noise Ratio (PSNR) measured in deciBels (dB) and the Structural Similarity index (SSIM), which is computed between generated SR images and the corresponding ground truths. Keeping up with the SR community, the RGB reconstruction results first transformed to $YCbCr$ space, and then just the luminance channel is considered to compute the PSNR and SSIM in our experiments.

Degradation Models

In order to demonstrate the efficiency of the proposed model, following the work of [192], three different degradation models created to simulate LR images and make fair comparisons with available methods. Degradation data was obtained as follow: Firstly, a bicubic (BI) down-sampling dataset with scaling factors [×2, ×3, ×4] has been created. Secondly, Blur-Downsampled (BD) has been created by applying Gaussian kernel 7×7, and $\sigma = 1.6$ to HR images and then downsampled images with scaling factor ×3. Aside from the BD, a more challenging degradation model has been created, referred to as Downsample-Noisy (DN). DN degradation model is down-sampling HR images with bicubic followed by adding 30% Gaussian noise.

Implementation Details.

In the training phase, RGB patches are provided as inputs with size of 64 × 64 from each of the randomly selected 32 low quality training images. Data augmentation applied on patches by means of horizontal random flips and 90 degree rotation. AdamP [59] optimizer has been employed with the initial learning rate 10^{-3} and its halved every 4×10^5 steps. $L1$ is used as loss function to optimize the model. Also, the configurations of our transformer encoder is as follow, we used 4 Transformer blocks within 6 Transformer layers for each block, Embedding dimension set to 64 and MLP ratio to 2 for all Transformer blocks. Also, a Conv1 × 1 is used inside each Transformer blocks. SRFormer developed by using PyTorch framework and trained on a single NVIDIA RTX 3090 GPU to achieve its performance.

6.3.2 Comparison with State-Of-The-Art Methods

In this section, SRFormer and SRFormer+ are compared to other lightweight state-of-the-art SR methods. Self-ensemble method [152] is also used to further boost the performance of the proposed SRFormer (denoted as SRFormer+).

Results on Bicubic Degradation

We present comparisons between the proposed method (SRFormer and SRFormer+) and several of the most recent lightweight SOTA CNN and Transformer based models: VDSR [78], DRCN [79], CARN [1], CBPN [194], FALSR [30], LAPAR-A [91], LatticeNet[109], MADNet [85], HDRN [74], DPN [94], A^2F[159], ESRT [106], and SwinIR[93] on the Bicubic (BI) degradation model for scale factors [×2, ×3, ×4]. Also, the number of network parameters and Multi-Adds operations are presented in Table 6.1 to demonstrate the complexity of the model and have a fair comparison with the existing methods. As can be seen, SRFormer produces superior outcomes in practically all circumstances when compared to the other methods mentioned above. This shows that SRFormer is capable of continuously accumulating these hierarchical characteristics to build more robust representative features that are well-focused on spatial context information. This trait can be confirmed by the obtained SSIM scores, which are based on the visible structures in the image and are therefore more accurate. Furthermore, it can be observed that using self-ensembles [152], the proposed SRFormer+ gains even more performance benefits. Several visual outcomes are presented in Fig. 6.3. As can be seen, the texture direction of the reconstructed images from all of the compared approaches is utterly incorrect while the text is blur in all the cases at different levels. However, the results obtained by SRFormer are similar to ground truth texture.

6.3.3 Results on BD and DN Degradation Models

We also provide the performance of SRFormer and SRFormer+ on the BD (Blurry) and DN (Noisy) benchmark datasets in Table 6.2 and Table 6.3 to illustrate the strengths of proposed model when it comes to challenging situation with SOTA models. Due to degradation mismatch the following methods SRCNN and VDSR are re-trained for both BD and DN. As can be seen, SRFormer outperforms all other lightweight SOTA models on challenging benchmark datasets, and it is particularly impressive when compared to other lightweight SOTA models. A high-capability model, RDN [192] is also listed, which is used to demonstrate the superior performance of our SRFormer in comparison to a deep and costly model in these challenging datasets. SRFormer performs better in both datasets notwithstanding, RDN is a significantly expensive network compared to the low-cost SRFormer. RDN is nearly ×20 more expensive in terms of computational complexity. Furthermore, a visual representation of both challenging BD and DN benchmark datasets is shown in Fig. 6.4 and Fig. 6.5 respectively. As can be seen our proposed method performs better in comparison with other SOTA methods in removing the noises and fuzzy regions from input image, which results generating a sharper with fine details SR images.

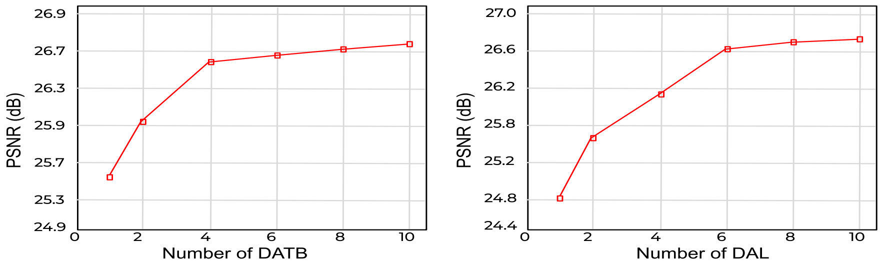


Figure 6.6: Performance investigation on different settings of SRFormer on Urban100 for scale factor ×4 .

6.4 Ablation Study

The performance of the propose model is further investigated through an extensive ablation study that includes in-depth examinations of impact of each module. The ablation study is designed to provide additional insight into the performance of the proposed model.

6.4.1 Relation Between Number of Transformer Blocks and Layers vs. Performance and Network Parameters

We investigate deeply the relation of number of Transformer blocks (DATB) and Transformer layer (DAL) on the performance and model size of our proposed model architecture in Fig 6.6. We discovered that the performance (PSNR) of the network has a positive relationship with aforementioned hyperparameters however performance gains by increasing the number of blocks and layers will not come for free. By increasing the number of blocks or layers while performance continuously improve, the overall number of network parameters and FLOPs increases, which makes the network computationally inefficient. Also, we can see that by increasing these hyperparameters, the performance benefit gets more and more limited until it is start to saturate progressively. Thus, we design our network by choosing four Transformer blocks and six Transformer layers inside of each block to still have a lightweight yet powerful feature extraction module.

6.4.2 Visualization on Influence of Conv Layer in Transformer Block

Figure 6.7 shows the average feature maps of each stage of our Dense Feature Extraction module to investigate the impact of conv layer when it stack up with Transformer layers. Each average feature map is the mean of F_{out} in channel dimension, which represents the output of Transformer block at each stage. The average feature maps without a conv layer are shown on the top row, and with a conv layer within Transformer blocks are illustrated on the bottom row. By visualizing the feature maps, we can first see that, using a conv layer within a Transformer, helps the Transformer to learn sharper representations compared to without conv layer. Second, as the network focuses more on high level information, feature maps tend to include more negative values at each stage, indicating a stronger impact of suppressing the smooth area of the input image, which further leads to a more accurate residual image.

6.4.3 Impact of Dual Attention

We further study the impact of both proposed SpAM and SeAM to illustrate the effectiveness of the proposed Dual Attention. We investigate the performance of SRFormer with the standard self-attention layer [156] and each sub branch of our Dual Attention layer. As can be seen in Table 6.4, the SRFormer with Dual Attention boosts the performance of the network while using less computational cost compared to when standard self-attention layer replace in the network. In contrast

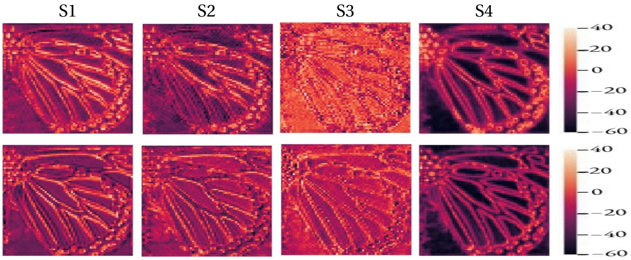


Figure 6.7: Average feature maps of Transformer blocks (DATB). Top: DATB without Conv layer. Bottom: DATB with Conv layer

to other self-attention layer, Dual Attention built up with two parallel branches, which able to encode the spatial information more efficient and enables the Dual Attention to preserve a rich representation while shrinking its depth to make further computation lightweight. Also, it helps the network train faster compare to other transformer based network.

Table 6.4: Influence of different setting of the Dual Attention layer on Urban100 scale factor ×4.

	SeAM	SpAM	Parameters(K)	PSNR(dB)↑
Meta-Former	–	–	953K	26.47
Dual Attention	–	✓	955K	26.38
	✓	–	942K	26.54
	✓	✓	**958K**	**26.61**

6.4.4 Influence of Gated MLP Feature Fusion

Table 6.5 shows the impact of our proposed lightweight Gated MLP Feature Fusion compared to without and with baseline MLP on the performance of proposed network. In addition, we investigate the impact of the usage of depthwise, pointwise conv layer, and gated mechanism in our Gated MLP Feature Fusion. As can be seen, SRFormer obtains performance gain compared to when the network does not contain any MLP module or even when it is compared to the baseline MLP with a less computation cost. The intuition behind that is, GMFF uses gated mechanism to allow gradients to backpropagate more easily through depth, and a Dw-Conv layer between the MLP layers to leak the location information, which lead the

Table 6.5: Gated MLP Feature Fusion performance investigation on Urban100 for ×4.

	Parameters(K)	Memory(M)	PSNR(dB)↑
w/o MLP	955K	2,631	26.52
Baseline MLP	962K	2,875	26.58
GMFF(Ours)	958K	2,739	**26.61**

Table 6.6: Impact of different Gated MLP Feature Fusion setting on Urban100 for ×4.

	PwConv	DwConv	Gated Mech.	Parameters (K)	PSNR (dB) ↑
GMFF	✓	–	–	959K	26.56
	–	✓	–	960K	26.59
	–	✓	✓	**958K**	**26.61**

Table 6.7: Perceptual index comparison between proposed method and recent lightweight state-of-the-art methods on benchmark datasets for scale factor ×4. The lower is better.

Methods	Parameters	Set5	Set14	B100	Urban100	Manga109
CARN[1]	1.5M	6.297	5.775	5.700	5.540	5.132
SRFBN-S[78]	0.6M	6.451	5.775	5.702	5.549	5.010
SRDenseNet[154]	2M	6.128	5.615	5.653	5.526	4.762
RFDN_L[100]	0.6M	6.124	5.644	5.659	5.531	4.810
A^2F_L[159]	1.3M	6.084	5.499	5.532	5.179	4.771
SRFormer(Ours)	0.9M	**4.931**	**4.821**	**4.474**	**4.649**	**3.880**

network to pay attention on positional information unlike the baseline MLP that uses positional encoding [38] to introduce the location information, which is not suitable when the test resolution is different from training resolution. Furthermore, we illustrate the performance gain of our Gated MLP Feature Fusion with pointwise, depthwise convolution layers, Gated Mechanism, and without GMFF. As shown in Table 6.6, the performance of our SRFormer boosts when depthwise convolution layer with gated mechanism used compare to other setting.

Table 6.8: Average running time (s) and memory consumption (MB) comparison on Urban100 for ×4.

Methods	Parameters	Memory	Running Time(s)↓	PSNR(dB)↑
CARN[1]	1.5M	1,116	0.032	26.07
SRFBN-S[92]	0.5M	2,154	0.031	25.71
SRDenseNet[154]	2M	5,531	0.221	26.05
RFDN-L[100]	0.6M	3,215	0.033	26.22
A^2F-L[159]	1.3M	3,015	0.032	26.32
RCAN[191]	16M	1,531	0.297	26.82
EDSR[95]	43M	2,731	0.085	26.64
SAN[92]	16M	3,015	0.224	26.79
RDN[192]	23M	5,015	0.172	26.82
SwinIR[93]	0.9M	3,340	0.216	26.47
SRFormer(Ours)	0.9M	2,739	0.102	**26.61**

6.4.5 Perceptual Index Metric

In order to assess the quality of the generated super resolution images, Perceptual Index (PI) is used, which is more accurate in reflecting human perceptions of image quality compared to other metrics (PSNR and SSIM). Table 6.7 illustrates the PI metric between SRFormer and SOTA methods with same order of magnitude in terms of network model size. It can be seen that the proposed model achieves lower results (lower is better) compared to other models. This demonstrates the ability of the proposed SRFormer for generating more realistic images.

6.4.6 Model Complexity and Inference Time Analysis

Table 6.8 illustrates the advantages of the propose SRFormer architecture in terms of Network Parameters (M) Inference Time (s) and Memory Consumption (MB) compared to existed light- and heavy-weight SOTA CNN and Transformer base architecture methods on Urban100. In order to make a fair comparison, all the models are measured with same configuration with their published source code and default hyper-parameters on a single NVIDIA RTX 3090 GPU. As shown, our model has the shortest inference time and less memory hunger per image compared to Transformer models. This comparison illustrates that our proposed model successfully strike a balance between performance and running time requirements.

6.5 Summary

In this paper we present a novel and efficient Transformer architecture based network called SRFormer. The proposed model is designed by using strength of both

Convolutional and Transformer layers to extract and preserve the fine details of the features while while remains memory efficient. To do so, we introduce Dual Attention layer, a Transformer layer, which generate the global attention map from two different branches (SpAM and SeAM) in order to capture both local context information and global dependency between sequences. Also, we introduce a lightweight Gated MLP Feature Fusion to aggregate the multi stage feature representation by focusing on inner spatial information before upsampling module. We demonstrate the efficiency of the proposed method through a series of ablation investigations. We have empirically demonstrated that our approach outperforms previous lightweight state-of-the-art methods on all benchmarks datasets, despite having a similar or fewer number of network parameters.

7 TnTViT: Transformer in Transformer Network for Guidance Super Resolution

This chapter presents the article submitted at:

Armin Mehri, Parichehr Behjati, and Angel D.Sappa. TnTViT: Transformer in Transformer Network for Guidance Super Resolution. IEEE Access, 2022. (under review)

Image Super Resolution is a potential approach that can improve the image quality of low-resolution optical sensors, which can lead to improved performance in a wide variety of industrial applications. It is important to emphasize that the most SOTA super resolution algorithms are often using a single channel of input data for training and inference. However, this practice ignores the fact that the cost of acquiring high-resolution images in various spectral domains can differ a lot from one another. In this paper, we seek to exploit complementary information from a low-cost channel (visible image) to increase the image quality of an expensive channel (infrared image), which nowadays infrared images become increasingly necessary in different sectors such as in visual surveillance, military, and security sectors. We propose a dual stream Transformer based SR approach to use the visible image as a guidance to super resolution of other spectral band images. To this end, we introduce Transformer in Transformer (TnTViT) an efficient and effective method that extracts the feature of each input images via different streams and fuse them together at various stages. Furthermore, it is worth to mention that, unlike other guidance SR approaches, TnTViT is able to generate SR images in arbitrary sizes. Extensive experiments on different datasets illustrate the advantage of proposed model compared to other state-of-the-art SR approaches.

7.1 Motivation

In recent years, image super resolution, has achieved a significant interest from both academic and industrial communities. The process of reconstructing a high

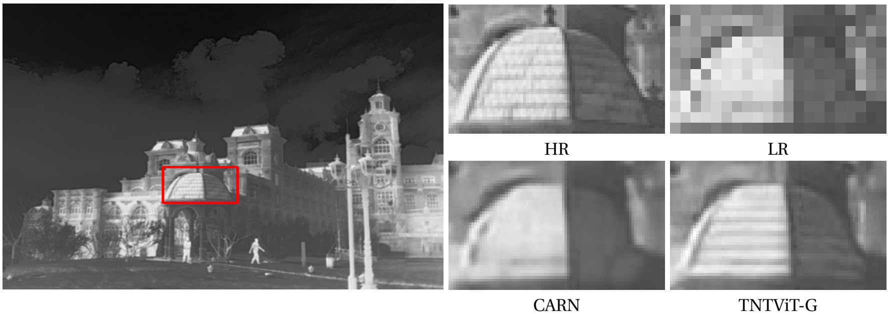


Figure 7.1: The visual comparison of super resolution results on M3FD dataset for scale factor ×4.

resolution (HR) image from its low resolution (LR) counterpart is referred to as the super resolution (SR) problem in the field of image processing. Due to the fact that a single LR image might have numerous mappings from LR to SR, SR is an ill-pose problem, which also known as a one-to-many. Thus, numerous SR methods have been introduced to reconstruct a high resolution image from its low resolution ones, such as traditional approaches like self-exemplars approach [66], anchoring neighborhood regression [151], sparse representation [173] and random forest [135].

More recently, by advancing the deep learning approaches, several Convolutional Neural Networks (CNNs) and Transformers networks are being used as a solution for the ill-pose SR problem. This is largely attributable to the recent successes of deep learning approaches in a variety of vision tasks, such as object detection, image recognition, semantic segmentation, image classification, and many others. The first work in this direction has been presented by Dong et al. [35], which developed a three-layer CNN model to train a nonlinear LR-to-HR mapping function, called SRCNN, which greatly outperforms the traditional machine learning-based methods. The majority of later expansions of SRCNN enhance SR accuracy by employing more complicated network designs (such as RDN [192], EDSR [192], RCAN [191], among others) or by utilizing a training dataset with better quality.

However, in real-world applications, the environment around us is dynamic and changing all the time due to many known and unknown reasons, which requires to deal with the various challenging conditions such as rain, fog, occlusions, poor lighting, low resolution, and many others. All these factors make difficult for an algorithm that uses only visible-band sensor (RGB) to achieve high performance under these conditions [127]. Therefore visible image is found to be insufficient for such a cases and cross-spectral images have become increasingly necessary in many applications as they are robust against obstacles in visual environments and provide support to the RGB images. Cross-spectral images (e.g., visible-thermal infrared; visible-near infrared) have been used in many range of specialized fields such as surveillance [166], military affairs [44], pedestrian tracking [10], firefighting [9] and many others. However, their associated costs of having such images (infrared images) grow significantly with the increase of their resolution.

Various approaches and algorithms have been proposed to improve the resolution of different infrared images using hardware or software. Employing low resolution cameras that are less expensive than high-end cross spectral domains cameras and using SR methods to increase the resolution of such images, is one strategy to boost the consumer applicability of such cameras to deal with challenging situation at a lower cost. However, as previously stated, single image SR (SISR) is a tricky operation that becomes even more difficult when the input image has a very poor resolution (such as the ones produced by inexpensive cross/multi-spectral sensors), which SISR techniques may hallucinate missing details from low resolution inputs and therefore yielding to artifacts [3].

To address the aforementioned problems, a fundamental solution is to take advantage of any additional information that can be found with the low resolution infrared images since most cross-spectral cameras are accompanied with an inbuilt visible RGB camera with higher resolution. As a result, it is permissible to use low-cost visible images as additional information to considerably improve the accuracy of SR results of the costly infrared images. For example, long-wave infrared (LWIR) detectors, required to capture thermal images, are sealed inside of their own separate vacuum packages in order to carry out high-precision thermal measurement, which is a procedure that is both time consuming and costly [132]. As the result, the cost of LWIR sensors is much higher than that of RGB ones with comparable spatial resolutions. The majority of commercially available LWIR cameras capture LR images (for example, 160x120 or even 80x60 pixels) [20], in which significant information are severely lost.

In this paper, we attempt to boost the performance of image restoration in the expensive channel by taking into account the complementary information captured by additional low cost visible sensor. The primary focus of this work is to build a deep learning model that applies multimodal sensor fusion using visible cross-spectral images—the proposed approach is evaluated with two different schemes (i.e., visible-thermal infrared (LWIR), and visible-near infrared (NIR)) but is is also valid for any other input data. The proposed model accept two images as inputs to integrate them in such a way that enhance the generated infrared image resolution with fine detailed by help of corresponding visible image. Thus, a guidance super

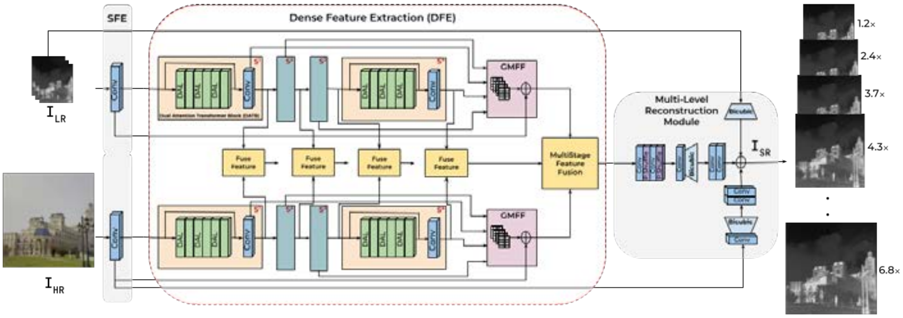


Figure 7.2: The overall network architecture of the proposed TnTViT-G.

resolution network ($TnTViT - G$) is proposed, to enhance the LR infrared image by integrating the rich information in the HR visual image. We show that HR visual images can help the model fill the missed values and generate higher frequency details in the reconstructed SR infrared image.

The main contributions of our work can be summarized as follow:

- We present TnTViT-G, an efficient dual stream Transformer based network for guidance super resolution (GSR) task. TnTViT-G Transformer blocks built on the top of the idea of vision Transformer with completely revising the self-attention layer.

- We present a lightweight Dual Attention layer, which significantly improves the reconstruction quality by generating global attention map from two local attention weights, which obtain individually by two branches in parallel while its not memory hunger.

- We present a high quality arbitrary upsampling module, which able to generate SR images in any scale factors.

- Extensive experiments show that TnTViT-G achieves state-of-the-arts on various datasets for GSR task compared to CNN/Transformer based networks.

The rest of the paper is organized as follows: Section 7.2 describes the proposed TnTViT-G and its core components in detail. Experimental comparisons against several state of-the-art methods are presented in Section 7.3. Model investigation presents in section 7.4. Section 7.5 concludes the paper.

7.2 Proposed Method

In this section, the overall network architecture of the proposed TnTViT-G is described. Later, the detailed information of Dual Attention layer is provided. TnTViT is designed for Single Image Super Resolution and TnTViT-G is siamese based network of TnTViT, which designs for guidance super resolution.

7.2.1 Overall Pipeline

The main objective of the proposed model is to design an efficient Transformer-based network for Guidance Super Resolution (GSR) that is capable of producing fine details high-quality images with the help of the guided images (e.g, visible images) to boost the performance of the network while staying computationally low. Thus, we employ the original Transformer structure but modify it in a way that the model achieves to a considerable performance over existing CNN and Transformer networks. The overall architecture of the TnTViT-G is illustrated in Fig. 7.2, which consists of two streams to extract the features of LR infrared input images and HR visible images. In particular, the proposed TnTViT-G consists of three modules: Shallow Feature Extraction (SFE), Dense Feature Extraction (DFE) and Multi-Level Reconstruction Modules. We defined I_{LR}^{IR}, I_{HR}^{Vis}, and I_{SR}^{IR} as the low-quality infrared, high-quality RGB inputs, and high quality output of our network, respectively.

7.2.2 Shallow Feature Extraction

Given the input images to the network, we apply a single 3 × 3 convolutional layer on each network's streams to the provided LR and HR visible inputs in order to map the input images space to a higher dimensional feature space and increase the performance of the network [168]. Therefore, we extract the shallow features as follow:

$$\boldsymbol{F}_0^{IR} = Conv_{3\times3}(\boldsymbol{I}_{LR}^{IR}), \boldsymbol{F}_0^{Vis} = Conv_{3\times3}(\boldsymbol{I}_{HR}^{Vis}), \tag{7.1}$$

where $F_0^{IR}(.)$ and $F_0^{Vis}(.)$ denotes the output of shallow feature extraction on both infrared and visible images.

7.2.3 Dense Feature Extraction

After mapping the inputs to a higher dimensional feature space, the features pass through the Dense Feature Extraction F_{DFE} to encode the information in order to understand the context of the sequences. The feature encoders of the proposed approach (i.e, Dense Feature Extraction) is a Transformer based network, which shares between the both input images (I_{LR}^{IR} and I_{HR}^{Vis}) to keep the network computationally efficient. However, each stream receive the same patch of input image with different sizes since LR images are relatively smaller than visible images. Particularly, Dense Feature Extraction design by using several Transformer blocks to extract abstract features and spotlights the high-level information. Each Transformer block consists of several Transformer layers and a 1 × 1 Conv layer with benefit of cascade connections to transfer the information from previous stage to current stage and help the gradient flow of the network. Thus, we extract the feature as follow:

$$\boldsymbol{F}_{DFE} = H_{DFE}(F_0^{IR}; F_0^{Vis}), \tag{7.2}$$

where $H_{DFE}(.)$ is Dense Feature Extraction with several Transformer blocks which can be formulated as

$$\boldsymbol{F}_i = Conv_{1\times1}(C[H_{DATB}(F_{i-1}), X_{i-1}], i = 1, 2, ..., K, \tag{7.3}$$

where $H_{DATB}(.)$ denotes the i_{th} Dual Attention Transformer Blocks. C stands for the concatenation operation between initial and output features of each $DATB$ block. $Conv$ denotes the convolutional layer after concat operation within each DATB. By using a convolutional layer in the Transformer block, help to transfer inductive bias from the convolution operation into the Transformer network and provide a more solid foundation for the later aggregation with shallow features.

After encoding the features through several DATB, the output feature maps of each DATB stage are concatenated together to highlight the positional information via GMFF module, which stands for Gated MLP Feature Fusion before reconstructing the SR images. GMFF module is shown in Fig 7.3, which design to generate a multi-stage representation feature map of Transformer blocks. Later, the feature map passes through a lightweight MLP network. However, unlike to standard MLP network, the GMFF´s MLP module is designed by using a 3 × 3 depthwise Conv layer and gating mechanism technique to first, leak the spatial information since highlighting such features are important in SR tasks to achieve high performance. Second, allowing the useful information pass through the network and suppress the less informative ones. Gating mechanism used by applying element wise product of two parallel routes of linear transformation layer that one is activated with the GELU. Thus, Gated MLP Feature Fusion can be seen as follow:

$$\boldsymbol{F}_{GMFF} = MLP(GELU(Conv_{3\times3}(MLP(F_i)))) + F_0, \tag{7.4}$$

where F_{GMFF} is the output of DFE with aggregation of the initial features, which later used by Multi Stage Feature Fusion Module.

7.2.4 Multi Stage Feature Fusion Module

After encoding the information of both LR infrared image I_{LR}^{IR} and HR visible image I_{HR}^{Vis} with a dual stream shared network, the LR features first scale up to the same spatial size of HR visible image before fusing the information, with a learnable bicubic upsampling that contains a conv layer before it; later, the aggregated features of all the stages are concatenating together to enhance the infrared LR images before upsampling it to the desired output size.

$$\begin{aligned} &\boldsymbol{F}_{MSFF} = \\ &Conv_{1\times1}(C[H_{UP}(F_{GMFF}^{IR}), F_{GMFF}^{Vis}, FF_{S1}, ..., FF_{S4}]), \end{aligned} \tag{7.5}$$

where $F_{MSFF}(\cdot)$ denotes the output of multi-stage feature fusion module of both TnTViT streams and the feature fusion of each stage.

7.2.5 Multi Level Reconstruction Module

Later, to upsample the LR infrared image after fusing the information, we propose a new inductive bias in GSR architectures to generate SR images more accurately with less artifacts compared with the other methods or naive interpolations techniques. To do so, we first pass the information through two pixel shuffle layers and a conv layer before each of them. Second, the upsampled features with pixel shuffle layers feed to a learnable bicubic interpolation to up scale the features to any arbitrary sizes. Later, the information aggregate with a shallow features of HR guided image, which also up scaled with learnable bicubic interpolation and directly up scaled feature of LR thermal image to grant access to the original LR information.

$$\boldsymbol{I}_{SR} = H_{Rec}^{\uparrow}(H_{UP}(F_0^{IR}) + F_{MSFF} + H_{UP}(I_{LR})), \tag{7.6}$$

where $H_{Rec}(\cdot)$ and I_{SR} denotes the up-sampling module and high quality reconstructed image respectively. Hence, the proposed module can learn how to refine the pixels more correctly via different level of up scaling to bring it closer to the actual high-resolution counterpart and beyond. The extensive experiments have been detailed on ablation study to show the efficiency of proposed reconstruct module over other approaches.

7.2.6 Loss Function

To keep the consistency with previous works, we use L_1 loss as a cost function during training to optimize the parameters of proposed TnTViT.

$$L_1(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N} \|I_{SR} - I_{HR}\|_1, \tag{7.7}$$

where I_{SR} obtained by taking low-quality infrared image as the input of our model and I_{HR} is the corresponding ground-truth.

In the next subsections, we provide more details about our Transformer layer.

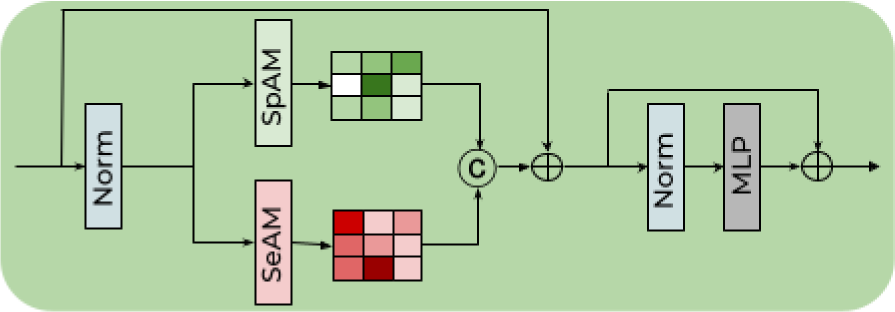


Figure 7.3: Illustration of Dual Attention Layer (DAL).

7.2.7 Dual Attention Layer

This section presents the proposed Dual Attention layer, an architecture abstracted from general multi-head Transformer layer [156] with revising the self-attention layer. As is generally known, self-attention is critical to achieve excellent performance in Transformer-based networks. However, self-attention might be troublesome for a variety of reasons. For example, the computational complexity of self-attention grows quadratically with the number of tokens to mix. Also, self-attention does not take into account the local contextual information due to nature of self-attention and treats the images as flatten sequences which ignore the structure of the image. Thus, we propose the Dual Attention layer to address the mentioned limitations by constructing a global attention map at a lower computational cost. Dual Attention layer creates a global attention map by combining two local attention maps, which are obtained in parallel by using a CNN-based Attention Module and a Transformer self-attention layer. Unlike the prior token mixer, the Dual Attention able to take to the account both long-range dependency and local contextual information with less computing cost.

As shown in Fig 7.3, we design our Dual Attention such that the channel information is distributed evenly across both attention module branches (SpAM and SeAM). Both attention branches get half of the input tensor from Norm layer tensor X to generate the local attention map independently. SeAM is a self-attention Transformer, which first generates the query (Q), key (K), and value (V) projections enriched with local context. Inspired by [180], we apply SeAM only across the channels rather than spatial dimensions. Our SeAM uses only depth-wise convolutions to emphasize the channel-wise spatial context before computing feature covariance to produce the attention map. Thus, Q, K, V are computed as:

$$Q = W_d^Q Y, K = W_d^K Y, V = W_d^V Y, \tag{7.8}$$

where $W_d^{(\cdot)}$ is the 3×3 depth-wise convolution. Next, query and key projections reshape in a way that their dot-product interaction generates a transposed-attention map. Thus, attention map generate as follow:

$$Attention(Q,K,V) = W_d(V.Softmax(K.Q/\alpha)) + X, \tag{7.9}$$

where X is the input feature map and α is a learnable scaling parameter that is used to regulate the magnitude of the dot product of K and Q before applying the Softmax function. Similar to previous works [93, 156, 177] we perform the attention function for h times to learn separate attention maps in parallel in our SeAM module.

Second branch of our Dual Attention layer is Spatial Attention Module (SpAM), which is an almost parameter free attention mechanism. SpAM receive the other half of the input tensor to generate the local attention map. The goal of SpAM module is to encode the spatial information, which represents the importance of each pixel in the input feature with negotiable cost. Given half of the input tensor information, the channels of the input tensor are reduced by mean and max operations, of which the shape is $1 \times H \times W$. The obtained features concatenated, then pass through a dilated convolution layer with kernel size of 3×3. After, a sigmoid activation layer apply to the output feature to generate the attention weights of shape $1 \times H \times W$, which are later multiply with the input tensor to refined tensors of shape $C \times H \times W$. Thus, the SpAM can be formulated as follow:

$$X = Sigmoid(Conv_{3\times3}[F_{Mean}(X), F_{Max}(X)]) * X, \tag{7.10}$$

where $F_{Mean}(\cdot)$ and $F_{Max}(\cdot)$ denotes for mean and max operations. Later, both generated local attention maps from SpAM and SeAM are concatenated together to obtain a unify global attention map with less computational cost. Thus, the generated attention map contains both long range dependency and local context information with enrich of spatial features.

Following that, a multi-layer perceptron (MLP) with two fully connected layers and a GELU non-linearity activation function between them is employed for further feature modifications. The norm layer is also added before MLP, and both modules

Table 7.1: Average PSNR, SSIM comparisons with SOTA CNN- and Transformer-based methods with the same range of network parameters on the Bicubic (**BI**) degradation for scale factors [×2, ×4, ×8]. Best results are **highlighted**.

Scale	Method	DM	G/S	M3FD		RGB-NIR	
				PSNR↑	SSIM↑	PSNR↑	SSIM↑
×2	Bicubic	BI	Single	37.74	0.9465	31.91	0.8792
×2	CARN [1]	BI	Single	38.82	0.9538	33.05	0.8982
×2	SwinIR [93]	BI	Single	37.89	0.9498	31.84	0.8863
×2	TNTViT [OURS]	BI	Single	38.91	0.9542	33.14	0.9002
×2	TNTViT-G [OURS]	BI	Guided	**39.01**	**0.9556**	**34.49**	**0.9152**
×4	Bicubic	BI	Single	30.79	0.8435	26.63	0.7129
×4	CARN [1]	BI	Single	31.58	0.8336	27.33	0.7284
×4	SwinIR [93]	BI	Single	30.82	0.8457	26.12	0.7177
×4	TNTViT [OURS]	BI	Single	31.64	0.8646	27.40	0.7395
×4	TNTViT-G [OURS]	BI	Guided	**32.00**	**0.8735**	**29.59**	**0.8252**
×8	Bicubic	BI	Single	26.77	0.7594	24.10	0.6142
×8	CARN [1]	BI	Single	27.41	0.7787	24.79	0.6348
×8	SwinIR [93]	BI	Single	26.81	0.7621	24.18	0.6153
×8	TNTViT [OURS]	BI	Single	27.50	0.7607	24.90	0.6348
×8	TNTViT-G [OURS]	BI	Guided	**27.88**	**0.7628**	**26.21**	**0.7835**

contains the residual connection between them. Thus, the entire procedure inside of our Dual Attention is as follow:

$$\begin{aligned} X &= (Norm(SpAM(X/2) + SeAM(X/2))) + X \\ Y &= MLP(Norm(X)) + X \end{aligned} \tag{7.11}$$

where $Norm(\cdot)$ stands for normalization layer and Y for the output feature map.

7.3 Experimental Results

7.3.1 Setting

Datasets

Two datasets have been used to perform the experiments, namely M3FD[101] and RGB-NIR [19]. The first dataset is M3FD, which newly released by [101]. The M3FD dataset contains pair of visible and thermal images. The dataset built with a synchronized system of one binocular optical camera and one binocular thermal sensor to capture corresponding two modality images. We use M3FD Fusion dataset which consists of 300 aligned pair images from different scenarios in Daytime, Night and Overcast. Also, The dataset consists of images from different scenes such as road, campus, street, forest, and many others.

The second dataset is RGB-NIR Scene [18] dataset. The RGB-NIR Scene dataset contains aligned pair of 477 RGB and near-infrared images, divided into 9 categories such as country, field, forest, indoor, mountain, old building, street, urban, and water. The images were acquired by utilizing different exposures from customized SLR cameras equipped with visible and near-infrared filters.

Evaluation Protocol

Two widely used quantitative metrics have been considered to measure the performance of our TnTViT compared to other approaches. We used Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) to measure reconstructed SR image accuracy. PSNR assesses the image by statistically calculating distortion levels between the reconstructed and ground-truth images. SSIM measures the structural similarity between two images based on luminance, contrast, and structure, which has a value range between [0-1]. Higher value, better for both PSNR and SSIM.

Degradation models

Degradation models have been created to replicate LR images in order to demonstrate the effectiveness of our proposed approach. The degradation model is bicubic downsampling (BI), which simulates LR images with the scale factors [×2, ×4, ×8] by applying bicubic downsampling to HR images.

Implementation Details

We randomly select 70%, 20%, and 10% of images of each datasets for the training, validation and test phases respectively. In the training phase, we provide the image patches as inputs with different sizes based on the size of each datasets from LR images and corresponding RGB images. The batch size has been set to 32 for the training. Horizontal random flips and 90 degree rotation data augmentation applied on patches of images. Adam optimizer has been used with the initial learning rate 10^{-3}. $L1$ is used as loss function to optimize the model and the network has been trained for $150K$ iterations. Also, the configurations of our transformer encoder is as follow, we use 4 Transformer blocks within 6 Transformer layers for each block, Embedding dimension set to 64 and MLP ratio to 2 for all Transformer blocks. TnTViT-G is developed by using PyTorch framework and trained on single NVIDIA RTX 3090 GPU to achieve its performance.

7.3.2 Comparison with State-Of-The-Art Methods

In this section, we compare our proposed SISR (TnTViT) and GSR (TnTViT-G) with other lightweight state of the art approaches on different datasets with different scale factors.

Experiments on Bicubic Degradation

Table 7.1 shows comparisons between the proposed approach (TnTViT and TnTViT-G) and SOTA CNN and Transformer based models, CARN [1] and SwinIR [93] on the Bicubic (BI) degradation model for scale factors [×2, ×4, ×8]. Also it is worth to mention that, these networks contain almost the same number of network parameters, allowing for a fair comparison. As can be observed, when the proposed method compared to the approaches mentioned above, TnTViT achieves better results without help of any guided image (Visible image). Furthermore, the proposed method (TnTViT-G) with the guidance of visible image information achieves superior results in all cases with major margin. This demonstrates that TnTViT-G continually accumulate these hierarchical information from different spectral images in order to construct more robust representative features that are well-focused on spatial context information since its the key of accurate SR image. The derived SSIM scores, which are based on the visible structures in the image and hence more accurate, validate this claim. Fig. 7.4 shows some qualitative results on M3FD, and RGB-NIR datasets on different scale factors. As can be seen, TnTViT produce image better than existing method in the SISR since the network is able to focus better on the spatial information. However, TnTViT-G is able to reconstruct high-frequency details significantly better than all the existing methods and generates more accurate SR infrared images, which are more similar to the ground truth images.

Table 7.2: Avargae LPIPS comparison between proposed method and the other methods on benchmark dataset for scale factors [×2, ×4]. The lower is better.

Methods	Scale	M3FD	RGB-NIR
CARN[1]	×2	0.1127	0.1365
SwinIR[93]	×2	0.2076	0.2291
TNTViT (Ours)	×2	0.1013	0.1224
TNTViT-G (Ours)	×2	**0.0916**	**0.0934**
CARN[1]	×4	0.2418	0.3371
SwinIR[93]	×4	0.3176	0.3985
TNTViT (Ours)	×4	0.2322	0.3262
TNTViT-G (Ours)	×4	**0.2119**	**0.2202**

7.4 Ablation Study

The proposed model is further studied by an extended ablation investigation to demonstrate its efficiency. The ablation study is intended to offer further information about the performance of the proposed approach.

7.4.1 Visualization on Impact of Guided Image

Fig. 7.5 shows the average feature maps of each stage of our Dense Feature Extraction module to investigate the impact of guidance image (i.e, visible image) when it stacked up with the LR feature map in each stage of DFE. Each average feature map reflects the output of the Transformer block at each stage in Dense Feature Extraction module. The average feature maps without guidance images are presented on the top row, while those with guidance images are shown on the bottom row. We can observe from the feature maps that using a guidance image helps the network acquire sharper representations. Second, as the network focus more on high-level information, feature maps tend to include more negative values at each stage, showing a greater influence of suppressing the smooth area of the input image, yields to a more accurate SR output.

7.4.2 Influence of Multi-Level Reconstruction Module

We investigate the advantages of using the proposed Multi-Level Reconstruction Module, as well as the impact of two widely upsampling and interpolation approaches on reconstruction results. We carried out following experiments: *i*) Di-

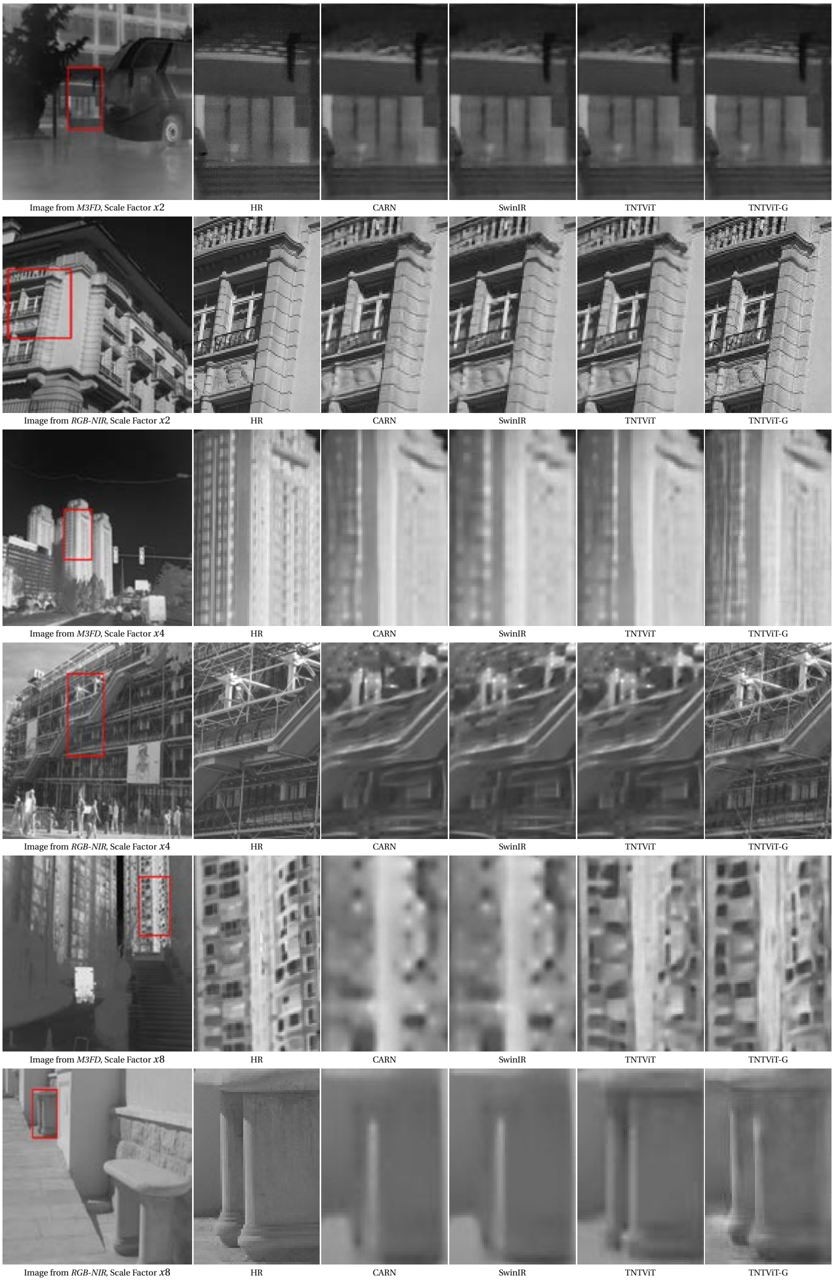


Image from M3FD, Scale Factor x2
HR
CARN
SwinIR
TNTViT
TNTViT-G
Image from RGB-NIR, Scale Factor x2
HR
CARN
SwinIR
TNTViT
TNTViT-G
Image from M3FD, Scale Factor x4
HR
CARN
SwinIR
TNTViT
TNTViT-G
Image from RGB-NIR, Scale Factor x4
HR
CARN
SwinIR
TNTViT
TNTViT-G
Image from M3FD, Scale Factor x8
HR
CARN
SwinIR
TNTViT
TNTViT-G
Image from RGB-NIR, Scale Factor x8
HR
CARN
SwinIR
TNTViT
TNTViT-G

Figure 7.4: Visual results of **BI** degradation model for scale factors [×2, ×4, ×8] on M3FD an RGB-NIR datasets respectively.

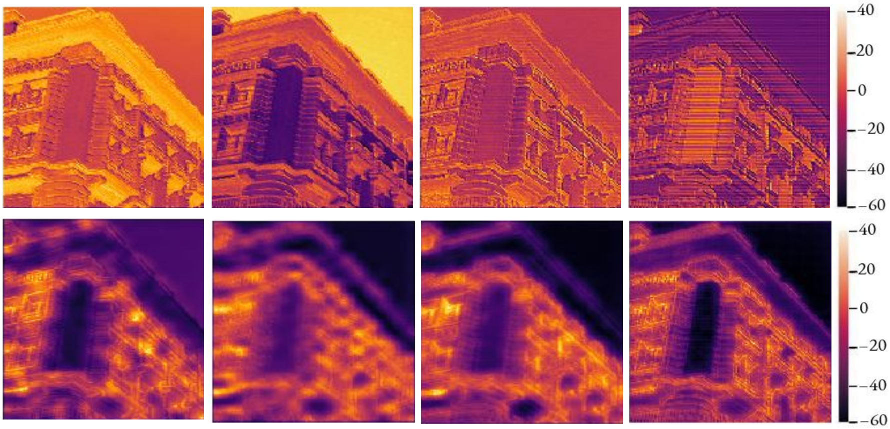


Figure 7.5: Average feature maps of TnTViT (*top*) and TnTViT-G (*bottom*) on different stages of Dense Feature Extraction.

Table 7.3: Average running time (s) and memory consumption (MB) comparison on RGB-NIR dataset for ×4.

Methods	Parameters(M)	Memory(MB)	Running Time(s)↓	PSNR(dB)↑
CARN[1]	1.5M	1,230	0.072	27.33
SwinIR[93]	1.0M	3,110	0.185	26.12
TNTViT (Ours)	1.2M	2,324	0.116	27.40
TNTViT-G (Ours)	1.3M	2,549	0.204	**29.59**

rectly employing Pixel Shuffle layer to produce images after fusing the information of both network's stream instead of our MLUP; *ii*) Using Pixel Shuffle layer followed by a conv layer and bicubic interpolation to scale the generated SR image to arbitrary scales. As can be seen in Table 7.4, when the suggested MLUP module is used for up scaling, superior results are obtained by a large margin compared to other upsampling techniques. These studies demonstrate that, opposite to common practice, the MLUP significantly improves reconstruction accuracy since the module is able to generate the SR images in multi level with the access of both direct and indirect shallow and abstract features which increasing the size of the encoder yields consistent improvements on benchmark datasets.

Table 7.4: Average PSNR results on RGB-NIR dataset for different upscaling methods with arbitrary scales. Best results are **highlighted**, second best underlined.

Experiment	**Scale**				
	×2	×2.2	×2.4	×2.6	×2.8
Pixel Shuffle	34.06	–	–	–	–
P.S. Bicubic	34.21	34.89	34.73	34.67	34.53
TNTViT MLUP	**34.49**	**35.24**	**35.07**	**34.90**	**34.71**
	×3	×3.1	×3.3	×3.5	×3.7
Pixel Shuffle	32.18	–	–	–	–
P.S. Bicubic	32.31	32.44	32.08	31.95	31.74
TNTViT MLUP	**32.57**	**32.45**	**32.29**	**32.15**	**32.02**
	×4	×4.2	×4.4	×4.6	×4.8
Pixel Shuffle	29.11	–	–	–	–
P.S. Bicubic	29.44	29.77	29.62	29.54	29.48
TNTViT MLUP	**29.59**	**29.85**	**29.81**	**29.73**	**29.64**

7.4.3 Learned Perceptual Image Patch Similarity

In Table 7.2, we provide the Learned Perceptual Image Patch Similarity (LPIPS) evaluation metric to evaluate the quality of the generated super resolution images since it has been demonstrated to correlate well with human perceptual similarity of image quality than other evaluation metrics (i.e, PSNR and SSIM). LPIPS is proposed by Zhang et al. [189], a deep-feature based evaluation metric that calculates the perceptual distance between two images. As can be seen, the proposed model achieves lower value than other approaches (lower is better). This shows the effectiveness of proposed TnTViT-G to generate more accurate and fine detailed SR IR images when HR visible images are accessible.

7.4.4 Model Complexity and Inference Time Analysis

Table 7.3 compares the proposed TnTViT-G architecture with existing CNN and Transformer based architecture approaches on M3FD test images in terms of Network Parameters (M), Inference Time (s), and Memory Consumption (MB). To provide a fair comparison, all models are tested using the same setup, including their public source code and default hyper-parameters, on a Intel Core i9-10900K CPU and a NVIDIA RTX 3090 GPU. As can be seen, TnTViT generate the SR images faster than other Transformer methods. This comparison shows that our proposed model properly balances performance and running time requirements.

7.5 Summary

This paper introduces TnTViT-G, a novel approach for guidance super resolution based on Transformer architecture. TnTViT-G is designed to accept two images of different domains, extract the information from each domain (infrared and corresponding visible image) with a separate stream and fuse them efficiently at different stages while remaining memory efficient. We propose a dense feature extraction, which contains both transformer self-attention layer and a convolutional attention module that can capture both global dependency and local context information at a lower computational cost while its well focusing on spatial features compared to other Transformer models. Furthermore, unlike to other GSR methods, TnTViT-G is able to generate the SR images in arbitrary sizes, while other methods only generate SR images in fixed sizes. Our experiments highlight that a high-cost, low-resolution spectral image (IR image) can be enhanced by a corresponding high-resolution, low-cost visible image (visible image). We have demonstrated that our approach achieves superior performance compared to other lightweight state-of-the-art methods on all benchmark datasets.

8 Conclusions and Future work

8.1 Conclusions

In addition to digital visible-spectrum imaging, IR imaging are making their way into numerous applications due to the spectacular advancements in digital sensors and the development of low-cost sensors. They offer a different perspective by including multi-modal spectral sensors, which prove crucial to boosting the precision of conventional digital vision and advancing machine vision beyond the visible spectrum. Construction of machine vision based on various modalities allows for the capture of features of physical objects, which are not visible to the human eye, including those in the thermal infrared spectrum.

Using several sensors for a single machine vision task generates a large amount of information. Therefore, it is vital to use the information from all the sensors and fuse them together to extract the greatest benefit from the provided data. However, the advancement of new generation multi-modal sensor networks camera presents a number of difficulties, the most significant one is the disparity in image quality and resolution between spectral regions, which prevents the fusion of details and integrated the images captured at different wavelengths.

This thesis presents my effort to investigate several challenging in image restoration tasks both single- and cross- domain by tackling the problem of Single Image Super Resolution, Guidance Image Super Resolution (i.e., cross domain), and Image Colorization through new deep learning approaches. In this research, we aim to advance the use of IR images by bringing image resolution closer to the human vision and expanding the scope of machine vision's possible uses. Also, we have studied regard to colorizing the IR images to map them as closely as possible to the human understanding of true colors since human perception is better at understanding such colors. In chapter 3, we propose a novel approach for colorizing near infrared images by using Generative Adversarial Network. The proposed model is able to map the NIR images to color image while there is not a paired dataset available. We introduce a new generator for CycleGAN and a new training strategy, which leads to generate a better visual quality compared to other SOTA approaches.

In chapter 4, we present a lightweight super resolution network to address the problem of Single Image Super Resolution by using Convlutional Neural Network. We present Multi-Path Residual Network, which contains set of Residual Concatenation Blocks with stacked by several Adaptive Residual blocks. The proposed network is able to adaptively extract important features and highlighting spatial context information, which is the key important in SISR task. Furthermore, a new attention module introduced to boost the performance of the network by spotlighting both informative features across the channel and spatial dimensions. We have experimentally shown that MPRNet outperforms previous state-of-the-art approaches in benchmarks datasets while maintaining relatively low computation and memory requirements. Moreover, in order to show the effectiveness of proposed model in other domains, we have experimented MPRNet for Thermal Image Super Resolution changeling, which shows a remarkable performance in infrared domain as well in chapter 5.

In chapter 6, we propose an effective and efficient network to tackle the problem of SISR by using Vison Transformer (ViT), called SRFormer, to mitigate the limitations of CNN models, which suffer from limited respective field and in-adaptability to the input content. Also, the proposed model cover the drawbacks of the Transformer based networks, which computational cost can increase dramatically high for HR images and the fact that transformer ignores the original structure of the input image. SRFormer is a powerful yet lightweight Transformer-based architecture that captures both local and global dependencies by redesigning several key importance of ViT without increasing computing costs and memory consumption. Furthermore, the Gated MLP Feature Fusion module introduced to aggregate the features of various phases of Transformer blocks by concentrating on their inter-spatial interaction. Extensive experiments shows SRFormer delivers huge performance gain compared to CNN based methods and superior results when it compared to Transformer based network.

Finally, in chapter 7, as our experiments showed success in enhancing the SR problem via a Convolutional Neural Network and Vision Transformer method, we turned our attention to extensive experiments to search for a better model architecture and other training procedures to advance the development of the super resolution problem on cross domain (IR and visible spectrum images). We extend the experiments by testing whether the rich texture details captured in the visible spectral images can contribute to the improvement and enhancement of the infrared images. Thus, we introduce a novel network, named TNTViT-G, which accept two domains as input and effectively aggregate the visible spectral image features with their inferred features counterparts to improve the quality of reconstructed super resolved infrared images. This experiment proved that high quality visible image can increase the resolution of low-quality IR image at lower cost than a high cost IR sensor.

The take-home message from this work is that by designing effective network with choosing correct training strategy, we are able to design fast, accurate, and lightweight networks which can increase the resolution of infrared images by using the help of the corresponding visible images at lower cost. We believe that the proposed approaches would have an important impact on the practical deployment.

8.2 Future Work

Despite the great success achieved by deep learning models in the single- and cross-domain image restoration tasks, there are still many unsolved problems. Thus, we will point out some of these problems explicitly and introduce some promising trends for future evolution in this section.

Real-world conditions provide significant challenges for image SR because to the fact that images captured in these conditions frequently exhibit indeterminate forms of image degradation, such as blurring, additive noise, and compression artifacts. Therefore, models that were trained on existing datasets that were done manually tend to have poor performance in real-world applications. As a result, developing an SR model that is capable of dealing with an unknown degradation is essential to boost the applicability of real-world applications. In addition, SR is not only limited to the use of domain-specific data and situations; it is also a significant assistance in the completion of other visual tasks. Therefore, applying SR to more particular domains, such as video surveillance, object detection, face recognition, medical imaging, and scene rendering, is also a potential avenue that should be pursued in the future.

Despite the fact that there has been experimental proof that low-cost channel (e.g., visible image) can be used to increase the resolution of expensive channel (e.g., infrared image), however this strategy relay on a well registered paired dataset, which is difficult to obtain such images since there is misalignment between multi-models sensors; and simple feed-forward network cannot deal with mismatch problem. Thus, image alignment technique is require as a pre-process to match the counterparts before encoding features of both domains and generate a super-resolved image. Therefore, there will be more research needs into this topic to solve the mismatch problem in a feed-forward network.

Furthermore, despite recent progress in unsupervised Image Colorization, there are a lot of work need to be done to map correct color components to infrared images. There are many possibilities of colors for each objects in a scene. For instance, human made objects can presents in different colors. We validates that an end-to-end deep learning architecture could be suitable for image colorization tasks. However, we believe that image colorization require some degree of human interactions, which it still a huge potential in the future and could eventually reduce hours of supervised work.

8.3 Scientific Articles

This dissertation has led to the following publications:

8.3.1 Published and Submitted Journals

- Armin Mehri, Parichehr Behjati, and Angel D. Sappa. SRFormer: Efficient Yet Powerful Transformer Network For Single Image Super Resolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. (*under review*)

- Armin Mehri, Parichehr Behjati, and Angel D.Sappa. TnTViT: Transformer in Transformer Network for Guidance Super Resolution. *IEEE Access*, 2022. (*under review*)

- Behjati, Parichehr, Pau Rodriguez, Carles Fernández, Isabelle Hupont, Armin Mehri, and Jordi Gonzàlez. "Single image super-resolution based on directional variance attention network." Pattern Recognition 133 (2023): 108997.

- Behjati, Parichehr, Pau Rodriguez, Carles Fernández Tena, Armin Mehri, F. Xavier Roca, Seiichi Ozawa, and Jordi Gonzàlez. "Frequency-Based Enhancement Network for Efficient Super-Resolution." IEEE Access 10 (2022): 57383-57397.

8.3.2 International Conferences and Workshops

- Mehri, Armin, and Angel D. Sappa. "Colorizing near infrared images through a cyclic adversarial approach of unpaired samples." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 0-0. 2019.

- Mehri, Armin, Parichehr B. Ardakani, and Angel D. Sappa. "MPRNet: Multipath residual network for lightweight image super resolution." In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 2704-2713. 2021.

- Mehri, Armin, Parichehr B. Ardakani, and Angel D. Sappa. "LiNet: A Lightweight

Network for Image Super Resolution." In 2020 25th International Conference on Pattern Recognition (ICPR), pp. 7196-7202. IEEE, 2021.

- Behjati, Parichehr, 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, pp. 2694-2703. 2021.

- Rivadeneira, Rafael E., Angel D. Sappa, and Armin Mehri et al. "Thermal image super-resolution challenge-pbvs 2021." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4359-4367. 2021.

- Rivadeneira, R. E., A. D. Sappa, and A. Mehri et al. "Thermal image superresolution challenge-pbvs 2020. In 2020 IEEE." In CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 432-439. 2020.

Bibliography

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

[2] F. Almasri. Towards the next generation of smart and visual multi-modal sensor, 2022.

[3] F. Almasri and O. Debeir. Multimodal sensor fusion in single thermal image super-resolution. In *Asian Conference on Computer Vision*, pages 418–433. Springer, 2018.

[4] A. Antoniou, A. Storkey, and H. Edwards. Data augmentation generative adversarial networks. *arXiv preprint arXiv:1711.04340*, 2017.

[5] S. Anwar, S. Khan, and N. Barnes. A deep journey into super-resolution: A survey. *ACM Computing Surveys (CSUR)*, 53(3):1–34, 2020.

[6] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 33(5):898–916, 2010.

[7] M. Arjovsky and L. Bottou. Towards principled methods for training generative adversarial networks. arxiv. 2017.

[8] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pages 214–223, 2017.

[9] B. C. Arrue, A. Ollero, and J. M. De Dios. An intelligent system for false alarm reduction in infrared forest-fire detection. *IEEE Intelligent Systems and their Applications*, 15(3):64–73, 2000.

[10] J. Baek, S. Hong, J. Kim, and E. Kim. Efficient pedestrian detection at night-time using a thermal camera. *Sensors*, 17(8):1850, 2017.

[11] J. T. Barron and B. Poole. The fast bilateral solver. In *European conference on computer vision*, pages 617–632. Springer, 2016.

[12] S. M. A. Bashir, Y. Wang, M. Khan, and Y. Niu. A comprehensive review of deep learning-based single image super-resolution. *PeerJ Computer Science*, 7:e621, 2021.

[13] M. Bates, B. Huang, G. T. Dempsey, and X. Zhuang. Multicolor super-resolution imaging with photo-switchable fluorescent probes. *Science*, 317 (5845):1749–1753, 2007.

[14] P. Behjati, P. Rodriguez, A. Mehri, I. Hupont, C. F. Tena, and J. 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.

[15] M. Bevilacqua, A. Roumy, C. Guillemot, and M. L. Alberi-Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. 2012.

[16] Y. Blau and T. Michaeli. The perception-distortion tradeoff. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6228–6237, 2018.

[17] Y. Blau, R. Mechrez, R. Timofte, T. Michaeli, and L. Zelnik-Manor. The 2018 pirm challenge on perceptual image super-resolution. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018.

[18] M. Brown and S. Süsstrunk. Multispectral SIFT for scene category recognition. In *Computer Vision and Pattern Recognition (CVPR11)*, pages 177–184, Colorado Springs, June 2011.

[19] M. Brown and S. Süsstrunk. Multi-spectral sift for scene category recognition. In *CVPR 2011*, pages 177–184. IEEE, 2011.

[20] Y. Cao, F. Wang, Z. He, J. Yang, and Y. Cao. Boosting image super-resolution via fusion of complementary information captured by multi-modal sensors. *IEEE Sensors Journal*, 22(4):3405–3416, 2021.

[21] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[22] G. Chantas, N. P. Galatsanos, R. Molina, and A. K. Katsaggelos. Variational bayesian image restoration with a product of spatially weighted total variation image priors. *IEEE transactions on image processing*, 19(2):351–362, 2009.

[23] G. K. Chantas, N. P. Galatsanos, and A. C. Likas. Bayesian restoration using a new nonstationary edge-preserving image prior. *IEEE Transactions on Image Processing*, 15(10):2987–2997, 2006.

[24] X. Chen, G. Zhai, J. Wang, C. Hu, and Y. Chen. Color guided thermal image super resolution. In *2016 Visual Communications and Image Processing (VCIP)*, pages 1–4. IEEE, 2016.

[25] Z. Cheng, Q. Yang, and B. Sheng. Deep colorization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 415–423, 2015.

[26] M. Chidambaram and Y. Qi. Style transfer generative adversarial networks: Learning to play chess differently. *arXiv preprint arXiv:1702.06762*, 2017.

[27] F. Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1251–1258, 2017.

[28] X. Chu, B. Zhang, H. Ma, R. Xu, J. Li, and Q. Li. Fast, accurate and lightweight super-resolution with neural architecture search. *arXiv preprint arXiv:1901.07261*, pages 59–64, 2019.

[29] X. Chu, B. Zhang, R. Xu, and H. Ma. Multi-objective reinforced evolution in mobile neural architecture search. *arXiv preprint arXiv:1901.01074*, 2019.

[30] X. Chu, B. Zhang, H. Ma, R. Xu, and Q. Li. Fast, accurate and lightweight super-resolution with neural architecture search. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 59–64. IEEE, 2021.

[31] T. Dai, J. Cai, Y. Zhang, S.-T. Xia, and L. Zhang. Second-order attention network for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11065–11074, 2019.

[32] Z. Daquan, Q. Hou, Y. Chen, J. Feng, and S. Yan. Rethinking bottleneck structure for efficient mobile network design. *arXiv preprint arXiv:2007.02269*, 2020.

[33] J. Distelzweig. How thermal imaging can combat false alarms, Oct 2018. URL https://www.securitysales.com/surveillance/thermal-imaging-false-alarms/.

[34] S. Dodge and L. Karam. Understanding how image quality affects deep neural networks. In *2016 eighth international conference on quality of multimedia experience (QoMEX)*, pages 1–6. IEEE, 2016.

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

[36] C. Dong, C. C. Loy, K. He, and X. Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307, 2015.

[37] C. Dong, C. C. Loy, and X. Tang. Accelerating the super-resolution convolutional neural network. In *European conference on computer vision*, pages 391–407. Springer, 2016.

[38] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[39] A. Fiandrotti, S. M. Fosson, C. Ravazzi, and E. Magli. Gpu-accelerated algorithms for compressed signals recovery with application to astronomical imagery deblurring. *International Journal of Remote Sensing*, 39(7):2043–2065, 2018.

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

[41] Q. Gao, Y. Zhao, G. Li, and T. Tong. Image super-resolution using knowledge distillation. In *Asian Conference on Computer Vision*, pages 527–541. Springer, 2018.

[42] R. Gerchberg. Super-resolution through error energy reduction. *Optica Acta: International Journal of Optics*, 21(9):709–720, 1974.

[43] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Region-based convolutional networks for accurate object detection and segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 38(1):142–158, 2015.

[44] A. C. Goldberg, T. Fischer, and Z. I. Derzko. Application of dual-band infrared focal plane arrays to tactical and strategic military problems. In *Infrared Technology and Applications XXVIII*, volume 4820, pages 500–514. SPIE, 2003.

[45] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, volume 27, pages 2672–2680, 2014.

[46] H. Greenspan. Super-resolution in medical imaging. *The Computer Journal*, 52(1):43–63, 2008.

[47] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems*, pages 5767–5777, 2017.

[48] B. K. Gunturk, A. U. Batur, Y. Altunbasak, M. H. Hayes, and R. M. Mersereau. Eigenface-domain super-resolution for face recognition. *IEEE transactions on image processing*, 12(5):597–606, 2003.

[49] M.-H. Guo, C.-Z. Lu, Z.-N. Liu, M.-M. Cheng, and S.-M. Hu. Visual attention network. *arXiv preprint arXiv:2202.09741*, 2022.

[50] Y. Guo, A. Yao, and Y. Chen. Dynamic network surgery for efficient dnns. *Advances in neural information processing systems*, 29, 2016.

[51] A. Gupta, N. Joshi, C. Lawrence Zitnick, M. Cohen, and B. Curless. Single image deblurring using motion density functions. In *European conference on computer vision*, pages 171–184. Springer, 2010.

[52] R. K. Gupta, A. Y.-S. Chia, D. Rajan, E. S. Ng, and H. Zhiyong. Image colorization using similar images. In *Proceedings of the 20th ACM international conference on Multimedia*, pages 369–378. ACM, 2012.

[53] L. Han and Z. Yin. Refocusing phase contrast microscopy images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 65–74. Springer, 2017.

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

[55] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, June 2016. ISBN 9781467388511. doi: 10.1109/cvpr.2016.90. URL http://dx.doi.org/10.1109/CVPR.2016.90.

[56] K. He, X. Zhang, S. Ren, and J. Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer, 2016.

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

[58] D. Hendrycks and K. Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

[59] B. Heo, S. Chun, S. J. Oh, D. Han, S. Yun, Y. Uh, and J.-W. Ha. Slowing down the weight norm increase in momentum-based optimizers. *arXiv preprint arXiv:2006.08217*, 2020.

[60] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, pages 6626–6637, 2017.

[61] Z. Hou and S.-Y. Kung. Efficient image super resolution via channel discriminative deep neural network pruning. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3647–3651. IEEE, 2020.

[62] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, et al. Searching for mobilenetv3. *arXiv preprint arXiv:1905.02244*, 2019.

[63] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[64] Y. Hu, J. Li, Y. Huang, and X. 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.

[65] Z. Hu, L. Xu, and M.-H. Yang. Joint depth estimation and camera shake removal from single blurry image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2893–2900, 2014.

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

[67] S. Huang, X. Jin, Q. Jiang, and L. Liu. Deep learning for image colorization: Current and future prospects. *Engineering Applications of Artificial Intelligence*, 114:105006, 2022. ISSN 0952-1976. doi: https://doi.org/10.1016/j.engappai.2022.105006. URL https://www.sciencedirect.com/science/article/pii/S0952197622001920.

[68] T.-W. Hui, C. C. Loy, , and X. Tang. Depth map super-resolution by deep multi-scale guidance. In *Proceedings of European Conference on Computer Vision (ECCV)*, pages 353–369, 2016.

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

[70] R. Ironi, D. Cohen-Or, and D. Lischinski. Colorization by example. In *Rendering Techniques*, pages 201–210. Citeseer, 2005.

[71] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[72] V. G. Jacob and S. Gupta. Colorization of grayscale images and videos using a semiautomatic approach. In *2009 16th IEEE International Conference on Image Processing (ICIP)*, pages 1653–1656. IEEE, 2009.

[73] X. Jia, X. Xu, B. Cai, and K. Guo. Single image super-resolution using multi-scale convolutional neural network. In *Pacific Rim Conference on Multimedia*, pages 149–157. Springer, 2017.

[74] K. Jiang, Z. Wang, P. Yi, and J. Jiang. Hierarchical dense recursive network for image super-resolution. *Pattern Recognition*, 107:107475, 2020.

[75] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016.

[76] A. Jolicoeur-Martineau. The relativistic discriminator: a key element missing from standard gan. *arXiv preprint arXiv:1807.00734*, 2018.

[77] P. Kansal and S. Nathan. Eyenet: Attention based convolutional encoder-decoder network for eye region segmentation. *arXiv preprint arXiv:1910.03274*, 2019.

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

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

[80] J. Kopf, M. F. Cohen, D. Lischinski, and M. Uyttendaele. Joint bilateral upsampling. *ACM Transactions on Graphics (ToG)*, 26(3):96–es, 2007.

[81] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6): 84–90, 2017.

[82] M. Kumar, D. Weissenborn, and N. Kalchbrenner. Colorization transformer. *arXiv preprint arXiv:2102.04432*, 2021.

[83] F. Lahoud, R. Zhou, and S. Susstrunk. Multi-modal spectral image super-resolution. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018.

[84] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 624–632, 2017.

[85] R. Lan, L. Sun, Z. Liu, H. Lu, C. Pang, and X. Luo. Madnet: A fast and lightweight network for single-image super resolution. *IEEE transactions on cybernetics*, 51(3):1443–1453, 2020.

[86] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017.

[87] B. Li, J. Liu, B. Wang, Z. Qi, and Y. Shi. s-lwsr: Super lightweight super-resolution network. *arXiv preprint arXiv:1909.10774*, 2019.

[88] D. Li, A. Zhou, and A. Yao. Hbonet: Harmonious bottleneck on two orthogonal dimensions. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3316–3325, 2019.

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

[90] K. Li, S. Yang, R. Dong, X. Wang, and J. Huang. Survey of single image super-resolution reconstruction. *IET Image Processing*, 14(11):2273–2290, 2020.

[91] W. Li, K. Zhou, L. Qi, N. Jiang, J. Lu, and J. Jia. Lapar: Linearly-assembled pixel-adaptive regression network for single image super-resolution and beyond. *Advances in Neural Information Processing Systems*, 33, 2020.

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

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

[94] Y. Liang, R. Timofte, J. Wang, S. Zhou, Y. Gong, and N. Zheng. Single-image super-resolution-when model adaptation matters. *Pattern Recognition*, 116: 107931, 2021.

[95] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 136–144, July 2017.

[96] M. Limmer and H. P. Lensch. Infrared colorization using deep convolutional neural networks. In *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 61–68. IEEE, 2016.

[97] D. Liu, Z. Wang, N. Nasrabadi, and T. Huang. Learning a mixture of deep networks for single image super-resolution. In *Asian Conference on Computer Vision*, pages 145–156. Springer, 2016.

[98] D. Liu, Z. Wang, Y. Fan, X. Liu, Z. Wang, S. Chang, and T. Huang. Robust video super-resolution with learned temporal dynamics. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2507–2515, 2017.

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

[100] J. Liu, W. Zhang, Y. Tang, J. Tang, and G. 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.

[101] J. Liu, X. Fan, Z. Huang, G. Wu, R. Liu, W. Zhong, and Z. Luo. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5802–5811, 2022.

[102] R. Liu, J. Lehman, P. Molino, F. P. Such, E. Frank, A. Sergeev, and J. Yosinski. An intriguing failing of convolutional neural networks and the coordconv solution, 2018.

[103] Y. Liu, J. Pan, J. Ren, and Z. Su. Learning deep priors for image dehazing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2492–2500, 2019.

[104] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.

[105] A. P. Lobanov. Resolution limits in astronomical images. *arXiv preprint astro-ph/0503225*, 2005.

[106] Z. Lu, H. Liu, J. Li, and L. Zhang. Efficient transformer for single image super-resolution. *arXiv preprint arXiv:2108.11084*, 2021.

[107] M. Lučić, K. Kurach, M. Michalski, S. Gelly, and O. Bousquet. Are gans created equal? a large-scale study. 2017.

[108] X. Luo, Y. Xie, Y. Zhang, Y. Qu, C. Li, and Y. Fu. Latticenet: Towards lightweight image super-resolution with lattice block.

[109] X. Luo, Y. Xie, Y. Zhang, Y. Qu, C. Li, and Y. 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.

[110] C. Ma, C.-Y. Yang, X. Yang, and M.-H. Yang. Learning a no-reference quality metric for single-image super-resolution. *Computer Vision and Image Understanding*, 158:1–16, 2017.

[111] Y. Ma, H. Xiong, Z. Hu, and L. Ma. Efficient super resolution using binarized neural network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[112] X. Mao, C. Shen, and Y.-B. Yang. Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections. *Advances in neural information processing systems*, 29:2802–2810, 2016.

[113] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2794–2802, 2017.

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

[115] Y. Matsui, K. Ito, Y. Aramaki, A. Fujimoto, T. Ogawa, T. Yamasaki, and K. Aizawa. Sketch-based manga retrieval using manga109 dataset. *Multimedia Tools and Applications*, 76(20):21811–21838, 2017.

[116] A. Mehri and A. D. Sappa. Colorizing near infrared images through a cyclic adversarial approach of unpaired samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[117] A. Mittal, R. Soundararajan, and A. C. Bovik. Making a “completely blind” image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012.

[118] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018.

[119] S. P. Mudunuri and S. Biswas. Low resolution face recognition across variations in pose and illumination. *IEEE transactions on pattern analysis and machine intelligence*, 38(5):1034–1040, 2015.

[120] A. Muqeet, J. Hwang, S. Yang, J. H. Kang, Y. Kim, and S.-H. Bae. Ultra lightweight image super-resolution with multi-attention layers. *arXiv preprint arXiv:2008.12912*, 2020.

[121] S. Nah, T. Hyun Kim, and K. Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3883–3891, 2017.

[122] S. Nathan and P. Kansal. Skeletonnet: Shape pixel to skeleton pixel. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRw)*, 2019.

[123] A. Odena, V. Dumoulin, and C. Olah. Deconvolution and checkerboard artifacts. *Distill*, 1(10):e3, 2016.

[124] M. Oliveira, A. D. Sappa, and V. Santos. A probabilistic approach for color correction in image mosaicking applications. *IEEE Transactions on image Processing*, 24(2):508–523, 2015.

[125] T. Peleg and M. Elad. A statistical prediction model based on sparse representations for single image super-resolution. *IEEE transactions on image processing*, 23(6):2569–2582, 2014.

[126] C. Plizzari, M. Cannici, and M. Matteucci. Skeleton-based action recognition via spatial and temporal transformer networks. *Computer Vision and Image Understanding*, 208:103219, 2021.

[127] F. Qingyun, H. Dapeng, and W. Zhaokui. Cross-modality fusion transformer for multispectral object detection. *arXiv preprint arXiv:2111.00273*, 2021.

[128] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[129] W. Ren, J. Pan, X. Cao, and M.-H. Yang. Video deblurring via semantic segmentation and pixel-wise non-linear kernel. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1077–1085, 2017.

[130] W. H. Richardson. Bayesian-based iterative method of image restoration. *JoSA*, 62(1):55–59, 1972.

[131] R. E. Rivadeneira, A. D. Sappa, and B. X. Vintimilla. Thermal image super-resolution: a novel architecture and dataset. In *Proceedings of the 15th International Conference on Computer Vision Theory and Application (VISAPP), Valletta, Malta*, pages 111–119, February 2020.

[132] A. Rogalski, P. Martyniuk, and M. Kopytko. Challenges of small-pixel infrared detectors: a review. *Reports on Progress in Physics*, 79(4):046501, 2016.

[133] O. Ronneberger, P. Fischer, and T. 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.

[134] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

[135] S. Schulter, C. Leistner, and H. Bischof. Fast and accurate image upscaling with super-resolution forests. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3791–3799, 2015.

[136] L. Shi, Y. Zhang, J. Cheng, and H. Lu. Decoupled spatial-temporal attention network for skeleton-based action recognition. *arXiv preprint arXiv:2007.03263*, 2020.

[137] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1874–1883, 2016.

[138] Z. Shi, C. Chen, Z. Xiong, D. Liu, Z.-J. Zha, and F. Wu. Deep residual attention network for spectral image super-resolution. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018.

[139] A. Shocher, N. Cohen, and M. 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.

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

[141] P. Song, X. Deng, J. F. Mota, N. Deligiannis, P. L. Dragotti, and M. R. Rodrigues. Multimodal image super-resolution via joint sparse representations induced by coupled dictionaries. *IEEE Transactions on Computational Imaging*, 6: 57–72, 2019.

[142] X. Soria, A. D. Sappa, and A. Akbarinia. Multispectral single-sensor RGB-NIR imaging: New challenges and opportunities. In *2017 Seventh International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pages 1–6. IEEE, 2017.

[143] J. Su, B. Xu, and H. Yin. A survey of deep learning approaches to image restoration. *Neurocomputing*, 487:46–65, 2022.

[144] P. L. Suárez, A. D. Sappa, and B. X. Vintimilla. Infrared image colorization based on a triplet DCGAN architecture. In *2017 IEEE Conference on Computer*

Vision and Pattern Recognition Workshops (CVPRW), pages 212–217. IEEE, 2017.

[145] P. L. Suárez, A. D. Sappa, and B. X. Vintimilla. Learning to colorize infrared images. In *International Conference on Practical Applications of Agents and Multi-Agent Systems*, pages 164–172. Springer, 2017.

[146] Y. Tai, J. Yang, and X. Liu. Image super-resolution via deep recursive residual network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3147–3155, 2017.

[147] Y. Tai, J. Yang, X. Liu, and C. Xu. Memnet: A persistent memory network for image restoration. In *Proceedings of the IEEE international conference on computer vision*, pages 4539–4547, 2017.

[148] K. Takumi, K. Watanabe, Q. Ha, A. Tejero-De-Pablos, Y. Ushiku, and T. Harada. Multispectral object detection for autonomous vehicles. In *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*, pages 35–43, 2017.

[149] M. Teutsch, A. D. Sappa, and R. I. Hammoud. Computer vision in the infrared spectrum: challenges and approaches. *Synthesis Lectures on Computer Vision*, 10(2):1–138, 2021.

[150] . Thermal Temp. Thermal body temperature security cameras va md washington dc. URL https://surveillancesecure.com/security-cameras/thermal-cameras/.

[151] R. Timofte, V. De Smet, and L. Van Gool. Anchored neighborhood regression for fast example-based super-resolution. In *Proceedings of the IEEE international conference on computer vision*, pages 1920–1927, 2013.

[152] R. Timofte, R. Rothe, and L. Van Gool. Seven ways to improve example-based single image super resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1865–1873, 2016.

[153] R. Timofte, E. Agustsson, L. Van Gool, M.-H. Yang, and L. 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.

[154] T. Tong, G. Li, X. Liu, and Q. Gao. Image super-resolution using dense skip connections. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4799–4807, 2017.

[155] L. Trottier, P. Gigu, B. Chaib-draa, et al. Parametric exponential linear unit for deep convolutional neural networks. In *16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 207–214. IEEE, 2017.

[156] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[157] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 568–578, 2021.

[158] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018.

[159] X. Wang, Q. Wang, Y. Zhao, J. Yan, L. Fan, and L. Chen. Lightweight single-image super-resolution network with attentive auxiliary feature learning. In *Proceedings of the Asian Conference on Computer Vision*, 2020.

[160] Y. Wang, F. Perazzi, B. McWilliams, A. Sorkine-Hornung, O. Sorkine-Hornung, and C. Schroers. A fully progressive approach to single-image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 864–873, 2018.

[161] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli, et al. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[162] Z. Wang, S. Chang, Y. Yang, D. Liu, and T. S. Huang. Studying very low resolution recognition using deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4792–4800, 2016.

[163] Z. Wang, J. Chen, and S. C. Hoi. Deep learning for image super-resolution: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43 (10):3365–3387, 2020.

[164] Z. Wang, X. Cun, J. Bao, and J. Liu. Uformer: A general u-shaped transformer for image restoration. *arXiv preprint arXiv:2106.03106*, 2021.

[165] T. Welsh, M. Ashikhmin, and K. Mueller. Transferring color to greyscale images. In *ACM Transactions on Graphics (TOG)*, volume 21, pages 277–280. ACM, 2002.

[166] W. K. Wong, H. L. Lim, C. K. Loo, and W. S. Lim. Home alone faint detection surveillance system using thermal camera. In *2010 Second International Conference on Computer Research and Development*, pages 747–751. IEEE, 2010.

[167] S. Woo, J. Park, J.-Y. Lee, and I. So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–19, 2018.

[168] T. Xiao, P. Dollar, M. Singh, E. Mintun, T. Darrell, and R. Girshick. Early convolutions help transformers see better. *Advances in Neural Information Processing Systems*, 34, 2021.

[169] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34, 2021.

[170] J. Xie, R. S. Feris, and M.-T. Sun. Edge-guided single depth image super resolution. *IEEE Transactions on Image Processing*, 25(1):428–438, 2015.

[171] Y. Xu, C. Xu, X. Chen, W. Zhang, C. Xu, and Y. Wang. Kernel based progressive distillation for adder neural networks. *Advances in Neural Information Processing Systems*, 33:12322–12333, 2020.

[172] C.-Y. Yang and M.-H. Yang. Fast direct super-resolution by simple functions. In *Proceedings of the IEEE international conference on computer vision*, pages 561–568, 2013.

[173] J. Yang, Z. Wang, Z. Lin, S. Cohen, and T. Huang. Coupled dictionary training for image super-resolution. *IEEE transactions on image processing*, 21(8): 3467–3478, 2012.

[174] Y. Yang and Y. Qi. Image super-resolution via channel attention and spatial graph convolutional network. *Pattern Recognition*, 112:107798, 2021.

[175] Z. Yi, H. Zhang, P. Tan, and M. Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2849–2857, 2017.

[176] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang. Generative image inpainting with contextual attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5505–5514, 2018.

[177] W. Yu, M. Luo, P. Zhou, C. Si, Y. Zhou, X. Wang, J. Feng, and S. Yan. Metaformer is actually what you need for vision. *arXiv preprint arXiv:2111.11418*, 2021.

[178] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, Z.-H. Jiang, F. E. Tay, J. Feng, and S. Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 558–567, 2021.

[179] Y. Yuan, S. Liu, J. Zhang, Y. Zhang, C. Dong, and L. 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.

[180] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang. Restormer: Efficient transformer for high-resolution image restoration. *arXiv preprint arXiv:2111.09881*, 2021.

[181] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.

[182] M. D. Zeiler, D. Krishnan, G. W. Taylor, and R. Fergus. Deconvolutional networks. In *2010 IEEE Computer Society Conference on computer vision and pattern recognition*, pages 2528–2535. IEEE, 2010.

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

[184] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE transactions on image processing*, 26(7):3142–3155, 2017.

[185] K. Zhang, W. Zuo, S. Gu, and L. Zhang. Learning deep cnn denoiser prior for image restoration. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3929–3938, 2017.

[186] K. Zhang, W. Zuo, and L. 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.

[187] R. Zhang, P. Isola, and A. A. Efros. Colorful image colorization. In *European conference on computer vision*, pages 649–666. Springer, 2016.

[188] R. Zhang, J.-Y. Zhu, P. Isola, X. Geng, A. S. Lin, T. Yu, and A. A. Efros. Real-time user-guided image colorization with learned deep priors. *arXiv preprint arXiv:1705.02999*, 2017.

[189] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[190] T. Zhang, A. Wiliem, S. Yang, and B. Lovell. Tv-gan: Generative adversarial network based thermal to visible face recognition. In *2018 international conference on biometrics (ICB)*, pages 174–181. IEEE, 2018.

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

[192] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu. Residual dense network for image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2472–2481, 2018.

[193] D. Zhou, R. Wang, J. Lu, and Q. Zhang. Depth image super resolution based on edge-guided method. *Applied Sciences*, 8(2):298, 2018.

[194] F. Zhu and Q. Zhao. Efficient single image super-resolution via hybrid residual feature learning with compact back-projection network. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[195] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2223–2232, 2017.

[196] W. W. Zou and P. C. Yuen. Very low resolution face recognition problem. *IEEE Transactions on image processing*, 21(1):327–340, 2011.