Silesian University of Technology

FACULTY OF AUTOMATIC CONTROL, ELECTRONICS AND COMPUTER SCIENCE

DOCTORAL DISSERTATION

Multi-image super-resolution reconstruction using deep graph neural networks

---

A thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy

mgr inż. Tomasz Tarasiewicz

Supervisor: **dr hab. inż. Michał Kawulok, prof. PŚ**

September 27, 2023

"And the Earth becomes my throne, I adapt to the unknown, under wandering stars I've grown, by myself but not alone."

James Hetfield

Contents

Chapter 1

Introduction

1.1 Understanding Image Resolution

At its core, image resolution defines the level of detail an image can hold. It is typically expressed in terms of pixels for digital images, indicating the number of individual pixels present in the image both horizontally and vertically. The more pixels an image contains, the more detailed it can be. This detail affects not only the clarity and sharpness but also the potential accuracy of any analysis derived from the image [38].

In the realm of digital photography and imaging, the resolution is paramount [11]. As images serve as a primary medium of conveying information in various domains, from medicine [23] to satellite imaging [55], the clarity and detail they provide can be crucial.

However, obtaining images of higher spatial resolution is not always feasible. Physical limitations of imaging sensors [35], constraints in bandwidth and storage [60], and unfavourable conditions during image capture (like atmospheric disturbances in satellite imaging) [121] can lead to images that lack the desired detail or clarity.

1.2 Motivation for Super-Resolution

Given the significant importance of image clarity across multiple domains and the inherent challenges in obtaining *high-resolution* (HR) images directly, there is a pressing need for techniques that can enhance the spatial resolution of existing images [156]. This is where the concept of *super-resolution reconstruction* (SRR) enters the frame. The ability to derive HR images from *low-resolution* (LR) counterparts can unlock new possibilities, from refining visual content for entertainment and education to enhancing critical decision-making in fields like medicine, geospatial studies, and defence.

Furthermore, as the world becomes increasingly digitized, the volume of visual data available for analysis grows exponentially [32]. There is a significant amount of LR data already in existence, and the ability to enhance these data can provide fresh insights and understanding. However, merely increasing the pixel count of an image does not necessarily equate to an improvement in its true resolution [131]. Specifically, in remote sensing applications, there exists a critical distinction between nominal and effective *ground sampling distance* (GSD) [93]. While the nominal GSD quantifies how much real-world distance each pixel represents, the effective GSD conveys the true clarity or discernibility of the image beyond mere pixel count. The challenge then becomes twofold: to increase the spatial resolution while simultaneously enhancing the authentic details and textures in the image. Achieving this demands sophisticated techniques capable of inferring and reintroducing details absent in the original LR image, rendering super-resolution reconstruction a complex yet highly rewarding endeavour.

The blend of deep learning with SRR provides a promising avenue to tackle this challenge [27]. With its ability to learn from vast amounts of data and model intricate, non-linear relationships, deep learning can potentially understand and recreate the details in LR images with unprecedented accuracy. In essence, the journey to refine and enhance image resolution using SRR techniques, especially those rooted in deep learning, is not just a pursuit of academic interest. It holds the promise of vast real-world applications, driving improvements in numerous fields and paving the way for innovations that hinge on the clarity and detail of visual data.

1.3 Differentiating Single-Image and Multi-Image Super-Resolution

The super-resolution field is divided into two primary branches: *single-image super-resolution* (SISR) and *multi-image super-resolution* (MISR). Both methodologies aim to create HR images from lower-resolution inputs, but they fundamentally differ in their approach and techniques.

1.3.1 Single-Image Super-Resolution

SISR, a key branch of SRR, works with a single LR image to generate an HR counterpart. The fundamental problem that it aims to solve is to predict the high-frequency components—the finer details, textures, and contours—that

are missing in the LR image. This is a non-trivial problem, because high-frequency information tends to be lost or distorted during the image degradation process, either through natural effects such as blurring or through digital processes such as compression.

To solve the SISR problem, various methods have been explored over time. Earlier approaches relied on non-learning-based techniques [133], such as bicubic interpolation [64], that use the surrounding pixel values to estimate those in higher resolutions. While straightforward and lightweight in the computational sense, these methods often produce smoothed images that lack finer details. As the field evolved, learning-based approaches became dominant [105]. Typically, these methods involve a training phase where models learn the mapping between low and HR image spaces from large-scale datasets of such pairs of images with different spatial resolutions. The models employed can range from classic machine learning techniques, like sparse coding [41, 101, 156] [12, 53], to advanced deep learning strategies, such as *convolutional neural networks* (CNNs) [75] or *generative adversarial networks* (GANs) [39, 76, 140].

Deep learning-based SISR methods have been recognized for superior performance in capturing intricate data details through complex, non-linear mappings [28, 142]. These methods employ advanced architectural designs like residual learning [66], attention mechanisms [2], and multi-scale processing [72] to better model the SRR problem. However, inherent limitations exist in SISR due to its fundamentally ill-posed nature; with multiple possible HR outcomes for a given LR image, the ambiguity in reconstruction increases with the magnification ratio. Although SISR models strive to infer missing details in the LR input using educated estimations from training, they are bound by the information in a single image, which might not suffice for accurately deducing all absent details [161].

1.3.2 Multi-Image Super-Resolution

Unlike SISR, MISR uses multiple LR images of a scene to achieve a more detailed image of higher resolution. These images have varying, complementary information due to minor capture differences. Ideally, these differences appear as slight, *sub-pixel* shifts between observations, but in reality, images taken at different times often vary in brightness, contrast, sensor noise or other factors and MISR methods have to minimize negative implications of such distortions [78].

MISR can be dissected into two main stages: *registration* and *fusion* [96, 109, 132]. Registration aligns input images, ensuring shared features are spatially matched [16]. It starts by extracting necessary affine transformations, like shift vectors or rotation matrices, and then applying them for alignment. For images that are only shifted, registration is categorized as full-pixel or sub-pixel, based on alignment granularity [128]. Full-pixel registration, correcting major misalignments without changing pixel values, is vital for preserving information for fusion [135]. On the contrary, sub-pixel registration, aiming for finer alignment by adjusting pixel values, must be cautiously handled to avoid compromising fusion information. Theoretically, ideal sub-pixel registration with only shifted images can reduce the volume of information required for MISR, turning the problem into a SISR one due to information loss; hence most MISR methods stick to full-pixel registration [22, 91, 111].

After registration, the images are fused into a single HR output, using collective information from multiple images to better render scene details. MISR's advantage is in the extra information from additional images, making super-resolution reconstruction more accurate and robust compared to SISR, and helping reduce noise and distortion effects [2, 56]. The MISR field has embraced machine learning, particularly deep learning, to significantly improve performance by leveraging complex image data relationships [61]. Unlike SISR, which focuses on understanding degradation, MISR concentrates on utilizing the diversity in multiple scene images, with deep neural networks designed to optimally merge this information.

1.3.3 Common Challenges in MISR

While MISR offers promising improvements over SISR, it introduces a unique set of challenges. In addition to registration, these challenges can be manifested as *temporal differences* between the multiple LR images taken at different points in time. This temporal variation between images adds a layer of complexity, particularly due to changes in the environment, such as various optical aberrations or object motion. For images spanning wider time frames, especially in remote sensing applications, more pronounced inconsistencies across the LR images can arise from changing weather patterns or man-made alterations like construction [104]. For instance, conflicting information between images—like a scene with a newly constructed building in one image

and an empty lot in another—poses the challenge of deciding which representation to prioritize during super-resolution.

Furthermore, common issues extend to occlusions, where parts of the scene are obscured in some images [170], noise interference with varying noise characteristics across different images [30, 87], and variable lighting conditions introducing disparities in object appearances [149]. These complications collectively make the task of multi-image super-resolution more demanding, necessitating robust algorithms to handle such complexities and achieve a high-quality super-resolved image. Notably, traditional CNNs often struggle with generating visual artefacts in regions of high temporal variance, such as rapidly moving objects, busy urban intersections, or areas with dense vegetation and rapid growth [152].

These challenges, ranging from alignment issues to temporal inconsistencies, make MISR a sophisticated task, demanding robust algorithms capable of handling such complexities to achieve a high-quality super-resolved image.

1.3.4 Importance of MISR

MISR, as a software-based solution, augments existing hardware limitations in a cost-effective and versatile manner, unlike traditional methods that entail physical enhancements to the imaging sensor's resolution [105]. By leveraging the inherent redundancy in multiple images of the same scene, captured from slightly different perspectives and at different times, MISR infers details not visible in any single LR image, delivering an HR output with more accurate and detailed information. This development of MISR techniques addresses the growing demand for higher-resolution images across various fields, marking a significant stride in digital imagery enhancement beyond the hardware limitations.

The wide-ranging applications of MISR techniques underscore their importance. Here are some of the key domains:

- **Remote Sensing and Geospatial Imaging**: In this field, MISR can help enhance the resolution of satellite or drone imagery, providing more detailed geographical data for environmental monitoring, urban planning, and resource management [100, 109, 141].

- **Medical Imaging**: In medical applications, MISR can be used to improve the resolution of images obtained from various imaging modalities such as magnetic resonance imaging, computed tomography, and

ultrasound. This can aid in more accurate diagnosis and treatment planning [119].

- **Surveillance and Security**: In surveillance systems, MISR can be used to enhance the quality of CCTV footage, which can assist in identifying faces, license plates, or other details that might be critical for security or forensic investigations [51, 116].
- **Entertainment and Media**: In the film and television industry, MISR can be used to upsample old, LR footage to a higher resolution, making it suitable for modern high-resolution displays [31, 90].
- **Astronomy**: MISR can be used to enhance the resolution of astronomical images, providing clearer views of distant celestial bodies and phenomena [43].
- **Document Processing**: MISR can enhance scanned documents, restore historical manuscripts, improve the accuracy of *optical character recognition* methods, and support forensic investigations. Its ability to produce clearer, HR versions of the documents is valuable across sectors like academia and law enforcement [103].

1.4 Central Themes of the Dissertation

When addressing the challenges associated with MISR, traditional CNNs have demonstrated proficiency in processing both spatial and temporal information [61]. However, their inherent design often lacks the finesse required to untangle the complex, asymmetric relationships among individual pixels based on spatial, spectral, or temporal dimensions. For instance, while CNNs can identify and process patterns, they have limited ability to discern the distinct connection of a pixel to its specific neighbour based on these various criteria. In MISR scenarios, where the unique relationship of every pixel, such as precise relative displacements, could be instrumental for accurate super-resolution, there emerges a potential for models adept in capturing and utilizing these nuanced relationships. Moreover, a fundamental limitation arises from the common methodology employed by CNNs in MISR, which involves stacking LR images in a matrix format. This representation provides no explicit information about the shifts between these images, especially the sub-pixel ones, obliging these models to deduce such shifts indirectly. Although some methods attempt to address this challenge in MISR,

for example, through a dedicated registration submodule [96], the matrix-based representation adopted by CNNs does not facilitate lossless sub-pixel registration.

This dissertation thus pivots towards exploring the potential of *graph neural networks* (GNNs), which inherently possess the capacity to encode such multifaceted relationships [168]. Unlike their CNN counterparts, GNNs consider the image as a graph, with each node representing a pixel or a region, and the edges signifying the relationships between them [70]. Moreover, this graph representation overcomes the limitations faced by CNNs, as it embeds the information about sub-pixel shifts directly into the graph without modifying the original input information, facilitating a more nuanced understanding and processing of the intricate relationships that are commonplace in MISR scenarios.

The pivotal assertion of this dissertation is that GNNs can significantly impact the course of advancements in MISR. This proposition is anchored on three core theses:

1. **When a set of LR images with sub-pixel shifts are represented as a graph, GNNs can process this graph to achieve super-resolution results comparable or better to those obtained by leading MISR architectures based on convolutional networks.**

2. **GNNs can enhance their MISR performance by incorporating techniques inspired by existing state-of-the-art MISR models based on CNNs. These techniques include individual feature extraction for each LR image, the use of attention mechanisms, and dynamic and trainable input registration.**

3. **GNNs can reconstruct a scene from a specific point in time by selecting a particular reference image from the input LR image set, with other images serving as supplementary information sources to enhance super-resolution accuracy. This approach can help reduce visual artefacts in regions of high temporal variability.**

The validation of these theses has been conducted both quantitatively and qualitatively to assess the effectiveness of GNNs in MISR tasks. Quantitative validation on traditional MISR benchmarks offers a measurable comparison of GNN-based models against existing methods. Qualitative validation sheds light on the visual quality and interpretability of the super-resolved images generated by different models. Especially for the last thesis on managing temporal variations, additional validation can be done by inspecting

the differences in the outputs of the same scene, generated using different reference images. This inspection demonstrates how GNNs handle temporal relations among input LR images, and how this capability affects the visual consistency and accuracy of the super-resolved outputs. Through analyzing these differences, the robustness and adaptability of GNNs in managing temporal dynamics in MISR can be substantiated.

Guided by these theses, the subsequent chapters of this dissertation explore GNNs in the context of MISR, aiming to provide both a detailed theoretical analysis and empirical validations through experiments. The results are evaluated quantitatively on traditional MISR benchmarks and qualitatively through visual inspections, particularly considering the temporal variations as outlined in the third thesis. Through this structured examination, this research aims to offer a clear argument for the potential of GNNs in addressing the challenges faced in MISR.

1.5 Published Works

The journey towards addressing the challenges in MISR through this dissertation was significantly informed and enriched by a series of published works. The papers listed below, in chronological order, either directly contributed to the thematic focus of this dissertation or provided essential theoretical and practical knowledge that propelled the research forward. It should be noted that these are among several other works published by the author of this dissertation, and have been selected for their relevance to the central themes of this study.

1. **A Graph Neural Network for Multiple-Image Super-Resolution**, in IEEE International Conference on Image Processing, 2021. *MNSIW: 70* [123]

2. **Deep Learning for Multiple-Image Super-Resolution of Sentinel-2 Data**, in IEEE International Geoscience and Remote Sensing Symposium, 2021. *MNSIW: 20* [62]

3. **Semi-Simulated Training Data for Multi-Image Super-Resolution**, in IEEE International Geoscience and Remote Sensing Symposium, 2022. *MNSIW: 20* [124]

4. **Extracting High-Resolution Cultivated Land Maps from Sentinel-2 Image Series**, in IEEE International Geoscience and Remote Sensing Symposium, 2022. *MNSIW: 20* [125]

5. **Transformer-Based Spectro-Temporal Fusion for Sentinel-2 Super-Resolution**, in International Conference on Systems, Signals and Image Processing, 2023. *MNSIW: 20* [107]

6. **Graph-Based Representation for Multi-Image Super-Resolution**, in International Workshop on Graph-Based Representations in Pattern Recognition, 2023. *MNSIW: 20* [122]

7. **A Real-World Benchmark for Sentinel-2 Multi-Image Super-Resolution**, in Scientific Data, In press. *MNSIW: 140* [69]

8. **Multitemporal and Multispectral Data Fusion for Super-Resolution of Sentinel-2 Images**, in IEEE Transactions on Geoscience and Remote Sensing, In press. *MNSIW: 200* [126]

1.6 Thesis Organization

The dissertation is organized as follows:

- **Chapter 2** provides a detailed literature review of MISR techniques, emphasizing deep learning-based approaches. This chapter also introduces GNNs and outlines the existing research gaps and challenges associated with current MISR techniques.

- **Chapter 3** delves into the proposed MISR technique using deep GNNs, highlighting its innovative approach in addressing the identified challenges.

- **Chapter 4** is dedicated to the datasets used in this research. It describes the real-world datasets and provides an in-depth explanation of the simulation process for generating data.

- **Chapter 5** details the training process of the models and the evaluation methodology. It also introduces and explains the metrics used for evaluation.

- **Chapter 6** presents the experimental results of the proposed method, comparing it with existing techniques. It discusses the implications of

these results, potential applications, and points towards possible directions for future research.

- **Chapter 7** concludes the dissertation, offering a summary of the findings and providing suggestions for future avenues of research.

Chapter 2

Related Work

The field of super-resolution reconstruction has witnessed considerable advancements over the years, fueled by the growing demand for HR imagery in diverse domains such as remote sensing, surveillance, and medical imaging. This area of study has seen an array of methods and techniques being proposed and validated, each with its unique strengths and limitations. The current landscape of SRR techniques is a testament to the relentless efforts of researchers worldwide to improve upon the existing methodologies and address emerging challenges. This section aims to provide a comprehensive overview of the key developments in this field, with a particular emphasis on the evolution of SISR (Section 1.3.1) and MISR (Section 1.3.2) methods and their underlying principles. Additionally, this chapter introduces GNNs, detailing their foundational concepts and illustrating their operational mechanisms, setting the stage for their application in the realm of SRR.

2.1 Single Image Super-Resolution

SISR has been an area of extensive exploration in the past years, with CNNs being central to the majority of the state-of-the-art solutions [142]. CNNs have been instrumental in advancing the performance of SISR solutions by effectively modelling the relationship between LR and HR images through feature representation and nonlinear mapping.

The first CNN proposed for SR, the *super-resolution convolutional neural network* (SRCNN) [28], was composed of just three convolutional layers. Subsequent advancements in the field gave rise to *SRResNet* [76], which notably introduced the incorporation of residual blocks to simplify training and enhance super-resolution performance. This design was inspired by the success of *ResNet*, the pioneering architecture that initially introduced the concept of deep residual learning [48]. Following this, the *multi-scale deep super-resolution* (MDSR) network was developed, offering an innovative approach

by integrating multiple scaling factors into a single model, allowing it to handle different upsampling tasks without switching models [80]. Building on these advancements, the *residual channel attention network* (RCAN) [166] was proposed, which further enhanced SISR by using a channel attention mechanism to weigh the importance of different channels and refine the feature representations, proving particularly beneficial for capturing intricate image details. Additionally, more complex models of much larger capacities, like the *enhanced deep SR network* (EDSR) [80], showcased the potential to model the LR-to-HR relationship even more effectively.

In addition to traditional CNNs, the application of GAN models to SISR has witnessed notable advancements. *SRGAN* [76], which utilised *SRResNet* as its generator and paired it with a single discriminator network, showcases the power of adversarial training in super-resolution, particularly in producing images with high perceptual quality. Building upon the foundation laid by SRGAN, *ESRGAN* [140] introduced architectural and loss function modifications, leveraging a robust adversarial loss and integrating residual-in-residual dense blocks. This resulted in images with enhanced details and sharper textures, setting new standards in the field. It is important to note that while GANs are particularly effective in reconstructing images of high perceptual quality, they do not necessarily recover the actual high-frequency information [117]. Despite this, GANs have been utilized in remote sensing [143] or medical imaging [44] applications, and it has been demonstrated that certain constraints on the adversarial loss can increase the reliability of the reconstruction outcome [65]. Another recent development in SISR is the use of vision transformers, which dynamically adjust the size of the feature maps, thus reducing the model complexity [86]. While this approach has shown potential, its broader implications and effectiveness in the context of SRR are still being explored.

2.2 Multi-Image Super-Resolution

The concept of MISR has its roots in early video processing techniques. In the late 1980s and early 1990s, the idea of using multiple LR frames to reconstruct a higher-resolution frame was first proposed [132]. This technique was initially known as multi-frame image restoration, and it was primarily used to enhance the quality of video footage.

The process exploited the natural movement in a video sequence, which caused each frame to capture slightly different information about the scene.

By aligning and combining these frames, it was possible to extract more details than could be seen in any single frame, resulting in a higher-resolution output. This concept was revolutionary at the time, as it provided a means to enhance the resolution of video footage without requiring any improvements in hardware. As computational power increased and algorithms became more sophisticated, the concept of MISR was extended to still images, resulting in the development of new techniques. MISR has since evolved into a complex field of study that involves various sub-disciplines, including image registration[95], image fusion [102], and machine learning.

Compared to SISR, its multi-image counterpart often achieves higher accuracy in reconstructing images [97, 158]. While SISR methods generate missing high-frequency details from a single image, MISR models leverage the information contained in multiple LR images of the same scene. This multi-image approach provides richer data, enabling the fusion of information from different images to enhance resolution.

MISR has witnessed significant progress during recent years, largely driven by the *Proba-V Super Resolution Competition* [91], an initiative of the European Space Agency in 2018 that ran for a duration of eight months. The primary objective of this competition was to super-resolve satellite images with a GSD of 300 meters to a finer, 100 m GSD resolution, thus it aimed to upsample the input images by a factor of three. GSD refers to the distance between two consecutive pixel centres measured on the ground, indicating the spatial resolution of an image [79]. The data for this challenge was sourced from the Proba-V satellite, with each scene composed of a varying number of observations. It introduced challenges such as obscured pixels from obstructions like clouds or their shadows, and uncaptured pixels for which the data was not observed. Even though the challenge concluded in 2019, the servers remain operational, allowing for so-called post-mortem evaluations. Current results from these ongoing assessments are recorded and showcased on a dedicated post-mortem leaderboard. A comprehensive overview of the dataset curated for this challenge can be found in Section 4.2.

Early in the deep learning era of MISR, *EvoNet* [61] emerged, applying convolutional neural networks for SISR combined with an evolutionary fusion strategy for multi-image integration [63]. It was succeeded by *DeepSUM* [96], a pioneering end-to-end deep learning model for MISR and the winner of the Proba-V challenge. Although powerful, DeepSUM's architecture is constrained to a fixed number of LR inputs and necessitates extensive training. *HighRes-Net* [22] subsequently addressed DeepSUM's limitations

by introducing a flexible model capable of handling variable input image counts. Its unique recursive fusion mechanism for latent representations marked a significant departure from earlier models. For a detailed exploration of HighRes-Net's architecture and contributions, refer to Section 2.4. Similarly, *RAMS* [111] model, while operational on a fixed number of LR images like DeepSUM, incorporated attention mechanisms [4], setting a new direction for MISR research. The *MISR-GRU* [109] model brought a different perspective by utilizing recurrent neural networks (RNNs) [110] to perform a fusion of information along the temporal dimension. It employed gated recurrent units (GRUs) to adaptively capture temporal dependencies, showcasing its potential in addressing MISR's inherent challenges related to handling temporal variations. Another groundbreaking model was *PIUNET* [134], focusing on permutation invariance and uncertainty in MISR tasks. Further innovations were witnessed with *TR-MISR* [2], which built upon HighRes-Net's foundation. Integrating visual transformers [136] into MISR showcased the capability of attention mechanisms in this domain. Detailed insights into TR-MISR's and PIUNET's approaches are elaborated in the subsequent section. It is worth noting that both of these models currently lead in the post-mortem Proba-V Super Resolution Competition, underscoring the efficacy of these state-of-the-art techniques in addressing MISR challenges.

Multitemporal fusion based on CNNs has been applied to both burst-image super-resolution [10] and video super-resolution [58], with a comprehensive overview provided by Liu et al.[82]. 3D CNNs have shown effectiveness in addressing video SRR challenges [67]. Moreover, contemporary methods are merging SISR and MISR techniques to enhance video streams more effectively [50, 114]. Such methodologies have also found applications in satellite imagery enhancement, particularly in processing satellite image bursts. These strategies typically leverage the consistent and known temporal frequency of input frames. They are often designed to handle moving objects by estimating motion fields, distinguishing them from multi-image fusion techniques that work with an unordered collection of images without specific timestamps, rather than a chronological sequence [100].

2.3 Spectral Fusion

Multispectral super-resolution plays a pivotal role in remote sensing applications, particularly in enhancing the spatial resolution of multispectral images. Among various spectral fusion approaches tailored for this task, pansharpening is a notable technique where a higher-resolution panchromatic channel is leveraged to enhance the resolution of other spectral bands [54]. This pansharpening-inspired strategy was notably employed in the *DSen2* model proposed by Lanaras et al. [74], amplifying the resolution of 20 m and 60 m GSD *Sentiel-2* bands to 10 m GSD. In scenarios devoid of an HR channel, the exploitation of spectral correlations via methods such as 3D convolutions or tensor decompositions becomes prevalent [77, 154]. These techniques aid in recovering lost spatial details while ensuring spectral consistency. While initially explored for hyperspectral images, these fusion techniques have also found applicability in multispectral images, broadening the scope of multispectral SRR methods.

Traditionally, spectral fusion has been conducted on single multi-channel images, aligning with the SISR framework [19]. A notable advancement in bridging the gap between multispectral super-resolution and MISR is the author's proposition of *DeepSent*, as detailed in [126]. Crafted for super-resolving multitemporal series of multispectral Sentinel-2 images, the uniqueness of this CNN lies in its capability to perform information fusion both in the spectral and temporal dimensions. This facilitates the enlargement of multispectral images and elevates all spectral bands to a unified 3.3 m GSD, achieving up to an $18\times$ upsampling factor. In [107], a transformer-based modification to DeepSent was also proposed, significantly reducing the model's parameter count without compromising the reconstruction performance. Empirical evidence showcases DeepSent's superior performance over state-of-the-art techniques, especially in real-world Sentinel-2 image enhancement, thereby extending the frontier of multispectral MISR.

2.4 Prominent Models in Super-Resolution

Super-resolution, encompassing both its single and multi-image variants, has seen the emergence of numerous models. Many of these have contributed to advancing the field, introducing novel methodologies or refining existing ones. In this section, the models that have substantially influenced the landscape of SRR, and particularly the direction of the research, are explored.

2.4.1 FSRCNN

The *fast super-resolution convolutional neural network* (FSRCNN) [24] represents a significant evolution in deep learning-based super-resolution. While many deep learning models aim for an end-to-end mapping from LR to HR images, FSRCNN distinguishes itself through its innovative architecture tailored for computational efficiency without sacrificing reconstruction quality. Unlike its predecessor, SRCNN, which employed an explicit bicubic interpolation step, FSRCNN eliminates this, leading to faster processing. The model's architecture is distinctively structured into several phases: an initial feature extraction phase, a shrinking phase to reduce the feature map dimensions, a series of convolutional layers for further feature extraction, and an expansion phase for upsampling the feature maps. This design ensures efficient feature representation and has set a benchmark for subsequent SISR models. Conceptually, FSRCNN can be viewed as being composed of three primary blocks:

Feature Extraction: At the heart of FSRCNN is its feature extraction phase, which employs a single convolution layer. The aim here is to capture the low-level features of the input LR image. This step ensures that the foundational features, which are essential for subsequent stages, are adequately represented.

Shrinking and Non-linear Mapping: Post feature extraction, FSRCNN introduces a shrinking layer which reduces the number of feature maps, ensuring a compact representation. This is followed by a series of non-linear mapping convolutional layers. These layers delve deeper into the captured features, enhancing their granularity and richness. The motivation behind this block is to refine the features before the expansion phase, ensuring that the model has a comprehensive feature set to work with during the upsampling process and that it is computationally efficient.

Expansion and Deconvolution: The final block of FSRCNN involves the expansion of the refined features back to a higher dimensional space, setting the stage for the deconvolution process. The deconvolution layer [159, 160], or transposed convolutional layer [29, 137], then upsamples the feature maps to produce the HR output. This design choice avoids the explicit bicubic interpolation used in traditional SR methods like SRCNN, resulting in a significant boost in computational efficiency.

The genius of FSRCNN lies in its ability to maintain a balance. While it

adopts a lightweight structure, it does not compromise the quality of super-resolved images. By intricately weaving together feature extraction, non-linear mapping, and deconvolution, FSRCNN sets a robust precedent, emphasizing efficiency without sacrificing performance.

2.4.2 DeepSUM

DeepSUM [96] marked a significant stride in the domain of MISR. Central to its design is a combination of CNN-based SISR techniques with an evolutionary fusion mechanism, enabling the model to assimilate details from multiple images effectively. A standout feature of DeepSUM is the *RegNet* submodule consisting of a series of shared 2D convolutional layers and a global *dynamic convolutional layer* (GDC). Instead of attempting to register the LR images directly, DeepSUM first employs its SISR component to upsample these images. RegNet then takes centre stage, ensuring alignment of these upsampled versions based predominantly on their spatial shifts. Specifically, for each upsampled image, except for a reference one, RegNet learns a filter G_i, which, when convolved with the image by the GDC, aligns it to the reference. The process of generating registration filters can be described as:

$$G_i = f_{\text{RegNet}}(Z^{ILR}_{[0,N-1]}, \theta_{\text{RegNet}}), \quad i = 1, \ldots, N-1, \tag{2.1}$$

which is followed by a registration process performed by GDC:

$$Z_i^{IRLR} = \begin{cases} Z_i^{ILR} & \text{if } i = 0 \\ G_i * Z_i^{ILR} & \text{if } i = 1, \ldots, N-1 \end{cases} \tag{2.2}$$

Here, $Z^{ILR}_{[0,N-1]}$ represents a stack of N input images, θ_{RegNet} denotes the learned parameters of the RegNet model, and $*$ stands for the 2D convolution operation. After the images are aligned to the first image, Z_0^{ILR}, they are then passed to subsequent modules for fusion to form an HR output.

DeepSUM's proficiency was unambiguously showcased when it emerged victorious in the Proba-V Super Resolution Competition, thus attesting to its efficacy in challenging real-world settings. However, one aspect of DeepSUM that warrants attention is its two-phased training approach. Initially, the RegNet submodule undergoes training, post which the overarching model is trained. While this sequential approach was pivotal in achieving the desired outcomes, it introduced additional complexity, particularly concerning training duration and computational requirements.

2.4.3 HighRes-Net

HighRes-Net [22] introduced an innovative approach to MISR by addressing the limitation of handling a fixed number of input images. One of its standout features is the formulation of shared representations and the embedding mechanism.

The backbone of HighRes-Net's feature extraction process is an architecture of its embedding block, represented as emb_θ, which encompasses a convolutional layer followed by two residual blocks, each activated by a *parametric rectified linear unit* (PReLU) [49] function. This architecture ensures consistent and robust feature extraction and is uniformly applied across all LR images independently. To derive a common representation for the set of LR images, the model computes a reference image by taking the median values across the entire set of LR frames. The value of a reference pixel at position (x, y) is defined as:

$$\text{ref}(x, y) = \text{median}(\text{LR}_0(x, y), \ldots, \text{LR}_{N-1}(x, y)). \tag{2.3}$$

This computed reference serves as a composite representation of the scene, synthesized from multiple vantage points. Each individual LR frame is then cohesively embedded with this reference. The resulting embedded states, originally denoted by s_0^i, capture the distinct features of each frame in comparison to the reference:

$$s_0^i = \text{emb}_\theta([\text{LR}_i, \text{ref}]). \tag{2.4}$$

Subsequent to the embedding phase, HighRes-Net unveils its recursive fusion mechanism. This procedure operates on the latent representations, amalgamating the information from multiple frames in a hierarchical manner. In instances where the number of LR frames does not align with a power of two, the model compensates by padding the set with zero-valued views, ensuring the fusion process remains consistent. The culmination of this fusion is a super-resolved image, enriched by the collective information from all the input LRs.

Through its unique approach of shared representation, embedding, and recursive fusion, HighRes-Net has solidified its place as a noteworthy contribution to the evolution of MISR techniques.

2.4.4 RAMS

The residual attention multi-image super-resolution network (RAMS) [111] represents a significant advancement in MISR, integrating attention mechanisms to refine the fusion process. Recognizing that different parts of the input images contribute variably to the super-resolved output, RAMS employs attention maps to amplify the most crucial regions, ensuring better feature utilization.

Central to RAMS is its series of attention blocks dispersed among its super-resolution layers. These blocks generate attention maps that gauge the interdependencies between different regions of the input images. By dynamically modulating the feature maps, RAMS ensures the subsequent network layers prioritize the emphasized regions, resulting in sharper and more accurate outputs.

However, a limitation of RAMS, similar to DeepSUM, is its reliance on 3D convolutional layers in the attention blocks, which fixes the number of input images it can process. Despite this, RAMS's introduction of attention mechanisms into MISR has not only elevated the standard for super-resolved images but also set a precedent for future models blending attention with super-resolution.

2.4.5 PIUNET

PIUNET [134], a significant advancement in the MISR domain, tackled two complex challenges: permutation invariance and uncertainty estimation. Recognizing the potential inconsistencies arising from different orders of input images, PIUNET ensured a consistent super-resolved output regardless of the sequence of its input images. This design is especially crucial in real-world scenarios where the order of image acquisition might differ, making consistent output a priority.

Furthermore, PIUNET brought an added dimension of interpretability to MISR by providing a measure of the uncertainty associated with its predictions. With an uncertainty map accompanying the super-resolved image, it highlights regions where the model is confident and areas where it is less certain. This not only aids in understanding the model's decision-making, but also offers a practical metric to assess the reliability of the super-resolved outputs in various parts of the image.

The architecture of PIUNET skillfully integrates these features, producing a model that excels in the quality of its outputs while also offering insightful

meta-information about its predictions. Its effectiveness is evident from its ranking in the Proba-V challenge, where it occupies the second-best position, highlighting PIUNET's capability and its importance in guiding future MISR research.

2.4.6 TR-MISR

TR-MISR [2], building upon the foundation laid by HighRes-Net, introduced the capabilities of transformers to the MISR domain. The central innovation in TR-MISR is its shift from the recursive fusion mechanism of its predecessor, HighRes-Net, to a transformer-based fusion. This change reflects the growing appreciation for the potential of attention mechanisms in deep learning, especially in tasks requiring detailed feature fusion.

Transformers, originally designed for natural language processing tasks, have been adopted across various domains due to their ability to handle long-range dependencies and intricate relationships between data points. In the context of MISR, transformers allow TR-MISR to effectively merge information from multiple LR images, capturing subtle details and relationships between features across these images.

By using self-attention mechanisms, TR-MISR can dynamically weigh the importance of features from different input images, leading to a more refined and contextually aware super-resolved output. This ability to adjust the focus on different parts of the input images based on their relevance allows TR-MISR to produce superior super-resolved images.

Highlighting its effectiveness, TR-MISR currently leads the post-mortem leaderboard of the Proba-V Super Resolution Competition, marking its position as one of the foremost models in the field of MISR. This achievement showcases the value of transformer architectures in pushing the boundaries in areas like multi-image super-resolution and provides a new standard for future research in this domain.

2.4.7 DeepSent

DeepSent is a model designed to be the first end-to-end architecture to perform a fusion of spatial, temporal and spectral information in the SRR domain [126]. Its aim is to super-resolve all Sentinel-2 bands, which originally have resolutions of 10 m, 20 m, and 60 m, to a uniform resolution of 3.3 m nominal GSD while preserving spectral relations between bands. The

architecture of DeepSent incorporates a feature extraction block and recursive fusion module, taking inspiration from the HighRes-Net architecture. DeepSent can be dissected into twelve branches, each corresponding to a specific spectral band, with the feature extraction block at the beginning of each branch.

The primary motivation behind DeepSent is to efficiently fuse information across both temporal and spectral dimensions to enhance the resolution of multispectral images. The recursive fusion module is employed to merge information initially in the temporal dimension as well as in the spectral dimension, ensuring a comprehensive fusion of data. Thanks to the recursive fusion blocks on both temporal and spectral levels, DeepSent shows significant flexibility when it comes to input data—it can operate on various numbers of LR images (this number can be different for each band) as well as handling scenarios in which some of the bands are not provided.

In the upsampling phase, DeepSent captures and propagates the intrinsic characteristics of each band to the output, ensuring the proper reconstruction of the entire multispectral image at a higher resolution. The final super-resolution module decodes the latent representation of a scene and upsamples it to reach the target resolution of 3.3 m GSD, maintaining the band-specific characteristics based on the latent representation. Through its architecture and fusion mechanisms, DeepSent addresses the challenges of super-resolving multispectral images, marking an advancement in the SRR domain.

2.5 Graph Neural Networks

A recent yet less explored area within super-resolution pertains to the use of GNNs. These networks have demonstrated considerable promise across various applications due to their inherent capability to model complex relational information in graph-structured data [70]. Super-resolution tasks fundamentally revolve around understanding and leveraging intricate relational information present in images [97]. While traditional methods, mainly CNNs, have predominantly focused on spatial dependencies, the true richness of relationships in image data goes beyond just spatial interactions. GNNs, with their ability to capture not just spatial, but more intricate and higher-order dependencies, are poised as potentially more suitable for super-resolution tasks. Their broader relational modelling capabilities can provide an edge in tasks where understanding nuanced relationships is paramount.

The advent of GNNs can be traced back to 2014 with the introduction of the *graph convolutional network* (GCN) [18]. GCNs represent a significant evolution in the realm of neural networks, adapting the foundational principles of traditional CNNs to suit graph-structured data. Unlike CNNs, which are optimized for grid-like data such as images, GCNs are designed to handle data that exists in non-grid formats, where entities have complex interrelationships. In essence, GCNs capture and process the relational information inherent in graphs. They achieve this by employing a form of convolution that respects the topology of the graph, allowing for the effective propagation of information across connected nodes. This unique capability has made GCNs particularly valuable in scenarios where understanding the relationships between data points is crucial e.g. node classification [68], graph classification [164], recommendation systems [157] or relational reasoning [113].

Incorporating the *graph attention networks* (GATs)[138] into the realm of GNNs marked a significant advancement. GATs introduced an attention mechanism that assigns different weights to various edges in a graph, diverging from the conventional uniform information aggregation method in traditional GCNs. This allowed GATs to process varying contributions from different connected nodes based on their relative importance. The unique attention mechanism not only enables the model to highlight salient features but also devalues less important ones, leading to more expressive and discriminative feature representations. This selective aggregation of information, driven by the attention mechanism, enhances GATs' capability to capture informative nodes effectively, finding applications in tasks like person re-identification [5], action recognition [155] or prompt-driven object detection [139].

Recurrent graph networks (RGNs) are another variant of GNNs that process graph-structured data by recursively updating node representations. This recursive nature allows the network to capture deeper and more intricate relationships within the graph. RGNs have been particularly beneficial for video analysis [81], where the temporal sequence of frames can be modelled as a graph, and the recurrent nature helps in capturing the temporal dependencies. Additionally, RGNs have shown promise in modelling social relationships [37], capturing intricate patterns and dynamics within social networks.

A significant milestone in the development of GNNs for computer vision tasks is the introduction of *SplineCNN*. This model presents a generalization of the convolution operation that respects the local structure and spatial characteristics of arbitrary graphs. The distinctive feature of SplineCNN is its use

of continuous B-spline kernels for convolution, effectively applying spline-based filters on graphs and efficiently utilizing spatial information of each node. By parametrizing the filters as splines, the model can better adapt to the characteristics of the graph data. These spline-based filters ensure smooth transformations, which are more expressive and robust against over-smoothing, a prevalent issue in traditional GCNs. SplineCNNs have been effectively used in the task of image graph classification, graph node classification and 3D shape recognition on point clouds [34].

While the use of GNNs in super-resolution is still in its nascent phase, the introduction of the *cross-scale internal graph neural network* (IGNN)[150]—that integrates traditional convolutional layers with the graph-based ones—highlights the immense promise of these networks for resolution enhancement tasks. The design of IGNN is particularly tailored for the SISR problem, proving especially advantageous for images that exhibit repetitive patterns across varying scales, akin to self-exemplar techniques like the one proposed by Huang et al.[52]. It establishes an internal graph structure from the input image and taps into both local and global correlations, employing a cross-scale approach to boost reconstruction quality. However, the application of GNNs for MISR remains an open area of research, presenting a potential avenue for groundbreaking discoveries and advancements in the domain.

Further enriching the landscape of GNNs in SRR is the introduction of the interlayer feature-representation-based GNN for image super-resolution, *LSGNN* [120]. This model, similarly to IGNN, is tailored specifically to tackle the SISR problem. While many CNNs for SISR have primarily focused on broader and deeper architecture designs, they often overlook the detailed information inherent in the image itself and the potential relationships between features captured at different stages of a network. The LSGNN model addresses these gaps by emphasizing the importance of understanding the interdependence between the extracted features of different layers. It introduces a layer feature graph representation learning module that captures this interdependence, enabling the extraction of deeper and more fine-grained image detail features.

2.6 Introduction to GNNs

In the multifaceted domain of machine learning, neural networks have been instrumental in facilitating substantial progress across a wide range of fields, from image and speech recognition to natural language processing. These

networks draw their design inspiration from the neural structure of the human brain, providing them with remarkable abilities to identify patterns and make data-driven predictions. However, traditional neural networks often face challenges when dealing with data that is not regularly structured, as is the case with images or time series, and often overlook the relational information inherent in the input data. This is the intersection where GNNs showcase their strength.

2.6.1 The Concept of Graphs

A fundamental understanding of graphs is essential for grasping the intricacies of GNNs. In the realm of data structures, a graph is defined as a collection of entities, termed as *nodes* or *vertices*, which are interconnected by links known as edges [47, 147]. The power of graphs lies in their ability to represent intricate relationships and systems, from social networks and molecular formations to the vast interconnectedness of web pages [168].

Within the domain of graph theory, various classifications of graphs exist based on their structural properties and the kind of relationships they depict [98]. Notable among them are:

- **Undirected Graphs**: These are graphs where edges do not have a direction. An edge between vertex A and vertex B is identical to an edge between vertex B and vertex A.
- **Directed Graphs (Digraphs)**: In these graphs, edges have a clear direction. An edge from vertex A to vertex B does not reciprocate an edge from vertex B to vertex A.
- **Weighted Graphs**: In such graphs, every edge is assigned a specific weight or cost. This weight can represent various metrics, such as distance, cost, or any domain-specific value.
- **Unweighted Graphs**: Contrary to weighted graphs, all edges in these graphs are identical and do not carry any specific weight.
- **Cyclic Graphs**: These graphs contain cycles, which are closed paths where the starting and ending vertices are the same.
- **Acyclic Graphs**: Such graphs are devoid of any cycles. A special kind of acyclic graph, termed a tree, ensures a unique path between any two vertices.

- **Connected Graphs**: In these graphs, a path exists between every pair of vertices.
- **Disconnected Graphs**: These are graphs where certain vertices might not be reachable from others.
- **Bipartite Graphs**: The vertices of these graphs can be divided into two distinct sets where no two vertices within the same set share an edge.
- **Complete Graphs**: In these graphs, a unique edge connects every pair of distinct vertices.

Mathematically, a graph $\mathcal{G}$ can be defined as a tuple $\mathcal{G} = (\mathcal{V}, \mathbf{H}, \mathcal{E}, U)$, where:

- $\mathcal{V}$ is a set of vertices (or nodes). Each vertex represents an entity in the graph. Formally, $\mathcal{V} = \{v_1, v_2, \ldots, v_n\}$ where n is the number of vertices.
- $\mathbf{H}$ is a matrix representing the features of each node in $\mathcal{V}$. For a given node v_i, its feature vector is represented as $\boldsymbol{h_i} \in \mathbb{R}^F$, where F is the number of features in each node, thus $\mathbf{H} \in \mathbb{R}^{n \times F}$.
- $\mathcal{E}$ is a set of edges that connect pairs of vertices. Each edge represents a relationship or connection between two vertices. Formally, $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$, and an edge e_{ij} exists if there is a connection from vertex v_j to vertex v_i. $\mathcal{E}$ can also be represented as a *sparse adjacency matrix* $\mathbf{A} \in \mathbb{B}^{n \times n}$.
- U is a set of attributes associated with the edges. Here, each edge e_{ij} has an attribute from the set U and such attribute is denoted as u_{ij}. Since those attributes can take different forms for different applications, such as single values (attention coefficients [14]), vectors (relative position to another node [33]) or even matrices (multifaceted attributes in knowledge graphs [113]), their dimensionality is not explicitly defined at this point.

If the graph is undirected, then the order of vertices in the edge pair does not matter, that is, $e_{ij} = e_{ji}$ what implies $u_{ij} = u_{ji}$. If the graph is directed, then the order is significant, thus e_{ij} and e_{ji} are distinct edges.

This understanding of graphs is not only fundamental but also intimately tied to the essence of this research.

2.6.2 Importance of Graph Neural Networks

GNNs represent a pivotal shift in the realm of neural networks. They are a specialized form of neural networks tailored to efficiently handle and process graph-structured data [15, 45]. This differs from traditional neural networks like CNNs or RNNs, which excel when the data can be structured in a regular form, such as images or time series. These structured data types inherently possess a local order and continuity in their relationships, which these networks leverage to capture patterns and dependencies. However, when the data deviates from this structural format, as is the case with graph data, these traditional networks encounter difficulties [7].

Graph-structured data does not adhere to the spatial or temporal consistency of images or time series data. Instead, graphs are non-Euclidean structures where nodes and their corresponding edges can represent diverse and complex systems [168]. The edges connecting the nodes can symbolize a myriad of relationships, from friendships in a social network [94] and hyperlink connections in the World Wide Web [13] to protein interactions in a biological network [6]. The fundamental challenge is to process this irregularly structured data, extract meaningful features, and identify complex patterns that capture the inherent relationships within the graph [40, 46].

GNNs rise to this challenge by working directly with data in a graph format, allowing them to exploit the unique properties and the richness of relationships that graphs offer [15, 112]. Unlike traditional neural networks that may lose relational information when applied to graph data, GNNs are designed to natively preserve and manipulate the relationships between the nodes [68]. They are capable of learning the topology of the graph and the attributes of the nodes and edges [36], which equips them with the ability to infer the intricate and often non-linear relationships within the data [7].

This capability of GNNs opens the door to enhanced performance across a broad array of tasks that involve relational or structured data. In essence, they are adept at tasks that require inference about relationships, relational reasoning, and the propagation of relational information. These tasks span the spectrum from node classification, link prediction, graph classification, and even complex system dynamics prediction [14].

The power of GNNs to handle and learn from graph-structured data signifies a paradigm shift in the realm of deep learning and artificial intelligence [151]. By processing data in its native graph format, they are not only better equipped to handle the data's complexity but also to unlock insights

from the intricate relationships within the graph. It is this transformative capability of GNNs that introduces a new era in machine learning, paving the way for novel applications and research directions.

2.6.3 The Mechanics of Graph Neural Networks

The operational philosophy of GNNs centres on principles known as *message passing* or *neighbourhood aggregation* [36]. This methodology, derived from the fundamental characteristics of graphs, forms the backbone of GNNs, enabling them to process and extract information from graph-structured data in an iterative and layered manner.

In essence, the message-passing framework postulates that a node's representation can be effectively informed by aggregating information from its immediate neighbourhood. This is formalized in a two-step process - a *message function* and an *update function* [112].

Message Function

Initially, each node v_i in the graph is assigned a feature vector $h_i^{(0)} \in \mathbb{R}^F$, derived from its attributes. This feature vector serves as the initial state of the node. As the network progresses through each layer of computation, the nodes perform the message-passing steps, updating their states based on the aggregated messages from their neighbours and their current states. A node v_j is considered a neighbour to v_i if there exists an edge e_{ij} between them, meaning $\mathbf{A}(i,j) = 1$. The capability to perform these operations repeatedly, often referred to as "stacking layers" in GNNs, allows the nodes to gradually incorporate information from an increasingly larger neighbourhood [68]. After several iterations, each node in the graph holds a feature representation that captures not only its own attributes but also the contextual information from its extended neighbourhood. This process is pivotal to learning in GNNs, with the final node representations serving as powerful feature vectors for downstream tasks.

Each node v_i computes a message $\boldsymbol{m}_i$ based on its current state, $\boldsymbol{h}_i$, and the state of its neighbouring nodes. This can be expressed mathematically as:

$$\boldsymbol{m}_i = \bigoplus_{j \in \mathcal{N}(i)} \phi(\boldsymbol{h}_i, \boldsymbol{h}_j, u_{ij}), \tag{2.5}$$

where $\bigoplus$ represents a differentiable and permutation invariant aggregation function (e.g. sum, product, mean), $\mathcal{N}(i)$ is the neighbourhood of node v_i, $\boldsymbol{h}_i$

and $\boldsymbol{h}_j$ are the features of nodes v_i and v_j, u_{ij} is the edge attribute from node v_i to v_j, and ϕ is the message function.

Update Function

Each node v_i updates its state based on its current state and the aggregated messages from its neighbours. This is represented as:

$$\boldsymbol{h}_i' = \gamma(\boldsymbol{h}_i, \boldsymbol{m}_i), \tag{2.6}$$

where $\boldsymbol{h}_i'$ is the updated feature of node v_i, and γ is the update function.

2.6.4 Influential Architectures

GNNs have evolved into a diverse array of architectures and methodologies, each tailored to process and understand graph-structured data in its unique way [7]. Among the vast landscape of GNN architectures, certain types and specific models have been identified as especially influential, not only in the broader research community but also in shaping the direction and methodologies of this research. The differences between these GNN architectures have been primarily defined by how information from a node's neighbours is aggregated and how features describing their type of connection are utilized—the defining aspect of graph neural processing [153].

In this section, emphasis is placed on two fundamental types of GNNs: GCNs and their attention-augmented variant, GATs. Also, SplineCNN is examined as a specific GNN architecture that has been found to be particularly impactful for graphs where spatial relationships are meaningful. While the foundation for processing graph data has been provided by GCNs and GATs, a novel approach has been introduced by SplineCNN that has significantly influenced the research perspective. Each of these architectures has offered unique insights and methodologies for handling graph-structured data, and their contributions to the field have been instrumental in guiding the approaches adopted in this research.

Graph Convolutional Networks

GCNs [18] operate under the principle that nodes closer to each other in the graph space should exhibit similar features. This intuition is materialized by

implementing a convolution operation in the graph domain. The convolution is achieved by applying a propagation rule at each layer, mathematically expressed as:

$$\boldsymbol{h}_i' = \sum_{j \in \mathcal{N}(i)} \frac{\mathbf{W}\boldsymbol{h}_j}{\sqrt{\tilde{d}_i \tilde{d}_j}}, i \in \mathcal{N}(i). \tag{2.7}$$

Here, $\tilde{d}_i$ and $\tilde{d}_j$ indicate node degrees, i.e. a total number of edges from nodes v_i and v_j, respectively. By incorporating a normalization factor based on node degrees, GCNs ensure that feature updates remain stable across layers. It is worth noting that GCNs assume self-connections for each node, hence $i \in \mathcal{N}(i)$. Additionally, $\mathbf{W} \in \mathbb{R}^{F' \times F}$ is the learnable weight matrix which plays a pivotal role in transforming F input features to F' output features. As the network progresses in depth, a node's representation begins to encapsulate features from more distant nodes, effectively capturing information from its multi-hop neighbourhood.

Graph Attention Networks

GATs introduced the concept of attention mechanisms to the world of graph networks [138]. Rather than simply averaging the features of neighbour nodes, GATs perform a weighted average where the weights are learned through the data itself. This ability to adaptively assign importance to neighbours results in a more expressive and flexible model. The attention-guided computation of features for node v_i in GATs can be formulated as:

$$\boldsymbol{h}_i' = \gamma \left(\sum_{j \in \mathcal{N}(i)} \alpha_{ij} \mathbf{W}\boldsymbol{h}_j \right), \tag{2.8}$$

for which the attention coefficient α_{ij} is a softmax-normalized weight across all choices of v_j in the neighbourhood of node v_i. It can be formulated as:

$$\alpha_{ij} = \frac{\exp\left(\text{LeakyReLU}\left(\boldsymbol{a}^T[\mathbf{W}\boldsymbol{h}_i || \mathbf{W}\boldsymbol{h}_j]\right)\right)}{\sum_{k \in \mathcal{N}(i)} \exp\left(\text{LeakyReLU}\left(\boldsymbol{a}^T[\mathbf{W}\boldsymbol{h}_i || \mathbf{W}\boldsymbol{h}_k]\right)\right)} \tag{2.9}$$

where $\boldsymbol{a} \in \mathbb{R}^{2F'}$ is a learnable weight vector serving as an attention mechanism in GATs, and the weight matrix $\mathbf{W} \in \mathbb{R}^{F' \times F}$, similarly to GCNs, is used as a higher-level feature extractor. The $||$ operator stands for concatenation operation, and $\cdot^T$ is a transposition. To achieve a higher level of non-linearity and stabilize the training, the calculation of attention coefficients α_{ij} employs the *leaky rectified linear unit* (LeakyReLU) activation function [88].

Spline Convolutional Neural Networks

SplineCNN [34] represents an important advancement in the field of GNNs, leveraging the power of B-spline basis functions to perform convolution operations on graph-structured data. The key innovation of SplineCNN is the use of continuous kernel functions, defined using these B-spline basis functions, which allows for more flexible and efficient handling of irregularly structured data.

The convolution operation in the SplineCNN takes into account the relative positions of the nodes with respect to their neighbours. This is achieved by using spatial relation vectors, or pseudo-coordinates, to define the kernel functions. These pseudo-coordinates are contained in the set of edge attributes U where the attributes of the edge e_{ij} are denoted by $\boldsymbol{u_{ij}} \in \mathbb{R}^D$ for a D-dimensional scenario. The ability to utilize spatial information is a key advantage of SplineCNN. This feature allows for the processing of graph-structured data in a way that is more similar to the processing of image data in traditional CNNs.

A pivotal aspect of the design and success of SplineCNN is its encoding of spatial relationships through edge attributes. This design philosophy draws parallels with the fundamental operations of traditional 2D convolutions. Just as a sliding window in traditional convolutions identifies neighbouring pixels based on their relative positions without knowledge of global position, SplineCNN discerns the relationships among nodes through the encoded spatial relationships in edge attributes. The graph structure inherently encodes not just the presence of a neighbour but also its direction and distance. This spatial encoding becomes especially pivotal for spline-based convolutions. Unlike traditional convolutional filters that are fixed, the relative spatial information embedded in the edge attributes facilitates a more fluid and adaptive convolutional operation in SplineCNN, making it adept at capturing intricate spatial relationships inherent in graph data.

SplineCNN has been successfully applied to a range of tasks, as demonstrated in the original paper [34]. The authors showcased the effectiveness of SplineCNN in image graph classification, graph node classification, and shape correspondence on meshes. In these applications, the architecture, with its spline-based convolution operator, proved particularly proficient.

SplineCNN operates by aggregating information from neighbouring nodes, with each node's contribution weighted by a kernel function $g_\theta : [a_1, b_1] \times \ldots \times [a_D, b_D] \to R^F$. It is crucial to mention that it performs a normalization step before the convolution operation, Specifically, the edge attributes $\boldsymbol{u_{ij}}$ are

normalized and then rescaled such that they align with the span of the kernel, i.e., $[a_1, b_1] \times \ldots \times [a_D, b_D]$. This ensures that the convolution is performed in a consistent space, leading to more meaningful feature aggregations. The computation of F' output features for the i^{th} node can be mathematically expressed as:

$$\boldsymbol{h}_i' = \frac{1}{|\mathcal{N}(i)|} \Big\|_{s=1}^{F'} \sum_{j \in \mathcal{N}(i)} \boldsymbol{h}_j{}^T g_{\theta_s}(\boldsymbol{u}_{ij}) \tag{2.10}$$

where $\|$ stands for a concatenation operation. The g_{θ_s} function is fundamentally based on D open B-spline basis functions of degree m, denoted as $((N_{1,i}^m)_{1 \le i \le k_D}, \ldots, (N_{D,i}^m)_{1 \le i \le k_D})$ and uniformly positioned based on equidistant knot vectors [106] with $\mathbf{k} = \{k_1, \ldots, k_D\}$ defining the number of basis functions for each dimension:

$$g_{\theta_s}(\boldsymbol{u}) = \Big\|_{r=1}^{F} \sum_{p \in \mathcal{P}} w_{s,p,r} \cdot B_p(\boldsymbol{u}). \tag{2.11}$$

Here, p is a D-element tuple of basis functions taken from the Cartesian product of $\mathcal{P} = (N_{1,i}^m)_i \times \cdots \times (N_{D,i}^m)_i$. Each tuple p has assigned a learnable parameter $w_{s,p,r}$ for r^{th} input feature and s^{th} output feature. $B_p(\boldsymbol{u})$ is the product of the basis functions in p at position $\boldsymbol{u}$:

$$B_p(\boldsymbol{u}) = \prod_{d=1}^{D} N_{d,p_d}^m(\boldsymbol{u}(d)) \tag{2.12}$$

Following the theoretical overview of SplineCNN, it is beneficial to delve into a visual exploration to illuminate its operations. Starting with a one-dimensional context provides an intuitive foundation before progressing to the more intricate two-dimensional scenario. In both cases, the basis functions have a constant degree $m = 2$.

In the one-dimensional context, Figure 2.1 depicts the positioning of individual B-spline basis functions along the continuum. When these functions are combined, they produce a consolidated spline-based surface. The interval $[a_1, b_1]$ marks the kernel's definition range, where the sum of all uniform basis functions equals one. This configuration provides a consistent backdrop for understanding the convolution operation in SplineCNN.

Building upon this, Figure 2.2 shows the outcome when each basis function is multiplied by its corresponding weight, $w_{s,p,r}$. The result is a fully defined kernel within the $[a_1, b_1]$ span. Notably, the kernel's profile deviates from the unity observed in Figure 2.1 due to the influence of the weights.

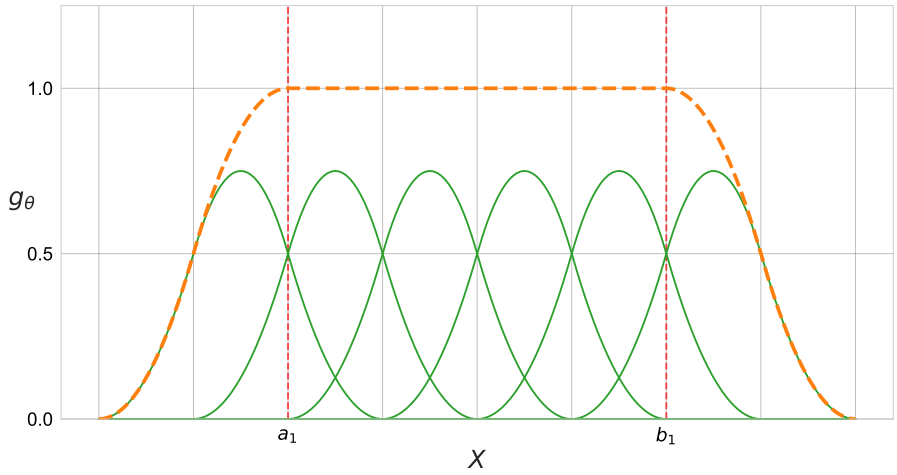


FIGURE 2.1: Positioning of B-spline basis functions in a 1D context. The orange line indicates a surface created by summing the $k_1 = 6$ spline bases of degree $m = 2$, and $[a_1, b_1]$ denote the lower and upper bounds of an interval in which the value of a spline surface equals one.

This modification allows the kernel to discern intricate patterns in the data, offering flexibility in the convolution operation.

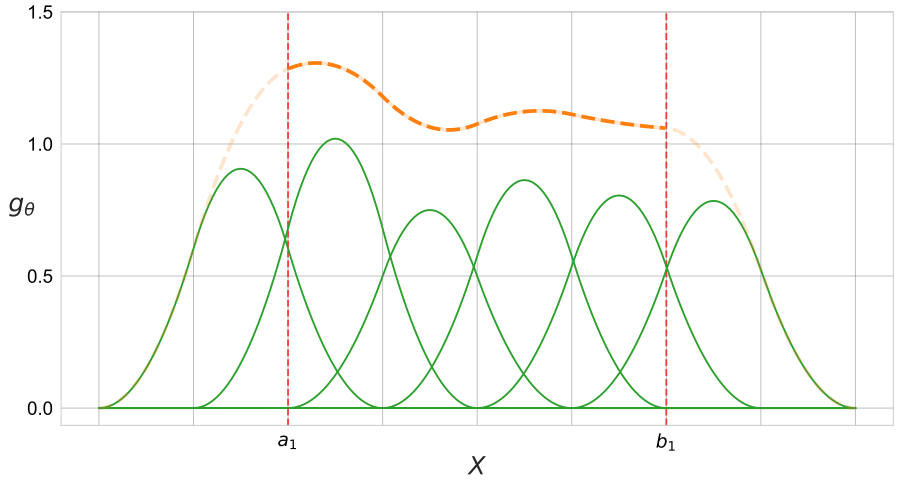


FIGURE 2.2: 1D basis functions multiplied by their corresponding weights. The continuous kernel is defined as a sum of the altered bases over the interval $[a_1, b_1]$.

Transitioning to the two-dimensional scenario ($D = 2$), Figure 2.3 illustrates the spatial arrangement of B-spline basis functions in a plane, culminating in the formation of a continuous 2-dimensional kernel. This kernel operates within a 2D span, defined by $[a_1, b_1] \times [a_2, b_2]$. Unlike the 1D setting

where each basis function is associated with an individual weight, in the 2D context, each weight corresponds to a pair of basis functions, one from each dimension. Thus, with 6 basis functions for each dimension, a total of 36 weights are defined.

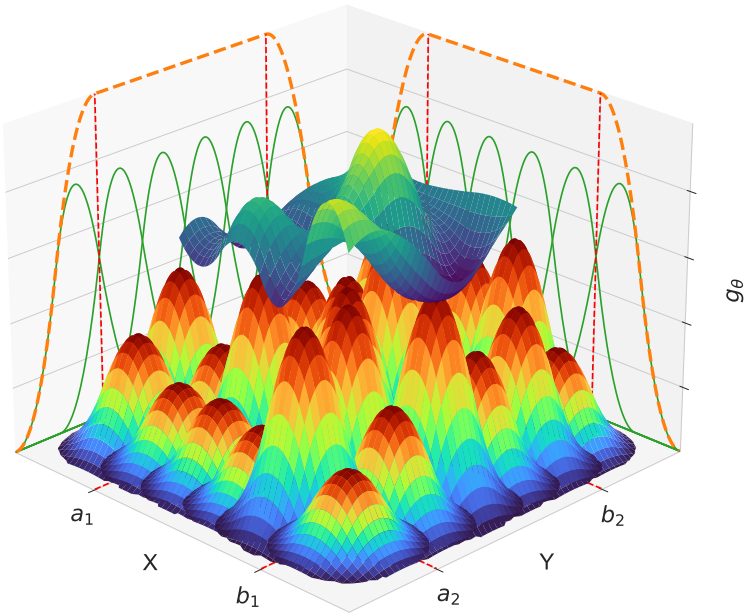


FIGURE 2.3: Creation of a continuous 2D kernel using B-spline basis functions. Here, each product B_p of basis functions in p is multiplied by its corresponding weight w_p and summed within a given interval to create the continuous kernel.

Lastly, Figure 2.4 demonstrates the application of the 2D kernel to a specific node in a graph. This visualization provides insight on how SplineCNN processes graph-structured data, drawing parallels to traditional CNNs' handling of image data. Through these illustrations, the capabilities and potential of SplineCNN in managing graph-structured datasets with irregularly located nodes become evident.

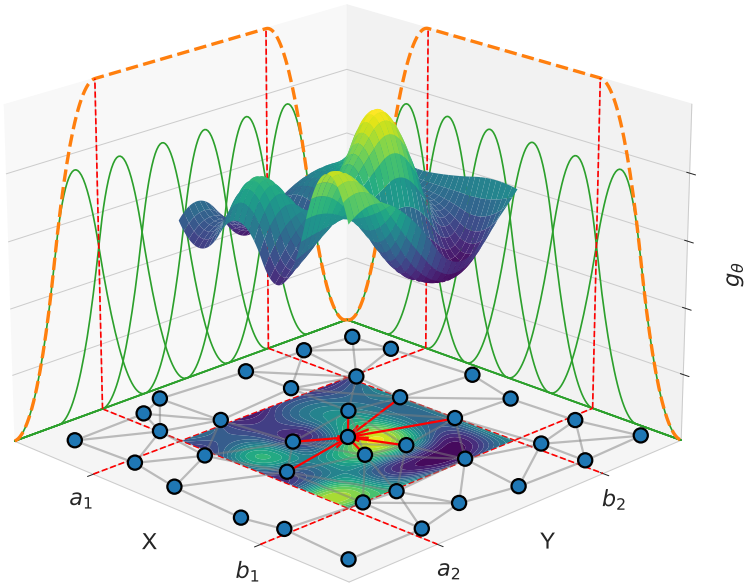


FIGURE 2.4: Application of the 2D spline-based kernel to a specific node in a graph. The red arrows indicate which neighbours within the kernel interval propagate their information to the centre node.

Chapter 3

Architecture Design

Super-resolution, particularly within the scope of MISR, is an ill-posed problem in optimization theory [100]. Addressing this challenge necessitates maximizing the information encapsulated in the input data to achieve high-quality outputs. A balance between effective data representation and algorithmic design is pivotal for harnessing this information.

In this chapter, the primary focus lies on the data creation process, elucidating the transformation of a stack of input images into a single, unified graph. The subsequent sections unravel the evolutionary sequence of model architectures. This journey commences with the foundational *MagNet* model, transitioning through its successive counterparts: *MagNet++* and $MagNet_{enc}$. The sequence culminates in the most advanced model, *MagNAt*. Further, this chapter introduces two additional models derived from *MagNAt*, namely $MagNAt_{no\_reg}$ and $MagNAt_{lead}$, each aimed at investigating specific aspects of the image super-resolution problem. This progressive evolution of MagNAt began with the foundational principles laid out by MagNet. With each subsequent model, new features and improvements were incorporated, highlighting the step-by-step advancements and providing insights into the motivations behind each architectural refinement.

3.1 Converting a Stack of LR Images into a Single Graph

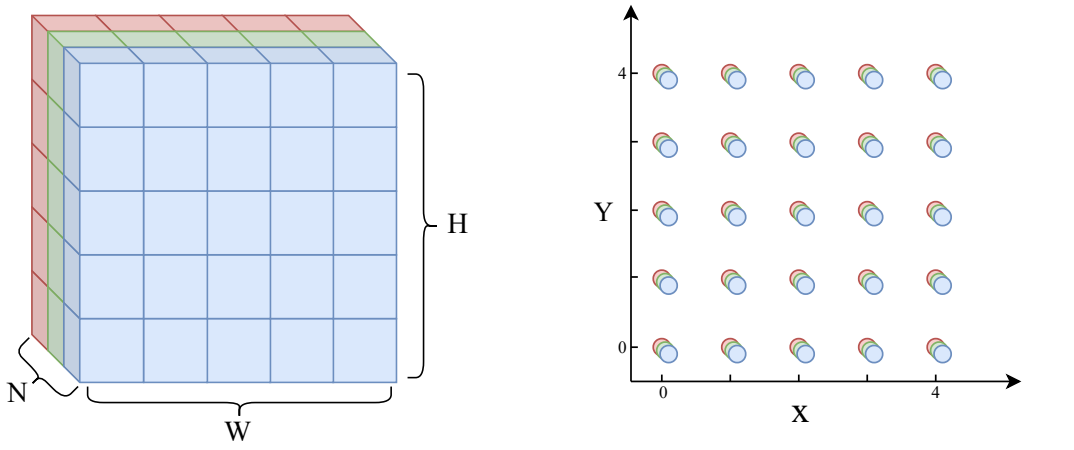
This section elucidates the transition from conventional image-based data structures to a graph representation, marking not merely a change in format but a fundamental shift in data representation. Initially, each pixel is transformed into a node on a mutual 2D plane, a step referred to as node positioning. Subsequently, these node positions are adjusted to account for shifts between LR images. Finally, nodes are interconnected to fabricate the

graph structure. This unique approach amalgamates multiple LR images into a single integrated graph, ensuring a lossless transition while preserving all original information.

3.1.1 Node Positioning

Establishing a meaningful position for each node within the eventual graph is paramount for preserving the inherent spatial relationships of the original images. At this preliminary stage, the challenge lies in mapping pixels to nodes on a unified plane, ensuring that the relative positions of these nodes reflect the spatial coherence of the original image.

To construct a graph, the primary components are initially defined. As introduced in Section 2.6.1, a graph $\mathcal{G}$ can be represented as $\mathcal{G} = (\mathcal{V}, \mathbf{H}, \mathcal{E}, U)$. At this initial stage, two of these components are defined: the set of vertices $\mathcal{V}$ and the matrix of F_{in} features for each node, $\mathbf{H} \in \mathbb{R}^{n \times F_{in}}$. The vertices are derived from the pixels of the LR images. Meanwhile, the input features, $\mathbf{H}$, depend on the number of channels in the LR images. For instance, in the case of RGB images, each pixel has three channels (Red, Green, and Blue); hence, each node in the graph would have three input features. Generally, if the LR images have F_{in} channels, each node would have the same number of input features.



(A) A stack of $N = 3$ input images of dimensions 5×5.

(B) Conversion of pixels to nodes on a mutual 2-dimensional plane.

FIGURE 3.1: Illustration of the process of converting a stack of LR images into a single graph.

As visualized in Figure 3.1, every pixel from the given set of LR images, shown in subfigure (A), is represented as a node on this plane. Its position on

the plane, depicted in subfigure (B), directly corresponds to its original position within its respective image. Specifically, if a node corresponding to the i^{th} pixel is denoted as v_i, its position on the plane can be defined by a function $loc : \mathcal{V} \rightarrow \mathbb{N}^2$ and the resulting node position $loc(v_i) = (x_i, y_i)$ corresponds to indices of a pixel from an image in a matrix form, thus it maintains the spatial integrity between the graph and the corresponding LR image. Hence, x_i and y_i are constrained by the image dimensions: $x_i \in \{0, \ldots, W-1\}$ and $y_i \in \{0, \ldots, H-1\}$, where W and H represent the width and height of the LR images, respectively.

Given N input images, each of size $W \times H$, there are in total $n = N \times W \times H$ nodes on this plane. However, it is crucial to note that at each occupied position on the plane, there are N vertices, each from a different image. Considering the nature of images and their grid-like structure, the closest horizontal neighbours of a pixel at (x, y) would be at $(x \pm 1, y)$, and the vertical neighbours would be at $(x, y \pm 1)$.

By this definition, at this initial phase, the node positions are direct representations of the pixel positions in the LR images without any adjustments. Subsequent steps involve refining these positions and establishing connections between these nodes to form a cohesive graph.

3.1.2 Displacement Calculation

The process of determining displacement vectors, essential to the node positioning strategy, seeks to identify both the magnitude and direction of each LR image's deviation concerning a reference image. This alignment, known as image registration, can be achieved using traditional registration algorithms or by leveraging neural networks specifically designed for this task. The techniques employed for registration can be broadly categorized into full-pixel and sub-pixel methods.

Full-pixel registration offers a direct methodology, translating pixels to align images while preserving their inherent values [16]. This approach is computationally efficient but might overlook the detailed nuances present at sub-pixel levels. On the other hand, sub-pixel registration delves deeper, adjusting pixel values to achieve more refined alignments, though at the potential cost of introducing noise or distortions [127].

Several traditional methods have been employed for image registration. Cross-correlation [71] serves as a foundational technique. Feature detection, such as SIFT [83], is crucial for facilitating feature matching methods [169].

Discrete Fourier transform (DFT)–based techniques also offer refined alignments [42]. In the realm of deep learning, models like RegNet [96] and ShiftNet [22] have emerged, along with convolutional networks designed specifically for geometric transformations in image registration [21]. The choice between these methodologies largely depends on the application's specific needs, available computational resources, and the desired level of precision.

In many MISR models, the inherent challenge is that they lack direct input information about sub-pixel shifts. Instead, these models must deduce such granular details independently in their feed-forward pass, often missing out on the nuanced information. In contrast, in this research, an efficient sub-pixel image translation registration by cross-correlation has been selected [42]. This choice was influenced by its notable advantages of relatively low computation time and high accuracy.

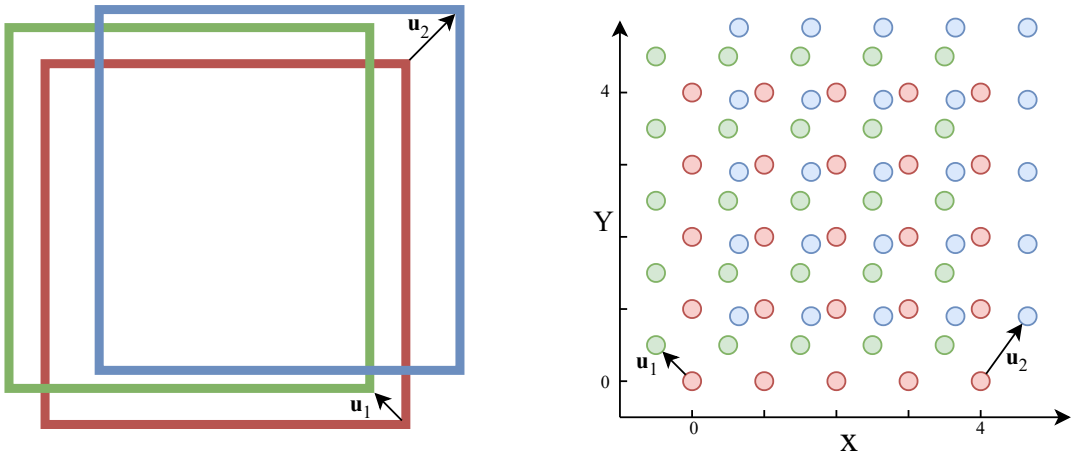
Regardless of the chosen method, the primary objective remains consistent: to determine the sub-pixel displacement for each LR image, ensuring the nodes' accurate positioning in the graph. Formally, for an ensemble of N images, where the image at index $i = 0$ is chosen as the reference, the displacement vectors $\vec{u_i}$ for each of the $N-1$ remaining images are computed. Each vector $\vec{u_i}$ is a two-dimensional representation capturing the horizontal and vertical shifts, given by $\vec{u_i} = [\bar{x_i}, \bar{y_i}]$, where i ranges from 0 to $N-1$, and $\bar{x_0} = \bar{y_0} = 0$. Building on this, the adjusted position of the node v_j, accounting for the displacement of its associated image, can be represented as $loc'(v_j) = loc(v_j) + \vec{u}_{img(v_j)}$. Here, $img(v_j)$ is a function that maps a node v_j to its associated image index, hence $img : \mathcal{V} \to \{0, \ldots, N-1\}$. Given $loc(v_j) = (x_j, y_j)$, this translates to:

$$\text{loc}'(v_j) = \begin{bmatrix} x_j + \bar{x}_{\text{img}(v_j)} \\ y_j + \bar{y}_{\text{img}(v_j)} \end{bmatrix} = \begin{bmatrix} x'_j \\ y'_j \end{bmatrix}, \quad j \in \{1, \ldots, n\}. \tag{3.1}$$

This mapping ensures that the spatial relationships between nodes are consistent with the displacements of their respective LR images.

As depicted in Figure 3.2, the left image (subfigure A) visualizes the resulting shift vectors derived from the image registration algorithms. Meanwhile, the right image (subfigure B) illustrates how these vectors have been applied to adjust the nodes' positions within the graph.

It is important to note that in more complex scenarios, where images exhibit rotational deviations relative to each other, the procedure can be extended to compute a distinct shift vector for each node using algorithms such as optical flow [163]. This approach accommodates rotational and other



(A) Calculated shift vectors using image registration algorithms.

(B) Node positions adjusted using the computed shift vectors.

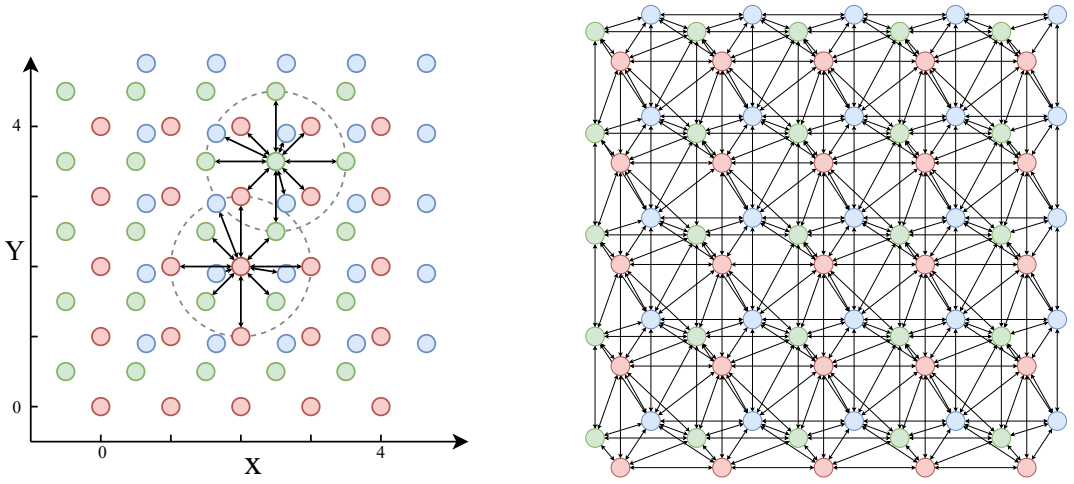
FIGURE 3.2: Illustration of the process of computing shift vectors and adjusting node positions based on these shifts.

complex transformations, determining spatial displacements on a per-node level. However, this method is not adopted in this research as the employed datasets exhibit only translations between the input images. Additionally, optical flow is sensitive to noise, potentially causing significant errors in images with high temporal variations, and is computationally more demanding [167]. Hence, this research focuses on addressing translations alone, utilizing the described displacement calculation mechanism to ascertain and compensate for spatial deviations between different LR images.

3.1.3 Graph Construction

In constructing the graph, the first critical step, post-node positioning, is the creation of edges, effectively binding the individual nodes into an integrated graph structure. As visualized in Figure 3.3, the establishment of connections is primarily dictated by proximity, employing the concept of Euclidean distance. Formally, for any two nodes v_i and v_j with positions $loc'(v_i) = (x'_i, y'_i)$ and $loc'(v_j) = (x'_j, y'_j)$, the Euclidean distance $d : V, V \rightarrow \mathbb{R}$ between them is defined as:

$$d(v_i, v_j) = \sqrt{(x'_j - x'_i)^2 + (y'_j - y'_i)^2}. \tag{3.2}$$



(A) Connecting procedure on two arbitrary nodes for radius $r = 1$.

(B) Fully prepared graph.

FIGURE 3.3: Visualization of node connection process.
The illustration provides a depiction of the graph construction process. The left subfigure (A) demonstrates the method of connecting nodes that fall within a specified radius, while the right (B) displays the completed graph, integrating both nodes and edges.

Using this distance metric, the graph's adjacency matrix $\mathbf{A}$ and the edge set $\mathcal{E}$ are updated as:

$$\mathbf{A}(i,j) = \begin{cases} 1 & \text{if } d(v_i, v_j) \leq r \\ 0 & \text{otherwise} \end{cases}. \tag{3.3}$$

Here, an edge e_{ij} is present if $\mathbf{A}(i,j) = 1$ and absent otherwise. The threshold r, set to one in this context, marks the boundary for node connection. Nodes within or on the circumference of a circle with radius r centred around any given node are regarded as neighbours.

By setting $r = 1$, the connections predominantly align with the pixel width, naturally preserving the spatial relationships inherent within individual images. Such a choice is both intuitive and computationally efficient: a broader radius would compound the number of connections, amplifying the computational burden.

With the foundational graph in place, enriched with the set of connections $\mathcal{E}$, edge attributes U are added not only to boost the graph's descriptive potential but also to fulfil the input requirements of the spline-based convolution highlighted in section . These attributes capture the spatial offsets between paired nodes, providing essential spatial relationship data that the spline-based convolution needs to operate efficiently.

For any two nodes v_i and v_j interconnected by edge e_{ij}, the attribute $\boldsymbol{u_{ij}}$ records their spatial displacement:

$$\boldsymbol{u_{ij}} = loc'(v_j) - loc'(v_i) = \begin{bmatrix} x'_j - x'_i \\ y'_j - y'_i \end{bmatrix} = -\boldsymbol{u_{ji}} \tag{3.4}$$

With these edge attributes, the graph's depth transcends mere structural connections, shedding light on relative spatial orientations between nodes. These attributes grant crucial context, particularly during graph processing or analysis, as the relative node positions can be inferred directly from edge data.

In conclusion, these steps come together to create a structured graph that accurately represents the original set of LR images. This representation not only maintains the inherent spatial dynamics and displacements, but also enriches the spatial relational details embedded within. Unlike the conventional matrix-based methods, where images are simply stacked in a sequence without adding any new information beyond raw pixel values, this graph-based approach is a significant move toward affirming the first thesis of this dissertation.

3.1.4 Benefits of Graph Data Representation for MISR

In theory, the graph-based data representation method offers numerous advantages that make it particularly suitable for MISR tasks. These potential benefits, including *permutation invariance*, the ability to handle *varying quantities of inputs* or *data heterogeneous in size*, and *flexible relationship modelling*, are discussed in this section.

Permutation Invariance

Permutation invariance in MISR signifies that the sequence of input LR images should not affect the super-resolution results. Regrettably, not every MISR approach inherently embodies this attribute. Many conventional methods necessitate a static ordering of input images, inadvertently inducing dependence on the specific sequence of LR inputs.

While contemporary techniques, particularly PIUNET and TR-MISR, do ensure permutation invariance, they do so through distinct architectural strategies. For instance, PIUNET employs self-attention mechanisms in the temporal dimension and shared convolutional layers. Conversely, TR-MISR utilizes

a transformer-based fusion mechanism to ensure input order does not affect outcomes.

However, the graph-based representation inherently guarantees permutation invariance by placing each pixel from all LR images on a unified two-dimensional plane, irrespective of their origin. Achieving this at the data-creation level eliminates the requirement for special architectural decisions to tackle this issue.

Varying Number of Input Images

A distinctive feature of the graph-based representation is its adaptability to various numbers of input images. In contrast, some traditional methods, such as RAMS, DeepSUM, and, to a certain extent, HighRes-Net, demonstrate limitations. Specifically, RAMS and DeepSUM can only handle the exact number of LR images during both training and inference as was available during their respective training sessions. Meanwhile, HighRes-Net frequently resorts to padding its input matrix with blank images to round up to the nearest power of two.

In the graph-based approach, the accommodation of varied numbers of LR images is straightforward and does not demand architectural adjustments. Regardless of the input count, they all form a single graph, with the only variable being the graph's density. However, it is worth noting that while this model demonstrates scalability, an exponential rise in complexity accompanies each additional input image. For an in-depth discussion on the computational implications, readers can refer to Section 6.5.

Handling of Heterogeneous Inputs

When it comes to real-world SRR tasks, there is often a mix of input images varying in aspects like resolution, size, or orientation, especially in remote sensing applications [74, 126]. Traditional MISR methods might grapple with this diversity, necessitating extra preprocessing steps to create a uniform dataset. This can be cumbersome and resource-intensive. In stark contrast, the graph-based approach can natively process a mix of different numbers, sizes, and orientations of LR images, assuming a proper node positioning procedure. This inherent flexibility means that whether the inputs are uniformly oriented LR images or a mix of varying resolutions and orientations, the graph-based model can process them without problems, positioning it as a versatile choice for diverse scenarios. However, a comprehensive

exploration of the adaptability to heterogeneous inputs remains beyond the scope of the current research.

Flexible Modelling of Relationships

Graphs are renowned for their ability to capture relationships between entities [15], an attribute that is highly beneficial in the context of MISR. In the adopted graph-based representation, nodes are interconnected through edges, which can represent various relationships, from spatial proximity to pixel value similarity. Such dynamic connection modelling enables the discernment and utilization of intricate interpixel relations, potentially augmenting the super-resolution results.

In the research presented, these relationships are characterized as relative displacements between pairs of pixels. While this serves as the primary relational feature for the current focus, in more advanced scenarios, other relational characteristics can be encoded. This is reminiscent of the approaches in more complex models, where, for instance, the significance of a node relative to its neighbours is encoded, hinting at the versatility of the graph-based approach. However, a detailed exploration of more advanced encoding methods is discussed in Section 3.5 in the context of the MagNAt architecture.

3.2 MagNet: A Comprehensive Analysis and Proof-of-Concept

MagNet [123], developed to support the propositions put forth in this thesis, pioneers in utilizing GNN for the purpose of MISR. In the briskly advancing domain of image super-resolution, MagNet introduces a unique application of GNNs for MISR, carving out a novel path for research in this sphere.

The architecture of the MagNet model seamlessly integrates Graph Neural Networks with the framework of the SISR model FSRCNN [25]. Even though FSRCNN was originally conceptualized as a SISR model, its structure lends itself efficiently to MagNet's needs. The graph data representation in MagNet essentially translates the multi-image super-resolution task into a single-graph super-resolution. This, in essence, means that the model, while processing a graph composed of various images, treats it similarly to a single image scenario. By harnessing this graph-centric representation, MagNet adeptly merges insights from various LR images, enhancing its overall performance in the MISR challenge.

As MagNet entered the MISR domain, its functionality and prowess were put to the test through an array of meticulously planned experiments. These tests did more than just highlight its operational efficiency; they emphasized its capability and prospective contributions in managing the intricate facets of multiple-image super-resolution.

It is pivotal to note that MagNet's present incarnation serves primarily as a proof-of-concept, illustrating the practicality of using GNNs in MISR. The initial rounds of evaluation were predominantly anchored on simulated data, extracted from a diverse pool of standard benchmark datasets. The primary reason for initially using simulated data, as detailed in Chapter 4, describing the employed datasets, was to assess the suitability of the architecture for the MISR problem in a controlled and more manageable testing environment.

However, leaning solely on simulated data is but a preliminary phase in MagNet's developmental journey. The inherent limitations of simulated data make it imperative to test the model in real-world scenarios. This transition to real-world data testing, detailed further in section 6.2, is expected to shed more light on the model's potential and areas needed for refinement.

In essence, while the simulated data-centric proof-of-concept phase was instrumental in unveiling MagNet's potential, the move towards real-world data promises a more comprehensive understanding of its capabilities, suggesting an imminent evolution in the world of MISR.

3.2.1 Dissecting the Architecture of MagNet

As depicted in Figure 3.4, the architecture of MagNet is influenced by the SISR network FSRCNN, discussed in Section 2.4.1. The proposed method of data representation, which merges multiple LR images into a solitary graph, makes this similarity possible. This unified graph can be perceived as a 'single super-image', possessing richer information than a conventional LR image used in SISR.MagNet employs spline-based convolutions, as introduced in SplineCNN (Section 2.6.4). This forms a composite model integrating traditional convolution methods with a message-passing technique inherent to GNNs. This hybrid design empowers the model to harness both the features and connection weights present in graph nodes and the spatial information often disregarded in standard GNNs, due to their relational-centric scenarios.

The very first layer in MagNet, serving as the feature extraction component, employs spline-based convolution kernels with a configuration of

$\mathbf{k} = \{3,3\}$. In this setup, the convolution utilizes two sets of B-spline basis functions, one for each spatial dimension, with each set having three basis functions. This configuration dictates the density of the weights within the kernel's spatial span and not the span itself. The spatial extent of this kernel, specifically from -1 to 1 in both dimensions, adheres to the connection radius described in section 3.1.1. Thus, it ensures that the convolution is influenced only by the nodes within this defined spatial range around each node. With this convolution, the input channels of each node are transformed to produce 56 features. The choice of outputting 56 features from this layer draws its inspiration directly from the FSRCNN architecture.

Subsequent to this, MagNet integrates a shrinking layer to streamline the architecture. This layer efficiently consolidates the previously mentioned 56 features, reducing them to a more compact 16 features, once again, a design choice influenced by FSRCNN. To achieve this feature reduction, the architecture employs 1×1 ($\mathbf{k} = \{1,1\}$) spline-based kernels. Notably, these kernels incorporate only a single pair of B-spline basis functions—one for each spatial dimension. Despite being influenced by just a singular weight

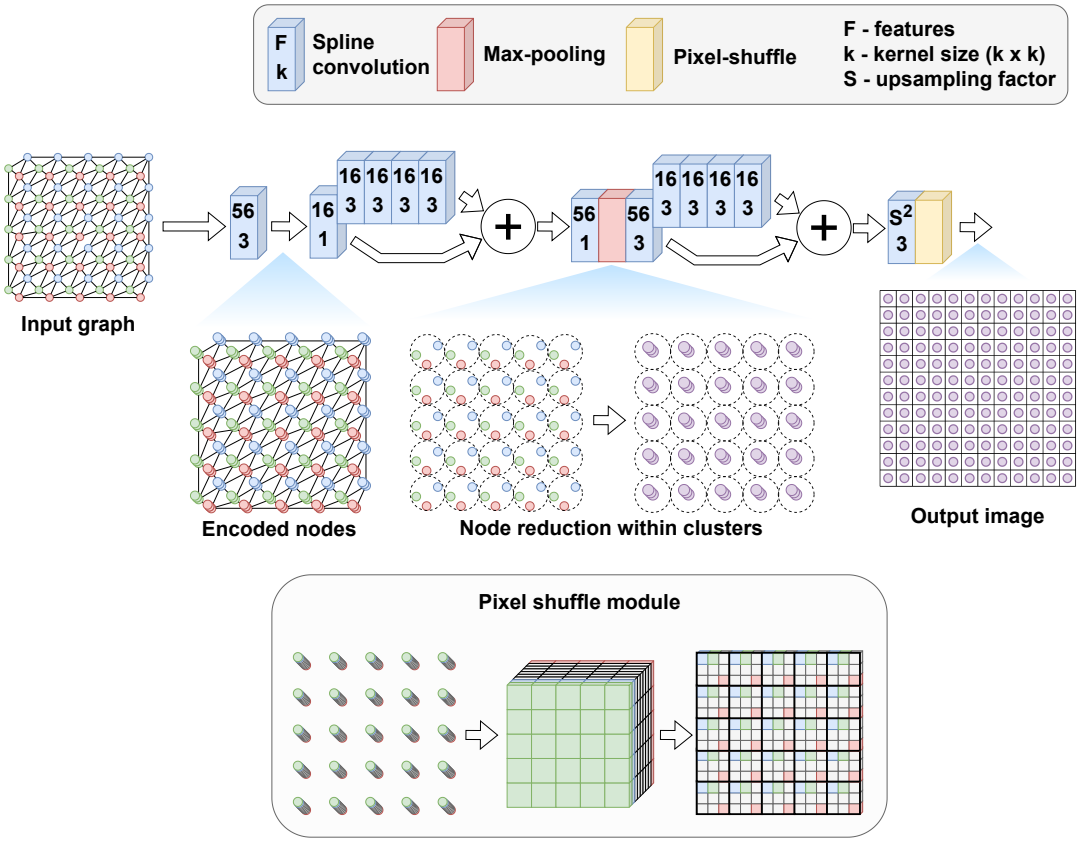


FIGURE 3.4: The architecture of MagNet.

for each node, this kernel design is ingeniously capable of assimilating information from different neighbouring nodes, a characteristic that distinctively sets it apart from conventional 1×1 convolutional layers. Subsequent to this, the architecture includes a convolutional block made up of four spline-based convolutional layers, interlinked with a skip connection to counteract vanishing gradient issues in deep neural networks [48].

In the next phase, the feature count is increased back to 56 through a single 1×1 spline-based convolutional layer. This is followed by a max-pooling operation applied to predefined node clusters. To understand the cluster formation, one can refer to the function introduced in Section 3.1, denoted as $loc(v_i)$. This function yields the discrete location of a node v_i. Considering multiple LR images, for each discrete position (x, y) where $x \in \{0, \ldots, W-1\}$ and $y \in \{0, \ldots, H-1\}$, there are N nodes, each corresponding to one of the N LR images. These nodes are assembled into a cluster based on their mutual spatial location, as defined by $loc(v_i)$.

By employing this clustering strategy, the architecture generates $W \cdot H$ clusters, and each cluster encompasses N nodes. Here's where the transformation becomes significant: the max-pooling operation consolidates each of these clusters, which consists of N nodes, into a single representative node. Consequently, the total number of nodes is reduced from $N \cdot W \cdot H$ to just $W \cdot H$.

This means that post max-pooling, for every discrete position (x_i, y_i) given by the function $loc(v_i) = (x_i, y_i)$, there exists only one corresponding node. Effectively, the architecture has transformed the graph into a regular, grid-like structure. This uniformity mirrors a standard image grid, with each node denoting a pixel-like entity at a particular spatial location. Such a structure not only makes further processing more streamlined but also enhances MagNet's scalability and adaptability, ensuring that the model remains efficient irrespective of the number of input images.

After the max-pooling step, the graph assumes a regular, grid-like structure. Before transitioning into a matrix format, MagNet conducts further spline-based convolutions on this grid-like graph, incorporating a residual block and a single spline-based convolution modifying the feature depth to $\mathcal{S}^2$, with $\mathcal{S}$ being the upsampling factor of the model. Upon the conclusion of these convolutional processes, the graph is converted into a matrix form, preserving the spatial integrity of each pixel according to the $loc(v_i)$ function, and preparing the data for the pixel-shuffle operation [115]. The operation works by reshuffling the channels of the input tensor to the spatial

dimensions. For an input tensor with shape $[H, W, F \cdot \mathcal{S}^2]$, where F is the number of features, the pixel-shuffle operation outputs a tensor with shape $[H \cdot \mathcal{S}, W \cdot \mathcal{S}, F]$. In the context of MagNet, given that the matrix representation post convolutions is of shape $[H, W, \mathcal{S}^2]$, the pixel-shuffle operation transforms it into $[H \cdot \mathcal{S}, W \cdot \mathcal{S}, 1]$. This effectively expands the spatial dimensions by a factor of $\mathcal{S}$ and produces a single channel, resulting in a representation resembling a high-resolution single-channel image.

In summary, the MagNet architecture, through its unique incorporation of graph-based techniques and spline-based convolutions, presents a distinct approach to addressing the MISR challenge. It effectively consolidates multiple LR images into a cohesive graph structure, applies convolutions to harness rich node information, and subsequently employs max-pooling and pixel-shuffling to produce a higher-resolution image.

3.2.2 Limitations of MagNet

The MagNet architecture, while being a promising approach to SRR tasks, poses certain challenges that need consideration. Central to these challenges is the application of max-pooling before the upsampling step. The primary objective of max-pooling in this context is to transform the graph into a more structured, grid-like configuration, essential for its conversion into a matrix form suitable for the pixel shuffle operation. However, while this transformation preserves the spatial boundaries, it reduces the graph's node count, potentially diminishing the amount of information embedded in the graph.

Consequently, the following pixel shuffle operation, applied to this now simplified matrix, might not fully utilize the complete spectrum of information present in the graph's state prior to the max-pooling. In light of these limitations, the next section introduces an evolved architecture, MagNet++, which aspires to retain the foundational strengths of MagNet while addressing its weaknesses, aiming for an optimized super-resolution output.

3.3 Graph-Based Upsampling in MagNet++

Built on the foundations of the MagNet model, a new and refined architecture named MagNet++ was designed and introduced by the author of this dissertation in [122]. While retaining core elements such as the feature extraction layer, shrinking layer, and a convolutional block with skip connections—all leveraging spline-based convolutions—its primary distinction lies in its

upscaling technique. This method is optimized for the use of input node information and is crafted to reduce the likelihood of information loss, thereby exploiting the unique aspects of this data representation more effectively.

In MagNet, the upscaling process involved the max-pooling of nodes corresponding to the same pixel position of the LR images. This operation reduced the number of nodes by a factor of N, transforming the nodes into a grid-like configuration. The transformed data was processed by a convolutional block, followed by a pixel shuffle operation to enhance the spatial resolution of the image.

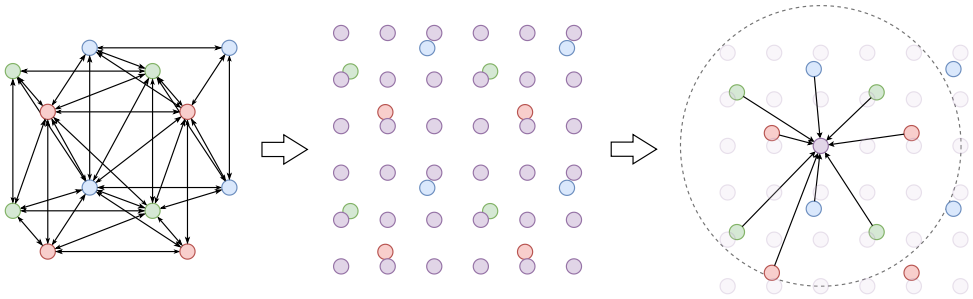


FIGURE 3.5: Schematic representation of the graph transformation and connection establishment for the bipartite upsampling procedure. The original nodes (red, green and blue) establish directed edges to the newly overlaid nodes (purple), ensuring a one-way flow of information in the new bipartite structure.

MagNet++, on the other hand, embraces a different approach. Instead of transforming the graph into a matrix prior to resolution enhancement, a fully graph-based approach is adopted here. This method utilizes a bipartite graph structure. To construct it, initially, the connections within the original graph are entirely removed, rendering it a null graph—a type of graph consisting of nodes but lacking any edges. Next, a new null grid-like graph is crafted having of $W \cdot H \cdot \mathcal{S}^2$ nodes, with a distance between each pair of neighbouring nodes (vertically or horizontally) equal to $\mathcal{S}^{-1}$. This new graph is then overlaid onto the original, now edge-less one, so that their centres are aligned. The connection strategy utilizes one-way directed edges, wherein the original nodes, retaining their information, connect exclusively to the new nodes. The chosen radius, $r = \sqrt{2}$, ensures that each new node establishes a connection to the original nodes, thus it avoids rendering a disconnected graph. A smaller radius, such as previously used $r = 1$, could risk nodes positioned centrally between original pixels (across both dimensions)

being too distant to form any connection. This could be particularly problematic in scenarios with a low count of input images or minimal displacements between them. These new edges ensure a one-way flow of information. Consequently, the newly overlaid nodes act solely as targets, receiving information from the original nodes without initiating any connections of their own. Visualization of this process can be seen in Figure 3.5, which provides a comprehensive schematic representation of the graph transformation and connection establishment.

In this bipartite graph structure, a single spline-based convolutional layer activated by the PReLU function is applied to gather information from the source nodes. Next, the overlaid grid-like graph is separated from the original one and transformed into a tensor of shape $[\mathcal{S} \cdot W \times \mathcal{S} \cdot H]$. With this method, information loss from the input graph is potentially reduced, addressing a notable limitation, discussed in Section 3.2.2, of the MagNet architecture.

3.3.1 MagNet++ Architecture

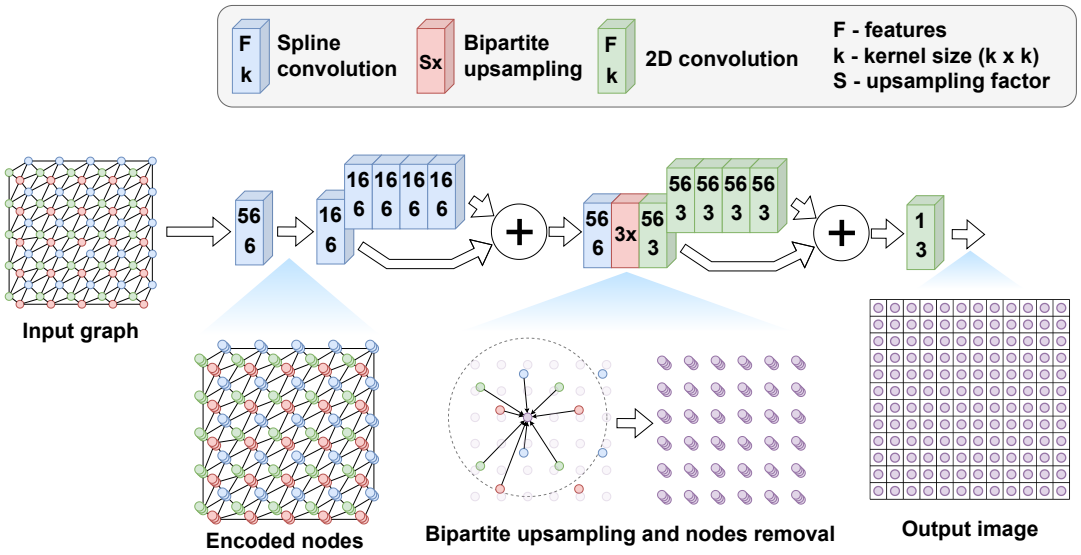


FIGURE 3.6: The architecture of MagNet++.

With the return of a single rectangular tensor from the bipartite upscaling operation, standard 2D convolutional layers can be applied instead of the spline-based ones to reduce the computational overhead of the model. The data is further processed through a single residually connected convolutional block, followed by a convolutional layer with a single kernel, producing the

final super-resolved image. The architecture of MagNet++ is shown in Figure 3.6.

A noteworthy advantage of MagNet++ is its flexibility with the shape and completeness of the input LR images. Unlike many models that necessitate rectangular input, MagNet++, theoretically, is able to accommodate non-rectangular images, providing greater adaptability in various application contexts. Furthermore, it also permits the processing of images with specific nodes removed, such as masked nodes. This functionality is particularly beneficial in cases like remote sensing applications, where certain nodes might depict unwanted or noise-inducing elements, such as clouds or uncaptured pixels. By introducing these new features, MagNet++ demonstrates a significant stride forward in the field of super-resolution imaging, offering a promising prospect for further advancements.

3.4 MagNet$_{enc}$ and Improved Feature Extraction

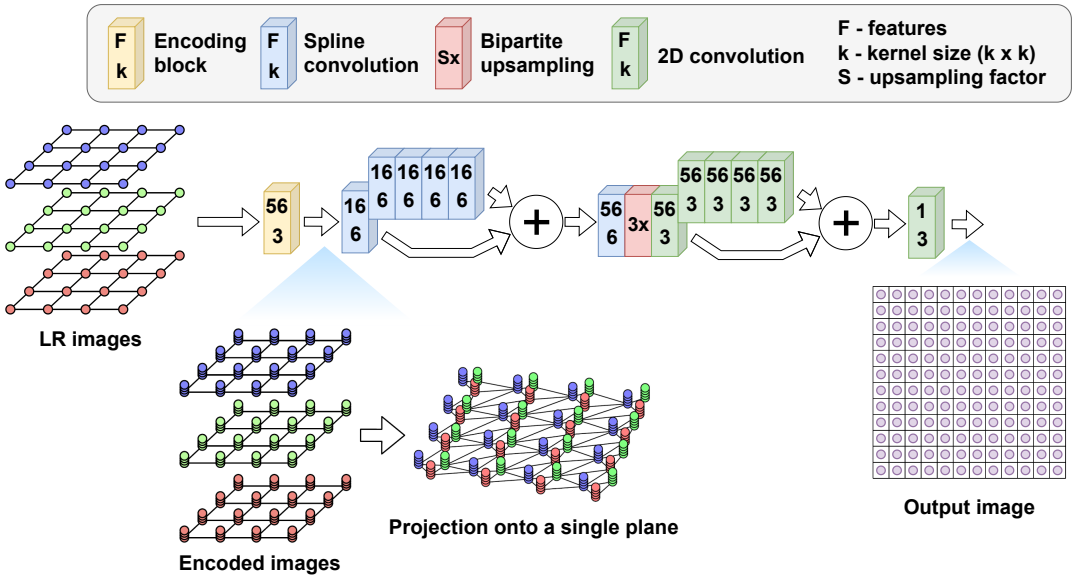


FIGURE 3.7: The architecture of MagNet$_{enc}$.

The successes of MagNet and MagNet++ paved the way for MagNet$_{enc}$, a model designed to refine super-resolution capabilities further, depicted in Figure 3.7. A core challenge in MISR is the diversity in LR images, which can differ in brightness, contrast, sharpness, and noise. This diversity, while offering a unique perspective of a scene, can sometimes cause neural networks to unintentionally prioritize some images over others based on their intrinsic characteristics.

To address this, $\text{MagNet}_{\text{enc}}$ incorporates an encoding block inspired by the HighRes-Net [22] and TR-MISR [2] models. As shown in equations 2.3 and 2.4, this block is anchored by a reference image derived from the median pixel value across multiple LR images, providing a consistent baseline. Individual LR images are then processed in conjunction with this reference. The encoding block's design, featuring an initial convolutional layer and two subsequent residual blocks with PReLU activations, effectively captures spatial patterns and hierarchies. Jointly processing the LR image with the reference image emphasizes differences across frames and aids in identifying high-frequency features. Subsequently, with each LR image independently processed, the resulting feature maps are merged into a unified graph. This combined graph is then passed through the same sequence of processing layers as detailed in the MagNet++ model. Through this approach, the core innovations of the MagNet++ architecture are retained, particularly the bipartite graph upsampling.

In conclusion, $\text{MagNet}_{\text{enc}}$ signifies a notable advancement from its predecessors, emphasizing a more detailed feature extraction phase. This model was crafted as a step towards validating the thesis concerning the integration of techniques utilized by current state-of-the-art MISR architectures to enhance the super-resolution performance of GNNs. By processing diverse LR images independently with a reference image, $\text{MagNet}_{\text{enc}}$ seeks to highlight differences across images, assisting in the recognition of high-frequency features. It aims to leverage the unique insights each LR image provides, towards generating a high-resolution output. This approach keeps the core innovations from MagNet++ intact, particularly the bipartite graph upsampling, while introducing an encoding block to better grasp spatial patterns and hierarchies.

3.5 Learnable Relationships in MagNAt

The motivation behind the development of the MagNAt model encapsulates two pivotal principles. Initially, the driving idea was that, while the relationships and dependencies between a pixel and its neighbours are crucial, not all pixels contribute equally to understanding the underlying scene structure; some relationships are more informative than others. The attention mechanism was seen as a tool to identify and weigh these different relationships based on their significance. By examining the relative importance of each

neighbouring node within a graph representation of an image, the model assigns weights that guide the subsequent processing steps. This ensures that the more relevant features are highlighted while the less relevant ones are toned down, enabling a more accurate and informed super-resolution process.

Moreover, traditional registration algorithms, while proficient in the realm of image super-resolution, exhibit potential shortcomings in scenarios with pronounced discrepancies between images [130]. This observation unveils a chance for inaccuracies that could influence the entire learning journey, possibly propelling the model to work with less-than-ideal data. Recognizing these sporadic challenges sparked an exploration into more adaptive alternatives. Within this narrative, the adaptive and trainable registration method is employed in the MagNAt model as a solution aiming to better address such specific scenarios. The dynamic node placement mechanism in the model adjusts translations of LR images relative to a reference image with each forward pass through the network, standing in contrast to conventional models where translations are usually set once during the data-loading phase and then remain static throughout the training.

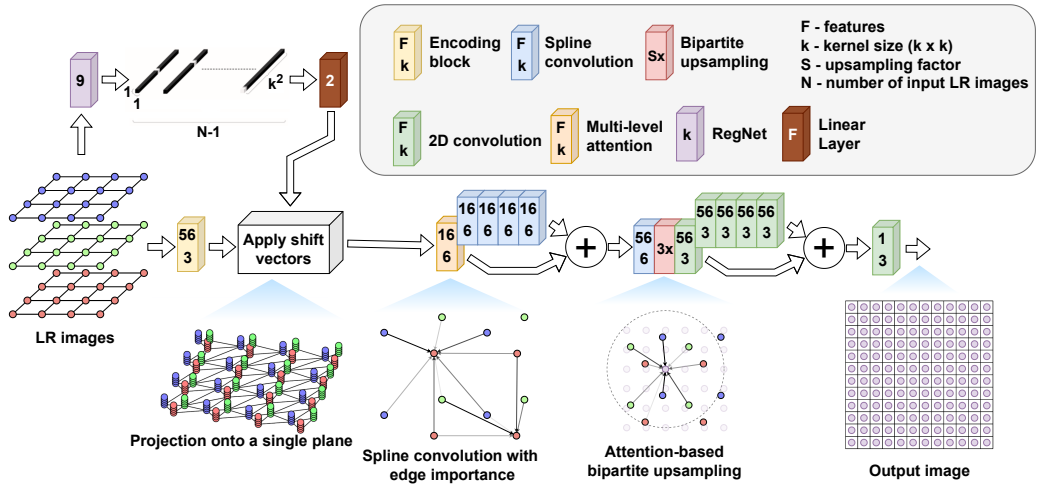


FIGURE 3.8: The architecture of MagNAt.

The following sections go into the details of the MagNAt's architecture, depicted in Figure 3.8, showing how the attention mechanism and adaptive registration work together to improve the super-resolution process, and providing a guide on how the model manages the convolutions, attention distribution, and dynamic node adjustments throughout its layers to achieve its goals. This approach was a necessary step towards substantiating the second thesis, stating that GNNs can markedly improve their MISR performance by assimilating techniques from existing state-of-the-art MISR models.

3.5.1 Attention-Based Convolution

The MagNAt model introduces a multi-level attention mechanism, specifically targeting both node features and edge attributes, into the field of graph-based image super-resolution. This mechanism, layered over the latent representation of an image mapped onto a graph structure, operates in tandem with the bipartite upscaling operator. Combining the principles of GATs, as detailed in Section 2.6.4, with spline-based convolution operations, this model is adept at assigning variable importance to neighboring nodes, thereby enhancing its expressiveness. For a given node v_i, the attention weight α_{ij} for each of its neighbours $j \in \mathcal{N}(i)$ is determined. Unlike the original attention coefficient methodology in GATs described by equation 2.9, the MagNAt model employs a modified approach that incorporates edge attributes, which can be defined as:

$$\alpha_{ij} = \frac{\exp\left(\text{LeakyReLU}\left(\boldsymbol{a}^T[\mathbf{W}\boldsymbol{h_i}||\mathbf{W}\boldsymbol{h_j}||\boldsymbol{u_{ij}}]\right)\right)}{\sum_{k\in\mathcal{N}(i)}\exp\left(\text{LeakyReLU}\left(\boldsymbol{a}^T[\mathbf{W}\boldsymbol{h_i}||\mathbf{W}\boldsymbol{h_k}||\boldsymbol{u_{ik}}]\right)\right)}. \tag{3.5}$$

In this equation, the inclusion of the two-element edge attribute vector $\boldsymbol{u_{ij}}$ necessitates an adjustment in the dimensionality of the learnable weight vector, so that $\boldsymbol{a}^T \in \mathbb{R}^{2F'+2}$.

Incorporating these attention weights into the convolutional process, the MagNAt model deviates from the traditional spline-based convolution, as given by the equation 2.10. The resulting feature vector for node v_i from this modified layer can be calculated as:

$$\boldsymbol{h_i'} = \Big\Vert_{s=1}^{F'} \sum_{j\in\mathcal{N}(i)} \alpha_{ij}\boldsymbol{h_j}^T g_{\theta_s}(\boldsymbol{u_{ij}}). \tag{3.6}$$

The primary adaptation in this layer revolves around the multiplication of the message from each neighbouring node j by its specific attention weight α_{ij}. Intrinsically, because the attention coefficients are obtained through a softmax function, their cumulative value for all neighbours of the node v_i converges to 1. Consequently, there is no need to divide the message by the count of neighbours of the node v_i. Instead, this mechanism effectively considers the incoming messages as a weighted sum.

3.5.2 Dynamic Registration

The MagNAt model employs a dynamic node placement technique, building upon the RegNet model from the DeepSUM framework [96]. From a set of

N input images, one serves as the *reference* image. The remaining images undergo processing through 3D convolutional layers followed by a global pooling stage, resulting in a single vector for each image. Conventionally, these vectors are reshaped into convolutional kernels that, when applied to the corresponding image, align it to the reference. However, in MagNAt, they are channelled through a linear layer to form 2-element shift vectors, which represent horizontal and vertical shifts, for each of the N-1 non-reference images.

At first, all images' nodes are placed based on the node positioning mechanism highlighted in Section 3.1.1. Using the shift vectors computed by RegNet, nodes in each image are adjusted, enabling dynamic alignment with the reference image. The entirety of this mechanism—including node adjustments, shift vector calculations, and the broader architecture—is depicted in Figure 3.8. While the overarching philosophy for graph creation remains unchanged, the innovation lies in the revamped algorithm that calculates these shifts, now a trained submodule. This dynamic alignment unfolds in real-time, underscoring MagNAt's adaptability and robustness.

Theoretical Advantages of Dynamic Registration

The dynamic node placement in the MagNAt model suggests several potential advantages based on theoretical reasoning, though these require further empirical validation:

1. **Robustness to Initial Errors**: Dynamic node placement might help the model continuously adapt to any initial registration errors, potentially providing resistance against inaccuracies seen in traditional static registration methods.

2. **Adaptive Learning**: With dynamic node placement, the model may have the ability to adjust to changes throughout training epochs, which could lead to better performance and improved super-resolution results.

3. **Improved Feature Extraction**: By better aligning LR images through dynamic node placement, the model might achieve more accurate feature extraction, potentially improving the quality of the super-resolution output.

4. **Increased Model Flexibility**: The adaptive nature of node placement could provide an additional level of flexibility, allowing it to optimize learning for the given task.

5. **Mitigating Overfitting**: By constantly adjusting the registration during training, the model may be able to reduce the risk of overfitting to incorrect translations, potentially leading to better performance on unseen data.

The MagNAt model significantly leverages a graph attention mechanism, underscoring the crucial relationships between pixels in image super-resolution tasks. This attention-centric approach is pivotal, especially in the initial residual block and the upsampling module, setting the MagNAt model apart from the $\text{MagNet}_{\text{enc}}$ design, and aiming for superior super-resolution outcomes. Furthermore, the adaptive registration component augments the model's robustness and performance by adaptively handling spatial discrepancies between input images. In the broader scope of this research, the MagNAt model, owing to its innovative features, has been chosen for extensive testing and evaluation, standing as a potential milestone in the MISR domain. Earlier versions of the model, although fundamental, are examined in their comparative analysis section 6.4, showing the step-by-step improvements and emphasizing the promising potential of the MagNAt model.

3.6 Modifications of MagNAt

In this section, two derivative models, $\text{MagNAt}_{\text{no\_reg}}$ and $\text{MagNAt}_{\text{lead}}$, are introduced, both fundamentally grounded on the MagNAt architecture. The creation of $\text{MagNAt}_{\text{no\_reg}}$ aims at evaluating the impact of adaptive registration on the model's performance, contributing to the validation of the second thesis postulated in this dissertation. The distinguishing feature of $\text{MagNAt}_{\text{no\_reg}}$ from MagNAt is the absence of the RegNet component, thus rendering the registration in this model analogous to other models utilizing pre-computed shift vectors. The architectural design of $\text{MagNAt}_{\text{no\_reg}}$ is illustrated in Figure 3.9.

3.6.1 Ensuring Temporal Consistency

$\text{MagNAt}_{\text{lead}}$ is devised to substantiate the third thesis, positing the feasibility of designating one of the LR images as the *leading* image to steer the model

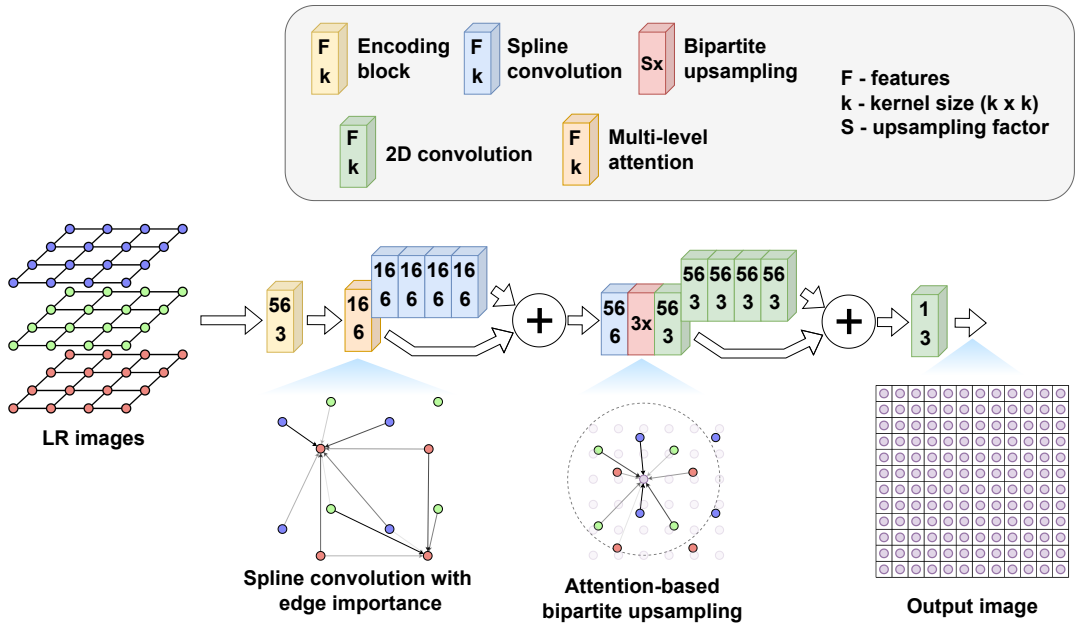


FIGURE 3.9: The architecture of $\text{MagNAt}_{\text{no\_reg}}$

in reconstructing the scene at a specific point in time, dictated by this leading image, and adeptly reconstructing regions of high-temporal variance. The structure of $\text{MagNAt}_{\text{lead}}$ mirrors that of MagNAt, although with minor deviations in the methodologies of registration, encoding, and upsampling. Regarding registration, the difference lies in merely selecting the leading image as the reference during the registration phase, thereby aligning all other images in relation to it. In the context of encoding, opposed to utilizing the median of all LR images as a reference, as denoted by equation 2.3, the leading image is solely used as such reference. Consequently, the modified general expression for the encoding block (equation 2.4) is formulated as:

$$s_0^i = \text{emb}_\theta([\text{LR}_i, \text{LR}_{\text{lead}}]) \tag{3.7}$$

where LR_{lead} represents the leading LR image.

Lastly, within the bipartite upsampling module, the modification relates to the positioning procedure of the new high-resolution graph laid onto the original one. Contrary to centring it with respect to all nodes of the initial graph, as denoted in Section 3.3, it is now aligned solely with the set of nodes corresponding to the leading LR image. In effect, there are $W \cdot H$ nodes of this high-resolution graph positioned identically to the nodes corresponding to the leading image. Theoretically, such an arrangement creates a pronounced bias towards the reconstruction of features corresponding to the leading image.

Although the $\text{MagNAt}_{\text{lead}}$ model suggests a potential improvement over MagNAt by aiming to address the particular problem of temporal consistency, it has not been adopted as a primary model within this dissertation. This decision stems from the fact that the current prevalent MISR datasets do not officially consider or provide a framework for evaluating temporal consistency. Nonetheless, the potential enhancements and performance characteristics of $\text{MagNAt}_{\text{lead}}$ are deliberated in Section 6.3.1, showcasing its promising capability to manage temporal variance in reconstructing scenes.

3.7 Comparison of Proposed Architectures

A progression is observed in the proposed models through iterative refinements, enhancements, and the incorporation of new concepts. For a consolidated overview, a tabulated comparison is presented below, showcasing the distinctive features of each architecture.

Model	**Bipartite upsampling**	**LR uncoding**	**Multi-level attention**	**Adaptive registration**	**Leading LR**
MagNet	✗	✗	✗	✗	✗
MagNet++	✓	✗	✗	✗	✗
$\text{MagNet}_{\text{enc}}$	✓	✓	✗	✗	✗
$\text{MagNAt}_{\text{no\_reg}}$	✓	✓	✓	✗	✗
MagNAt	✓	✓	✓	✓	✗
$\text{MagNAt}_{\text{lead}}$	✓	✓	✓	✓	✓

TABLE 3.1: A comparison of distinctive features across the proposed architectures.

Chapter 4

Data Description and Simulation

High-quality data is fundamental to the success of any machine learning or deep learning project. When it comes to super-resolution tasks, the choice of data and its preparation become even more critical due to the challenges inherent in this domain. This chapter focuses on the datasets used in this research, specifically detailing the process of data simulation, which played a key role in training and evaluating the developed models.

4.1 Simulated Dataset

This research strategically incorporates both simulated and real-world datasets. The simulated datasets afford advantages in generating a plethora of training examples from a limited set of HR images. They provide a controlled environment where specific features of the LR images, such as degradation levels, can be manipulated. Furthermore, they permit emulation of certain real-world conditions, particularly minor shifts in the image, crucial for MISR.

Two distinct simulated datasets, SRRB and SRRB_{enh}, were curated for this work. Their distinctions arise from their simulation parameters and intended purposes. The SRRB dataset solely simulates sub-pixel shifts between LR images, designed to assess the capability of models in capturing and addressing these minute translations. In contrast, the SRRB_{enh} dataset simulates not just translations, but also global variations like brightness, contrast, and noise. This was intended to approximate real-world scenarios where LR images of the same scene exhibit distinct global attributes. While SRRB_{enh} moves closer to mimicking real-world datasets, it still diverges due to inherent temporal consistencies. Since simulated LR images stem from a single HR source, they inherently share identical temporal information, unlike real-world observations, which inevitably possess such temporal variations.

In the context of this research, a pivotal decision concerned the downsampling factor for simulating LR images. Opting for a $3\times$ downsampling factor, the resultant LR images are three times smaller compared to their HR versions. This configuration mirrors the Proba-V super-resolution challenge and its corresponding dataset employed in this study (Section 4.2). Notably, numerous state-of-the-art models, which were employed and evaluated in this work, were originally tailored for this challenge and its specific upsampling factor. To preserve the original training conditions of the state-of-the-art models, the upsampling factor, $\mathcal{S}$, was consistently set at three for retraining and evaluation in this study.

Both simulated datasets source images from established databases. Specifically, the DIV2K [1] database was employed for training the models, while others, namely BSDS100 [3], historical [73], Manga109 [92], Set5 [9], Set14 [162], and Urban100 [73], were utilized for benchmarking purposes. The selected databases are renowned in super-resolution research, ensuring the incorporation of diverse image types and thereby enabling a comprehensive evaluation of the proposed models.

4.1.1 Training and Validation Datasets

The DIV2K dataset was chosen as the primary resource for the training and validation phases of the models. DIV2K is widely recognized in the super-resolution domain for its diverse and extensive collection of high-quality images. The original dataset comprises 1000 RGB images that display a range of scenes, objects, and textures.

For the purpose of this study, all images from the DIV2K dataset, as well as those from all other included datasets, were converted to grayscale. This conversion was undertaken to simplify the research process, concentrating on the core problem of image super-resolution without the added complexity of colour information.

Additionally, a preprocessing step was introduced to ensure uniformity in image sizes, a vital aspect for facilitating batch creation and collation during the training phase. The smallest image size in the dataset was identified, and all larger images were divided into non-overlapping sub-images of this determined dimension. In this research, the chosen size was 222×222 pixels. As a result of this cropping strategy, the effective number of images in the dataset was expanded from the original 1000 to 1409.

An 80:20 split was then applied to the preprocessed DIV2K dataset, with 80% of the images reserved for training and the remaining 20% designated for validation. This commonly adopted ratio ensures a substantial majority of data aids in discerning core features and patterns, while a significant portion is preserved to evaluate the model's capabilities and adaptability on previously unseen data during the validation stage, thus mitigating overfitting risks. This allocation remained consistent in both the SRRB and SRRB_{enh} datasets.

4.1.2 Benchmark Datasets

In addition to DIV2K, several other datasets were used for benchmark purposes. These datasets - BSDS100, historical, Manga109, Set5, Set14, and Urban100 - are well-known in the super-resolution community and widely employed for model evaluation, especially in SISR [129, 26, 66].

Each benchmark dataset offers unique image characteristics, providing a diverse testing ground. Notably, the numbers in the names of datasets like BSDS100, Manga109, Set5, Set14, and Urban100 indicate the number of images they contain. BSDS100 comprises natural images with complex textures and structures. The historical dataset, distinctively, consists of 10 images that showcase older, classical styles and textures. Manga109 is characterized by manga-style images demanding the preservation of sharp edges. Set5 and Set14, although smaller in size, offer a diverse and broadly used collection of images for benchmarking in the super-resolution community and image processing in general. Urban100 features images of urban scenes, predominantly with man-made structures and patterns. Using these varied datasets ensures that the performance of the models is scrutinized across different image types and conditions, thereby measuring their robustness and adaptability - factors vital for effective deployment in real-world SRR tasks.

4.1.3 The Process of Generating Simulated Datasets

This section details the methodologies behind the creation of simulated LR images for the SRRB and SRRB_{enh} datasets. While the SRRB is centred on image shifting and downsampling, the SRRB_{enh} goes a step further, introducing real-world challenges such as contrast adjustments, blurring, and noise. Each dataset plays a distinct role in evaluating MISR models, differing in their complexity levels. The subsequent subsections provide a thorough overview of the image generation processes for each dataset.

Simulated Dataset

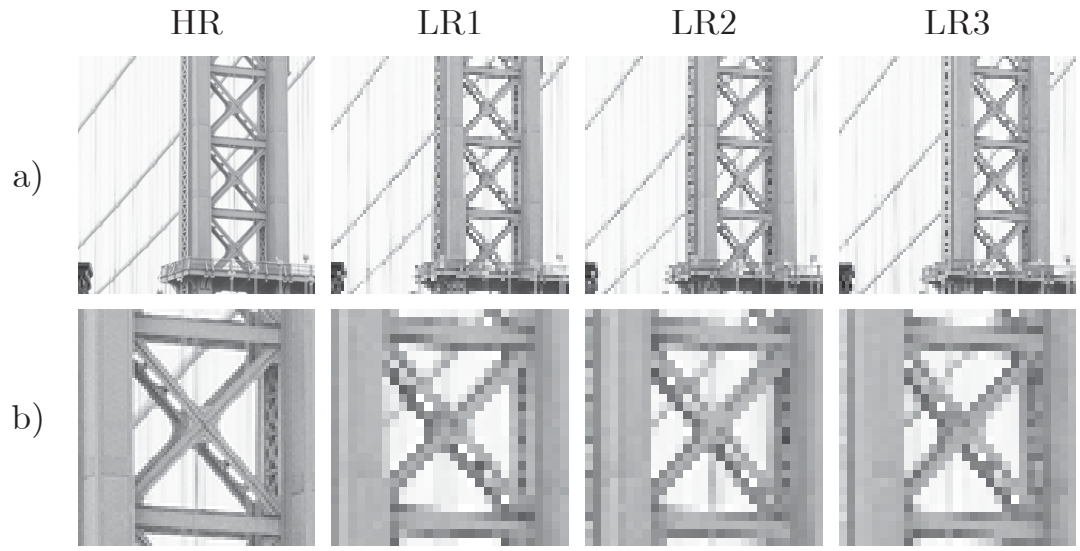


FIGURE 4.1: Illustration of the images synthesized for the SRRB dataset. The first row (a) presents the full-sized images, while the second row (b) offers magnified segments, emphasizing the sub-pixel shift variations between them.

The process of generating simulated LR images for the SRRB dataset involved two crucial steps: *image shifting* and *downsampling*. In the initial stage, every HR image was subjected to a series of shifts, resulting in nine distinct shifted versions for each. These shifts aimed to emulate the variability frequently observed in real-world situations where factors such as changes in perspective or environmental variables might modify a scene's representation. Shift vectors, applied to each HR image independently, were derived from a uniform distribution within the range of $[-1.2, 1.2]$ for both dimensions individually. This range was selected based on calculations and observations from the real-world Proba-V dataset. Specifically, sub-pixel shifts between LR images rarely surpass the interval $[-0.4, 0.4]$. Given the chosen downsampling rate of three, the range for HR images effectively matches this observed interval for the LR images. Following the shifting operations, the HR images were downsampled $3\times$ to produce the respective LR counterparts.

From this method, nine LR images were produced for each original HR image, intending to introduce foundational intricacies associated with MISR tasks. The entire procedure of generating multiple LR images from a singular HR example is detailed in Algorithm 1, and the resulting images are depicted in Figure 4.1.

Algorithm 1 Algorithm used to simulate multiple LR images.

Initialize empty set LRs
Set number of shifts $n \leftarrow 9$
Set shift range $r \leftarrow [-1.2, 1.2]$
Set downscale factor $d \leftarrow 3$
for $i \leftarrow 1$ to n **do**
    $x\_shift \leftarrow$ random number within r
    $y\_shift \leftarrow$ random number within r
    $shift\_vector \leftarrow (x\_shift, y\_shift)$
    $shifted\_image \leftarrow$ shift HR by $shift\_vector$
    $LR \leftarrow$ downscale $shifted\_image$ by factor d
    Add LR to the set LRs
end for
return LRs

Enhanced Simulated Dataset

The procedure for generating LR images for the SRRB_{enh} dataset is a slight enhancement of the one used for the SRRB dataset. This includes the steps of shifting HR images, adjusting contrast and brightness, downsampling, blurring, and finally, adding noise.

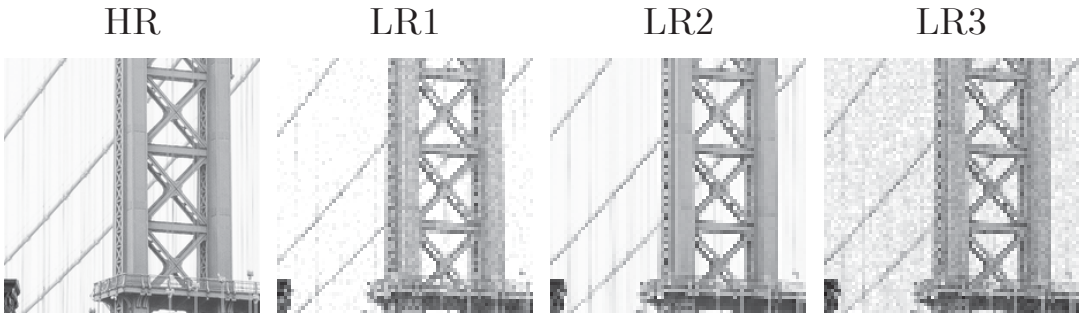


FIGURE 4.2: Illustration of the LR images obtained for the SRRB_{enh} dataset. The images underwent a series of transformations, including shifting, contrast and brightness adjustments, blurring, and noise addition.

Each image alteration parameter is independently sampled from the normal distribution $\mathcal{N}(\mu, \sigma^2)$ for every generated LR image. Using a distribution rather than a fixed value enables variations across different images, augmenting the dataset's diversity. In the normal distribution, two parameters are defined: the *mean* (μ), denoting the central value, and the *standard deviation* (σ), expressing the dispersion of values.

The contrast adjustment is performed using values from $\mathcal{N}(1.0, 0.05^2)$, while brightness adjustment employs values from $\mathcal{N}(0.0, 5.0^2)$. After these alterations, the image is downsampled and then blurred using a Gaussian

kernel with a standard deviation of 0.2. The noise is finally added using values from $\mathcal{N}(0.0, 10.0^2)$. The specific values for each parameter were carefully chosen based on the author's insights gathered in previous research [124, 126]. This series of steps increases the complexity of the SRRB_{enh} dataset, providing a more rigorous testing environment for the super-resolution models. Figure 4.2 showcases the LR images derived through the procedure outlined in Algorithm 2.

Algorithm 2 Algorithm used to simulate multiple LR images for the SRRB_{enh} dataset.

Initialize empty set LRs
Set number of shifts $n \leftarrow 9$
Set shift range $r \leftarrow [-1.2, 1.2]$
Set downscale factor $d \leftarrow 3$
Set contrast parameters $\mu_c \leftarrow 1.0$, $\sigma_c \leftarrow 0.05$
Set brightness parameters $\mu_b \leftarrow 0.0$, $\sigma_b \leftarrow 5.0$
Set noise parameters $\mu_n \leftarrow 0.0$, $\sigma_n \leftarrow 10.0$
for $i \leftarrow 1$ to n **do**
    $x\_shift \leftarrow$ random number within r
    $y\_shift \leftarrow$ random number within r
    $shift\_vector \leftarrow (x\_shift, y\_shift)$
    $shifted\_image \leftarrow$ shift HR by $shift\_vector$
    $c \leftarrow$ sample from $\mathcal{N}(\mu_c, \sigma_c^2)$
    $b \leftarrow$ sample from $\mathcal{N}(\mu_b, \sigma_b^2)$
    $enhanced\_image \leftarrow shifted\_image * c + b$
    $LR \leftarrow$ downscale $enhanced\_image$ by factor d
    $LR \leftarrow$ blur LR with Gaussian kernel ($\sigma = 0.2$)
    $noise \leftarrow$ matrix sampled from $\mathcal{N}(\mu_n, \sigma_n^2)$ with the same size as the LR
    $LR \leftarrow LR + noise$
    Add LR to the set LRs
end for
return LRs

4.1.4 Reflection on Data Simulation

Both the SRRB and SRRB_{enh} datasets were crafted to emulate the intricacies of real-world imaging, albeit with distinct intentions. The SRRB dataset, derived through random shifts and downsampling of HR images, offers a foundational training ground for super-resolution models. These shifts embody the natural variability found in multiple captures of real scenes, ensuring model exposure to diverse training conditions.

Enhancing upon the SRRB's premise, the SRRB_{enh} dataset integrates further intricacies, including image modifications like contrast and brightness

adjustments, noise addition, and mild blurring. These alterations simulate environmental influences and the various challenges faced in real-world captures, such as blurring from atmospheric distortions or noise due to sensor constraints and lighting conditions [17].

However, an essential nuance in these datasets is that the simulated LR images, despite their differences, represent the same temporal instance. Real-world captures often span different time points, encompassing shifts due to camera movements, environmental variations, or differing exposure times. For remote sensing, factors like satellite position alterations and atmospheric shifts further complicate the scenario. The current SRRB_{enh} dataset does not capture these temporal dynamics, suggesting a potential enhancement avenue for subsequent versions.

In essence, while the SRRB offers foundational training, the SRRB_{enh} propels model challenges by imitating a broader spectrum of imaging challenges. Using them in tandem ensures a comprehensive model evaluation, balancing fundamental training with advanced adaptability.

4.2 Real-World Dataset

Real-world data is pivotal in this research, enabling the testing and validation of super-resolution models under genuine conditions. To achieve this, the Proba-V multi-image super-resolution dataset [91] was chosen.

Several factors guided the selection of the Proba-V dataset. Firstly, it offers real-world images captured by the Proba-V satellite, creating an authentic and challenging environment for developing and testing super-resolution models. This dataset incorporates the inherent complexities typical of satellite imagery, including data gaps and variations in image quality due to atmospheric conditions like cloud coverage. These characteristics make it a true reflection of real-world scenarios.

Moreover, the Proba-V dataset provides an element that the simulated datasets SRRB and SRRB_{enh} lack: temporal variation. As the images in the Proba-V dataset were captured at different points in time, they expose the models to the natural alterations that occur between subsequent captures, including changes in lighting, movement of objects, and potential growth or alteration of the landscape. This temporal variation is an indispensable factor in real-world SRR tasks, making the Proba-V dataset a crucial tool for this research.

Lastly, the Proba-V MISR challenge has been a significant driving force in the field of MISR. Numerous state-of-the-art models have emerged as a result of this challenge, providing innovative techniques and benchmarks for evaluating and comparing the models developed in this research.

4.2.1 Proba-V MISR Dataset Structure

The Proba-V MISR dataset, provided for the challenge, consists of satellite data from 74 regions around the globe, which were hand-selected at different points in time within a 30-day window. A total of 1450 scenes are contained in this dataset, of which 1160 scenes are allocated for training and 290 for testing; however, the latter set is devoid of HR target images.

Each scene consists of exactly one HR (100 m GSD) image of shape 384×384, which is accompanied by multiple LR (300 m GSD) of shape 128×128. Additionally, each image has a corresponding quality map. Their role is to indicate which pixels in the image are obscured (e.g., by clouds, cloud shadows, ice, water) and which ones remain clear.

Inclusion criteria for an image in this dataset dictate that at least 75% of its pixels must be clear for 100 m resolution images, and 60% for 300 m resolution images. The number of LR images per scene can vary from 9 to 32, with an average count of 19. The original test set, encompassing 290 scenes, is not accompanied by HR reference images and is specifically structured for evaluations on the competition's servers. For the sake of enabling local evaluations, the original training subset was divided into three distinct subsets: training, validation, and test, adhering to an 80:10:10 ratio, respectively. Through this division, comprehensive evaluations of the models at different developmental stages were facilitated.

4.2.2 Spectral Bands of Proba-V Dataset

The Proba-V MISR dataset includes two spectral bands: *near-infrared* (NIR) and *red* (RED). The data consists of radiometrically and geometrically corrected *top-of-atmosphere reflectances* for these spectral bands.

The NIR band, with wavelengths between 780 and 2500 nm, is often used in remote sensing and satellite imaging to study vegetation, water bodies, and atmospheric properties. Images in the NIR band can reveal details about plant health, water content, and other features that are not visible in the RGB spectrum. The RED band, with wavelengths approximately between 620 and 750 nm, is particularly sensitive to chlorophyll and can provide information

about vegetation's photosynthetic activity. This makes it useful for monitoring plant growth, assessing crop yields, and tracking changes in forest cover, among other applications. In this study, both NIR and RED subsets, comprising 566 and 594 scenes, respectively, were independently utilized for training, validation, and testing of the models.

4.2.3 Handling Real-World Challenges

The Proba-V MISR dataset embodies various real-world challenges, underscoring the imperative for advanced super-resolution models. The models developed must not only replicate real-world dynamics but also adapt to myriad inconsistencies to ensure their practical relevance.

A significant challenge in the dataset is the presence of areas obscured by clouds, cloud shadows, ice, and water. To assist in managing this, the dataset provides clearance masks for both LR and HR images. During the training phase, these masks play an invaluable role, allowing the exclusion of obscured or irrelevant pixels. During the super-resolution reconstruction phase, these masks ensure a better input by excluding the irrelevant pixels.

Another nuance in the dataset is its data storage method. The image files use a 16-bit depth, but the actual image content occupies only 14 bits. Consequently, areas without captured content reach the maximum of the 16-bit range. This specific representation is not complexity in itself but a feature of the dataset. It's essential for super-resolution models to recognize this aspect, ensuring these pixels are properly managed to prevent distortions.

Geometric intricacies, inherent in satellite imagery, appear as shifts and rotations due to factors like satellite position and Earth's rotation. The Proba-V MISR dataset addresses this by utilizing the Plate Carrée projection [118]. While this projection significantly rectifies geometric distortions using elevation data and the sensor's metadata, minor shifts—primarily at the subpixel level—persist between the LR images. These shifts, however, can be seen as an advantage. Offering varied views of the same scene, they can potentially enrich the super-resolution process. Models should capitalize on these shifts, extracting enhanced information and producing more detailed outputs. To visually comprehend the nuances of the Proba-V MISR dataset, Figure 4.3 showcases representative images from the dataset alongside their respective clearance maps.

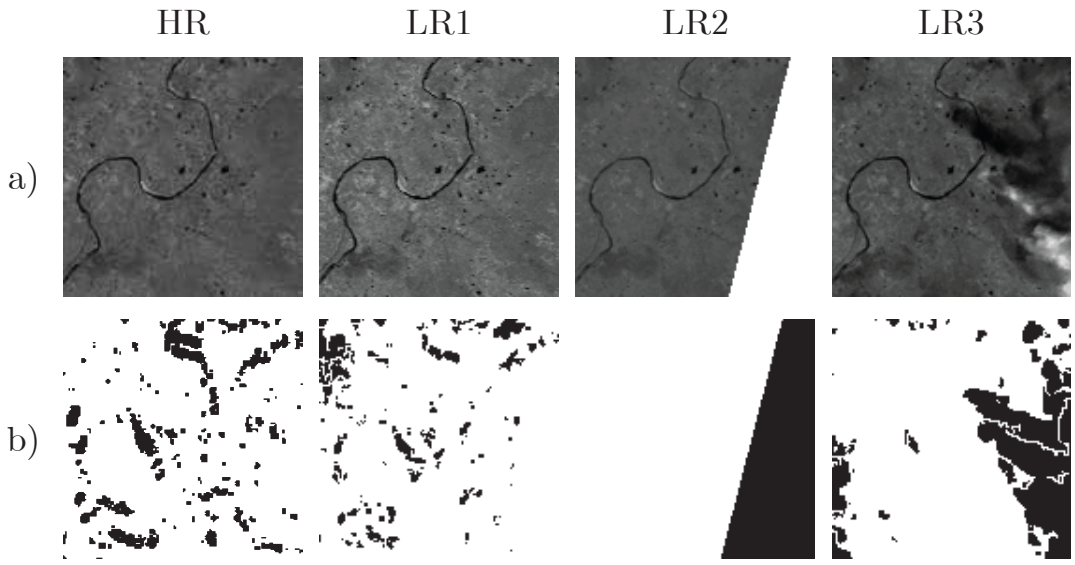


FIGURE 4.3: Illustrative samples from the Proba-V MISR dataset. The top row (a) displays selected images, while the bottom row (b) presents the corresponding clearance maps, highlighting the areas of clarity (white) versus obscurity (black).

4.2.4 Importance of the Proba-V MISR Dataset

The Proba-V MISR dataset's utilization in this research is seen as instrumental for the development and evaluation of super-resolution models under authentic real-world conditions. Characterized by its broad geographical coverage and diverse atmospheric conditions, this dataset can be considered a pivotal resource for super-resolution studies. Further advancements in the field have been significantly catalyzed by the Proba-V MISR challenge, with an array of innovative techniques and cutting-edge models being introduced. By engaging with this dataset, existing knowledge from these developments was built upon, and further contributions to this dynamic field were facilitated.

Although the Proba-V dataset is not the only resource available for MISR, other notable datasets have been introduced, such as *MuS2* [69] and the *semi-simulated Sentinel-2 dataset* [124]. MuS2 was designed for super-resolving multiple Sentinel-2 images, using WorldView-2 as the HR reference. However, it faces challenges due to the distinct sensor characteristics of the Sentinel-2 and WorldView satellites, leading to subtle discrepancies between the low and high-resolution images. Moreover, while the semi-simulated Sentinel-2 dataset approximates real-world conditions better than typical simulated datasets, it does not capture the full authenticity of purely real-world data. A

distinguishing factor of these datasets is their orientation towards multispectral MISR, which presents a different set of challenges and considerations. Given these considerations, the Proba-V dataset was favoured for this research. This choice was also influenced by the state-of-the-art MISR models specifically tailored for Proba-V, which were pivotal for comparison in this study. Additionally, the Proba-V dataset ensures a uniform sensor profile for both its resolution scales. Coupled with its intrinsic real-world challenges and the unbiased evaluation potential via the challenge's servers, the Proba-V dataset stands out as the principal dataset for this dissertation.

4.3 Dataset Comparison

Understanding the differences between datasets is crucial for appreciating the challenges and opportunities they present. To shed light on the distinctions between SRRB, SRRB_{enh}, and Proba-V datasets, a comparative table has been developed below.

Parameter	**SRRB**	**SRRB_{enh}**	**Proba-V (NIR)**	**Proba-V (RED)**
Training examples	1409	1409	510	535
Test examples	338	338	56	59
Images/Scene	9	9	9–32	9–32
LR Size	74	74	128	128
HR Size	222	222	384	384
LR Shifts	✓	✓	✓	✓
Global LR Differences	✗	✓	✓	✓
Local Variations	✗	✗	✓	✓

TABLE 4.1: Comparison of the datasets used in this research.

Given the outlined characteristics of each dataset, the subsequent chapter details how this data was employed for training the super-resolution models. The methodology adopted and the metrics chosen for assessment are also discussed to provide a comprehensive view of the experimental approach.

Chapter 5

Training Methodology and Evaluation Metrics

The methodology adopted for training significantly influences a model's performance. This chapter delves into the detailed training processes utilized for the proposed and various state-of-the-art models. The training was conducted on three distinct datasets: SRRB, SRRB_{enh}, and the real-world Proba-V dataset, each presenting its own unique challenges. Notably, for each architecture, four independent models were trained, one for each simulated dataset and one for each Proba-V band subset. The procedures followed while handling the training for each dataset, the choice of hyperparameters, and the techniques used to overcome various challenges are discussed in detail in the subsequent sections.

5.1 Loss Function

In the training phase of deep learning models, the loss function plays a pivotal role in optimizing the model to produce outputs closely resembling the target images. To train the proposed models, the *corrected PSNR* (cPSNR) is used as a loss function. cPSNR is a specialized version of the *peak signal-to-noise ratio* (PSNR) [146, 108] and was originally introduced for the Proba-V competition to address the unique challenges posed by its corresponding dataset in MISR tasks. A key feature of this metric is the exclusion of concealed pixels, which is particularly useful for real-world satellite imagery to avoid reconstructing obscured areas, such as clouds, that can introduce inaccuracies. Additionally, cPSNR incorporates brightness correction and full-pixel registration, ensuring that image quality assessment remains consistent regardless of variations in brightness and positional discrepancies between the compared images.

5.1.1 cPSNR Computation

The foundation of cPSNR lies in its emphasis on image registration. Given the inherent complexities of the multi-image super-resolution (MISR) scenario, where a series of mutually shifted LR images is provided without a precise HR image reference, it would be unjust to penalize the model for minor registration deviations. To account for this, cPSNR permits full-pixel shifts, aligning the super-resolved and HR images and adjusting for possible shifts during the super-resolution phase.

In the context of cPSNR computation, the HR images are divided into patches using indices i and j. These patches are created by shifting the HR image within a $[-3, 3]$ range in both dimensions, resulting in 49 distinct patches. Concurrently, the super-resolved image is cropped to obtain a single, central patch. It is important to note that each patch has a diminished effective size due to a 3-pixel exclusion on all sides. This exclusion ensures that border effects, resulting from shifting, do not influence the computation. The focus is primarily on comparing patches of these shifted HR images with the corresponding central patch of the super-resolved image.

First, the brightness correction term b is calculated by subtracting the average intensities of the cropped super-resolved image from the HR patches, considering only clear pixels. This is done for each of the 49 HR image patches denoted by $HR_{i,j}$:

$$b = \frac{1}{|\text{clear}(HR_{i,j})|} \left(\sum_{x,y \in \text{clear}(HR_{i,j})} HR_{i,j}(x,y) - SR(x,y) \right). \tag{5.1}$$

Next, the corrected mean squared error (cMSE) between the cropped super-resolved image and each HR patch is computed, with a correction for brightness. The cMSE accounts for the average intensity difference between the super-resolved and HR images and, analogously to the brightness correction term, is calculated only over clear pixels:

$$\text{cMSE}(HR_{i,j}, SR) = \frac{1}{|\text{clear}(HR_{i,j})|} \sum_{x,y \in \text{clear}(HR_{i,j})} (HR_{i,j}(x,y) - (SR(x,y) + b))^2. \tag{5.2}$$

Finally, the cPSNR is computed by finding the maximum value calculated for different HR patches:

$$\text{cPSNR}(HR, SR) = \max_{i,j \in 0,\dots,6} \Big\{ -10 \log_{10}(\text{cMSE}(HR_{i,j}, SR)) \Big\}. \tag{5.3}$$

In the following section, a detailed discussion on the training methodologies employed for the models is provided.

5.2 Training Details of Super-Resolution Models

During the course of this study, for each super-resolution architecture, four models were trained—one for each of the datasets: SRRB, SRRB_{enh}, Proba-V NIR, and Proba-V RED. The specific training configurations adopted for these models are detailed in Table 5.1.

TABLE 5.1: Training parameters for super-resolution models across different datasets.

Dataset	Model	Learning Rate	LR Patch Size	LR Images	Batch Size	Loss Function	Preprocessing
SRRB **SRRB_{enh}**	HighRes-Net	0.0007	64×64	9	32	cPSNR	Rescaling
	RAMS	0.0005	32×32	9	32	cL1	Standardization
	PIUNET	0.0001	32×32	9	24	cL1 with uncertainty	Standardization
	TR-MISR	0.0005	64×64	9	24	cPSNR	Rescaling
	MagNAt	0.001	32×32	9	32	cPSNR	Rescaling
Proba-V	HighRes-Net	0.0007	64×64	9-32	32	cPSNR	Rescaling
	RAMS	0.0005	32×32	9	32	cL1	Standardization
	PIUNET	0.0001	32×32	9	24	cL1 with uncertainty	Standardization
	TR-MISR	0.0005	64×64	9-24	4	cPSNR	Rescaling
	MagNAt	0.0005	32×32	9-15	16	cPSNR	Rescaling

Hyperparameters for the state-of-the-art models—HighRes-Net, RAMS, PIUNET, and TR-MISR—were adopted without alteration based on their corresponding papers and repositories. This adherence to original specifications encompasses preprocessing as well: while images for certain models underwent *rescaling*, adjusting image values to lie between 0 and 1, others utilized *standardization*. The latter was conducted using the overall mean and standard deviation of all training set images. Furthermore, the Adam optimizer was consistently employed for all models.

In terms of loss function optimization, most models were trained to improve the cPSNR value. However, RAMS and PIUNET took a different path, aiming to minimize the *cL1* loss function. This cL1 is an adapted version of the *L1* loss, adjusted with brightness and shift corrections analogous to the transition from PSNR to cPSNR. Additionally, PIUNET introduces the *uncertainty loss* component during training. Further insights into this specific aspect of PIUNET can be found in [134].

A notable divergence in the adopted approach was observed in terms of training duration. Instead of following epoch counts as outlined in the corresponding literature for each model, a consistent training period of 1000 epochs was established for all models. To counter potential biases, model weights that demonstrated optimal performance on a validation subset were chosen, regardless of when this peak performance occurred during the training process. This approach was devised to address potential overfitting challenges that could arise from the extended training duration.

Compared to Proba-V, the simulated datasets' uniform structure of always having nine LR images per scene allowed for the batch sizes to be increased for certain models, specifically TR-MISR and MagNAt. By maximizing the batch size within hardware memory limits, more efficient gradient approximations can be achieved, potentially accelerating convergence and improving generalization.

When it came to validation, a shift was made from the patch-based evaluation to using entire images from the validation subset. This strategic deviation ensured a holistic assessment of the models, gauging their capability to super-resolve entire images, which is more aligned with practical applications. It should be highlighted that the validation metric employed consistently for all models was cPSNR.

5.3 Evaluation

The post-training assessment spanned both simulated and Proba-V datasets, focusing on the adaptability of models to diverse data conditions.

For the simulated datasets, evaluations were based on image dimensions up to 160×160 for LR inputs and 480×480 for HR images. Images larger than these sizes were either cropped or split into non-overlapping patches. This method aimed to gauge the models' adaptability beyond training constraints and their scalability for diverse image sizes. The chosen maximum patch size was primarily dictated by the MagNAt's memory limitations, as further discussed in Section 6.5. In the context of Proba-V, evaluations engaged with original-sized images: 128×128 for LR images and 384×384 for HR references, offering a hands-on evaluation of the models' super-resolution abilities on entire images.

The prudent selection of evaluation metrics, particularly for satellite imagery, is pivotal to ensure that super-resolution techniques align with real-world applications [8]. In this research, the metrics employed include the

cPSNR, previously discussed in Section 5.1.1, and four additional metrics: *structural similarity index measure* (SSIM) [145], *learned perceptual image patch similarity* (LPIPS) [165], *mean gradient error* (MGE) [85], and *the blur effect* (TBE) [20]. Except for TBE, which is a non-reference metric calculated solely from the super-resolved image, other metrics use their 'corrected' version, denoted by the 'c' prefix. This correction, mirroring the cPSNR's adjustment process, encompasses brightness rectification and full-pixel registration, ensuring consistent, precise model performance comparisons.

5.3.1 Structural Similarity Index Measure

The SSIM is a perceptual metric used to assess super-resolved image quality against HR references. While traditional metrics like PSNR emphasize signal fidelity, SSIM focuses on preserving structural information, reflecting a more human-centric perception.

SSIM operates in a unique manner, moving a window of size $K \times K$ pixel by pixel over the entire image and evaluating the similarity between local regions in the super-resolved image and the corresponding regions in the HR image. It is a location-dependent analysis providing detailed insights into where the super-resolution model excels and where it falls short.

The SSIM index is calculated for each pair of windows x and y, both of size $K \times K$, taken from the super-resolved image and the HR reference image, respectively:

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{x,y} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}. \tag{5.4}$$

Here μ_x and μ_y are the averages of windows x and y, respectively; σ_x^2 and σ_y^2 are the variances of x and y; $\sigma_{x,y}$ is the covariance of x and y; $c_1 = (k_1 L)^2$ and $c_2 = (k_2 L)^2$ are two variables to stabilize the division with a weak denominator; L is the dynamic range of the pixel-values; $k_1 = 0.01$ and $k_2 = 0.03$ by default. SSIM values range from 0 to 1, with 1 denoting perfect similarity.

In the context of image super-resolution, SSIM stands out for its perceptual focus, emphasizing structural and contrast changes over pure signal fidelity. This aligns it more with human vision, which prioritizes structural detail. Its sliding window approach also highlights specific areas of dissimilarity for a nuanced evaluation.

However, SSIM is not without limitations. The sliding window approach introduces a higher computational cost. Additionally, while it offers a perceptual quality score, it does not necessarily align perfectly with human judgment in all cases, given the vast complexity of vision and the many factors that influence human perception of image quality [144].

5.3.2 Learned Perceptual Image Patch Similarity

LPIPS is a state-of-the-art metric designed for perceptual similarity assessment, differentiating it from traditional metrics such as PSNR and SSIM that target low-level image features. It uses deep learning to concentrate on high-level structures and details. It employs a neural network trained on human-rated images to estimate perceptual similarity between a pair of images. Notably, the LPIPS score ranges between 0 and 1, with 0 indicating perfect perceptual similarity and 1 signifying maximal perceptual dissimilarity. Capturing both pixel-level accuracy and nuanced visual characteristics, LPIPS has been recognized in recent research [69] as effective for assessing the quality of super-resolved images, especially when combined with PSNR and SSIM. However, its dependence on the training data and increased computational demand are acknowledged limitations.

Comparing SSIM with LPIPS highlights distinct approaches. SSIM measures structural similarity by considering aspects like contrast, luminance, and structure. In contrast, LPIPS assesses perceptual differences based on human judgment. While LPIPS offers a deep perceptual analysis, SSIM's deterministic approach delivers results without the potential biases from training data and is less computationally intensive. Both metrics serve their purpose: SSIM for structural similarity and LPIPS for deeper perceptual differences. Using both provides a thorough evaluation of super-resolved images in terms of both structure and visual perception.

5.3.3 Mean Gradient Error

The MGE is a metric uniquely tailored to evaluate edge sharpness in images—a critical component of perceptual quality [87]. While metrics like PSNR, SSIM, and LPIPS provide a broader perspective on image quality, MGE directly contrasts the gradient magnitudes between super-resolved and HR reference images, offering a specialized assessment of detail sharpness.

Edge sharpness delineates objects and communicates essential boundary and texture details. Given its capability to detect changes in edge sharpness,

MGE's importance in super-resolution is underscored. Traditionally utilized as a loss function [85, 84], through the evaluation of gradient magnitudes, MGE measures the preservation of edge sharpness in super-resolved outputs. Hence, a lower MGE score is preferable, indicating reduced sharpness differences between the images.

The MGE score between an HR image and a super-resolved image is defined as follows:

$$MGE(HR, SR) = \frac{1}{HW} \sum_{i=1}^{H} \sum_{j=1}^{W} (G(i,j) - \widehat{G}(i,j))^2. \tag{5.5}$$

In this equation, $G(i,j)$ and $\widehat{G}(i,j)$ represent the gradient magnitude at pixel location (i,j) in the HR and super-resolved images, respectively, with H and W indicating the total number of pixels in the vertical and horizontal dimensions. The gradient magnitude at a given pixel location is calculated using the formula:

$$G(i,j) = \sqrt{G_x^2(i,j) + G_y^2(i,j)} \tag{5.6}$$

Here, G_x and G_y represent the horizontal and vertical gradients, respectively. These gradients are obtained by convolving the image Y with the Sobel operators[57], which are well-established edge detection filters:

$$G_x = Y * \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix}, \tag{5.7}$$

$$G_y = Y * \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} \tag{5.8}$$

with $*$ denoting convolution operation. This gradient information serves as the basis for calculating the MGE, thereby allowing for a direct assessment of edge sharpness in the images.

5.3.4 The Blur Effect

TBE metric is designed to offer a no-reference evaluation of image sharpness by quantifying perceived blur. Unlike traditional metrics that require a reference image, TBE operates independently, eliminating the need for correction

procedures such as full-pixel registration and brightness adjustments. Its relevance and applicability have been demonstrated in other super-resolution studies, including those by Maier et al. [89] and Karmakar et al. [59].

TBE's foundation is based on the concept that blurring significantly affects perceived image quality, originating from various factors like camera focus errors, motion blur, atmospheric effects, or compression artefacts. Designed to capture perceptual aspects of the blur, TBE evaluates local contrasts, especially in edge regions, which are most perceptually significant. The following steps are involved in the computation of TBE:

1. **Edge Detection**: Edges within the image are identified using specific techniques. Such edges, marking boundaries between distinct regions, are vital for perceiving sharpness.

2. **Local Contrast Assessment**: The local contrast around detected edges is evaluated. Sharp edges show pronounced intensity changes over short distances, whereas blurred edges present gradual transitions.

3. **Spatial Distribution Analysis**: The distribution of local contrasts across the image is analyzed. This analysis provides a holistic view of the image's sharpness.

4. **Metric Aggregation**: Local contrast measurements and their spatial distribution insights are combined into a single metric value, indicating the overall perceived blur in the image.

MGE and TBE have been incorporated to offer a well-rounded sharpness assessment. While MGE centres on edge structures through gradient magnitude comparisons, TBE focuses on the broader aspect of a perceived blur. MGE is sensitive to registration quality, introducing potential variability, whereas TBE's non-reliance on references minimizes such concerns. Given TBE's alignment with human perceptions of blur and validation against human judgments, using both metrics ensures a comprehensive and robust evaluation for super-resolution model assessment.

Chapter 6

Experimental Results and Discussion

In this chapter, I present a comprehensive evaluation of the models proposed in this research, as well as various state-of-the-art models, namely HighRes-Net, RAMS, PIUNET and TR-MISR. The evaluation was conducted on both the simulated datasets (SRRB and SRRB_{enh}) and the real-world Proba-V dataset.

Each model was tested on the test subsets of each dataset. The quality of the super-resolved images was assessed based on several reconstruction quality metrics, as discussed in Chapter 5. Each metric offers a unique perspective on the quality of the super-resolved images, thereby enabling a comprehensive evaluation of the models' performance. In addition to these quality metrics, the time efficiency of each model was examined. This involved measuring the time taken by each model to produce a super-resolved image, which provides insights into their practical applicability in real-world scenarios where time efficiency can be crucial.

In addition to the state-of-the-art deep learning models, a multi-image version of bicubic interpolation was employed as a baseline for comparison. This version operates by interpolating each LR image independently before averaging the stack of upsampled images to generate the final super-resolved image. This process aligns with the methodology delineated in [123], allowing for a methodical comparison of bicubic interpolation with more advanced MISR models. The following sections delve into the detailed results of these evaluations, starting with the experiments on the simulated datasets.

6.1 Results on the Simulated Datasets

The super-resolution models were evaluated on both the SRRB and SRRB_{enh} datasets. Tables 6.1 and 6.2, respectively, detail the performance metrics for

each benchmark dataset independently, and when combined. To facilitate a direct comparison of model performance across these datasets, Table 6.3 provides the overall mean scores, consolidating results from both SRRB and $SRRB_{enh}$.

TABLE 6.1: Performance of super-resolution models on SRRB dataset. The table presents the mean scores of quality metrics across each benchmark dataset included in SRRB. The best scores are highlighted in bold, while the second-best scores are underlined.

Metric ↓	**Model** → **Dataset** ↓	Bicubic	HighRes-Net	RAMS	PIUNET	TR-MISR	MagNAt
cPSNR	BSDS100	25.46	29.11	31.27	31.42	29.67	**32.41**
	Manga109	25.56	32.11	34.81	34.35	33.35	**36.17**
	Set14	25.96	29.31	31.70	31.72	30.16	**32.55**
	Set5	28.91	33.77	35.10	36.41	34.79	**36.76**
	Urban100	23.11	27.43	29.59	28.74	27.87	**29.79**
	historical	22.10	26.53	28.50	28.72	26.99	**29.43**
cSSIM	BSDS100	0.737	0.878	0.921	0.925	0.885	**0.939**
	Manga109	0.852	0.959	0.975	0.976	0.966	**0.980**
	Set14	0.770	0.884	0.927	**0.930**	0.899	0.915
	Set5	0.865	0.954	0.967	**0.974**	0.964	**0.974**
	Urban100	0.739	0.887	**0.927**	0.915	0.894	0.924
	historical	0.704	0.885	0.92	0.922	0.889	**0.931**
cLPIPS	BSDS100	0.418	0.088	0.058	0.066	0.085	**0.052**
	Manga109	0.284	0.027	**0.017**	0.023	0.025	**0.017**
	Set14	0.358	0.065	**0.043**	0.051	0.060	0.048
	Set5	0.250	0.029	0.020	**0.016**	0.026	0.023
	Urban100	0.392	0.071	**0.048**	0.074	0.071	0.058
	historical	0.475	0.078	**0.054**	0.061	0.076	0.056
cMGE	BSDS100	0.080	0.028	0.017	0.017	0.025	**0.013**
	Manga109	0.110	0.022	0.012	0.014	0.016	**0.007**
	Set14	0.078	0.026	0.015	0.018	0.021	**0.014**
	Set5	0.053	0.012	0.006	0.006	0.008	**0.005**
	Urban100	0.172	0.053	**0.033**	0.044	0.048	**0.033**
	historical	0.152	0.049	0.037	0.035	0.046	**0.028**
TBE	BSDS100	0.384	0.302	0.297	0.301	0.304	**0.293**
	Manga109	0.412	**0.313**	0.318	0.326	0.320	0.315
	Set14	0.426	0.339	0.339	0.345	0.342	**0.337**
	Set5	0.474	0.391	0.395	0.397	0.397	**0.385**
	Urban100	0.406	0.318	0.317	0.328	0.321	**0.312**
	historical	0.352	0.273	**0.271**	0.278	0.278	0.272

The evaluation of super-resolution models across the simulated datasets provided several insights into their performance and adaptability. As observed for both simulated datasets, the bicubic interpolation trails behind the other models across all metrics. This consistent underperformance underscores its limitations when faced with the intricacies of the super-resolution task. The bicubic interpolation lacks the sophistication to accurately reconstruct HR details from multiple LR images. This deficiency becomes more

TABLE 6.2: Performance of super-resolution models on $SRRB_{enh}$ dataset. The table presents the mean scores for image similarity metrics for each model across each benchmark dataset. The best scores are highlighted in bold, while the second-best scores are underlined.

Metric ↓	**Model** → **Dataset** ↓	Bicubic	HighRes-Net	RAMS	PIUNET	TR-MISR	MagNAt
cPSNR	BSDS100	25.31	27.68	28.71	28.70	27.87	**29.39**
	Manga109	25.25	29.95	31.25	30.96	30.21	**32.17**
	Set14	25.66	27.63	29.05	28.78	27.89	**30.20**
	Set5	28.22	30.88	31.20	30.90	31.11	**31.58**
	Urban100	22.98	25.90	27.26	27.07	26.12	**28.54**
	historical	22.15	25.00	26.53	26.32	25.56	**27.08**
cSSIM	BSDS100	0.711	0.817	0.848	0.844	0.820	**0.861**
	Manga109	0.815	0.918	0.932	0.928	0.921	**0.935**
	Set14	0.743	0.822	0.861	0.849	0.829	**0.876**
	Set5	0.828	0.890	0.893	0.891	0.892	**0.895**
	Urban100	0.702	0.834	0.867	0.861	0.838	**0.887**
	historical	0.690	0.834	0.873	0.865	0.846	**0.886**
cLPIPS	BSDS100	0.447	0.180	**0.144**	0.152	0.175	0.149
	Manga109	0.314	0.082	0.075	0.073	0.079	**0.067**
	Set14	0.373	0.159	0.136	0.139	0.152	**0.129**
	Set5	0.280	0.112	0.108	**0.102**	0.107	**0.102**
	Urban100	0.428	0.122	0.102	0.107	0.121	**0.093**
	historical	0.490	0.132	**0.101**	0.117	0.126	0.102
cMGE	BSDS100	0.085	0.039	0.030	0.030	0.036	**0.025**
	Manga109	0.117	0.032	0.023	0.024	0.028	**0.016**
	Set14	0.084	0.041	0.027	0.030	0.035	**0.021**
	Set5	0.056	0.027	0.026	0.026	0.025	**0.018**
	Urban100	0.182	0.076	0.056	0.054	0.067	**0.037**
	historical	0.159	0.069	0.050	0.052	0.058	**0.041**
TBE	BSDS100	0.376	0.330	**0.323**	0.331	0.330	**0.323**
	Manga109	0.403	0.333	0.335	0.342	0.334	**0.331**
	Set14	0.415	0.371	0.368	0.375	0.368	**0.366**
	Set5	0.449	**0.426**	0.429	0.435	0.427	0.430
	Urban100	0.398	0.335	0.334	0.340	0.335	**0.331**
	historical	0.357	0.287	**0.280**	0.290	0.287	**0.280**

pronounced when handling the $SRRB_{enh}$ dataset, where the LR images in a stack contain variations not only in spatial alignment but also in global attributes. The struggle of the bicubic interpolation with these datasets emphasizes the need for more complex models to effectively tackle the super-resolution task.

HighRes-Net consistently outperformed the bicubic interpolation in every metric across all datasets, showcasing its enhanced capability to tackle super-resolution challenges. When compared with more advanced models, while its scores were commendable, especially in terms of minimizing blur (TBE metric), it rarely secured the leading position.

TABLE 6.3: Aggregated performance metrics for super-resolution models, combining results from both SRRB and $SRRB_{enh}$ datasets. This table presents the overall mean scores for each metric, across all benchmark datasets combined. The best scores are highlighted in bold, and the second-best scores are underlined.

Dataset	Model	cPSNR	cSSIM	cLPIPS	cMGE	TBE
SRRB	Bicubic	24.56	0.783	0.355	0.129	0.405
	HighRes-Net	29.60	0.913	0.057	0.036	<u>0.314</u>
	RAMS	<u>31.96</u>	<u>0.945</u>	**0.038**	<u>0.021</u>	0.315
	PIUNET	31.49	0.941	0.052	0.026	0.323
	TR-MISR	30.38	0.920	0.055	0.031	0.318
	MagNAt	**32.81**	**0.948**	<u>0.041</u>	**0.019**	**0.310**
$SRRB_{enh}$	Bicubic	24.35	0.749	0.386	0.136	0.396
	HighRes-Net	27.83	0.863	0.118	0.051	0.334
	RAMS	<u>29.10</u>	<u>0.888</u>	<u>0.100</u>	<u>0.037</u>	<u>0.333</u>
	PIUNET	28.90	0.884	0.103	0.038	0.340
	TR-MISR	28.06	0.867	0.115	0.045	0.334
	MagNAt	**30.12**	**0.901**	**0.094**	**0.026**	**0.330**

The RAMS model consistently demonstrates an impressive performance across all metrics in both simulated datasets. It particularly stands out for its low cLPIPS score, highlighting its proficiency at maintaining perceptual similarity during the super-resolution process. A key feature contributing to RAMS's notable performance is its use of 3D convolutions. These convolutions not only process individual image spatial information but also adeptly capture variations across image stacks. By effectively harnessing these relationships and intricacies among different images, RAMS manages to produce high-quality super-resolved images, underlining its robustness in confronting the nuanced challenges of super-resolution.

PIUNET's performance was consistently strong across both datasets, emphasizing its versatility and reliability. Although it registered high scores, especially in cPSNR and cSSIM metrics for the SRRB dataset, it did not always outpace every other model in all metrics. Such outcomes underline the close competition among super-resolution models, with each having unique strengths.

By examining the results, TR-MISR typically outperforms HighRes-Net across the evaluated metrics on simulated datasets. However, when compared to other state-of-the-art models, its performance seems to lag slightly. These findings could be indicative of the broader challenges associated with

the model's underlying design choices. TR-MISR, built on a transformer-based architecture, is inherently optimized for capturing complex dependencies and variations in data. Nevertheless, the nature of the simulated datasets, characterized predominantly by global image modifications, might not align seamlessly with the local complexities that transformers excel at modelling. Such a mismatch suggests that while transformers have shown significant promise in various domains, their application to specific super-resolution tasks might require a more tailored approach, especially when the dataset's nuances do not fully align with their strengths.

The MagNAt model frequently secured high rankings, often achieving the highest or near-highest scores across various metrics in both datasets. Its strengths in maintaining signal fidelity, retaining structural nuances, minimizing blur effects, and matching the perceptual attributes of the original HR image were evident. The commendable performance of MagNAt on both SRRB and SRRB_{enh} datasets highlights its potential in the super-resolution landscape.

Upon comparing the performance metrics across the SRRB and SRRB_{enh} datasets, a distinct trend emerges. Given the SRRB_{enh} dataset's more complex nature, it is evident that all models faced increased challenges, resulting in generally worse performance scores than when evaluated on the simpler SRRB dataset. This underscores the escalating difficulty of super-resolution tasks as dataset complexities rise.

6.1.1 Distribution Analysis

While averaging scores oftentimes gives a sufficient view of a model's performance, it is also valuable to look at the spread and distribution of the results. Figures 6.1 and 6.2 illustrate this effectively using half-violin plots, which combine two metrics in a single visual, providing a clearer picture of each model's performance on the SRRB and SRRB_{enh} datasets. The width of each violin shows how data is distributed: wider areas have more data points, while narrower ones have fewer. Instead of traditional box details inside, lines mark quartiles (dotted lines) and the median (solid line).

Upon examining the plots, a consistent trend emerges across both datasets. The distribution of scores for bicubic interpolation stands out, being distinctly wider, with its median positioned considerably distant from other

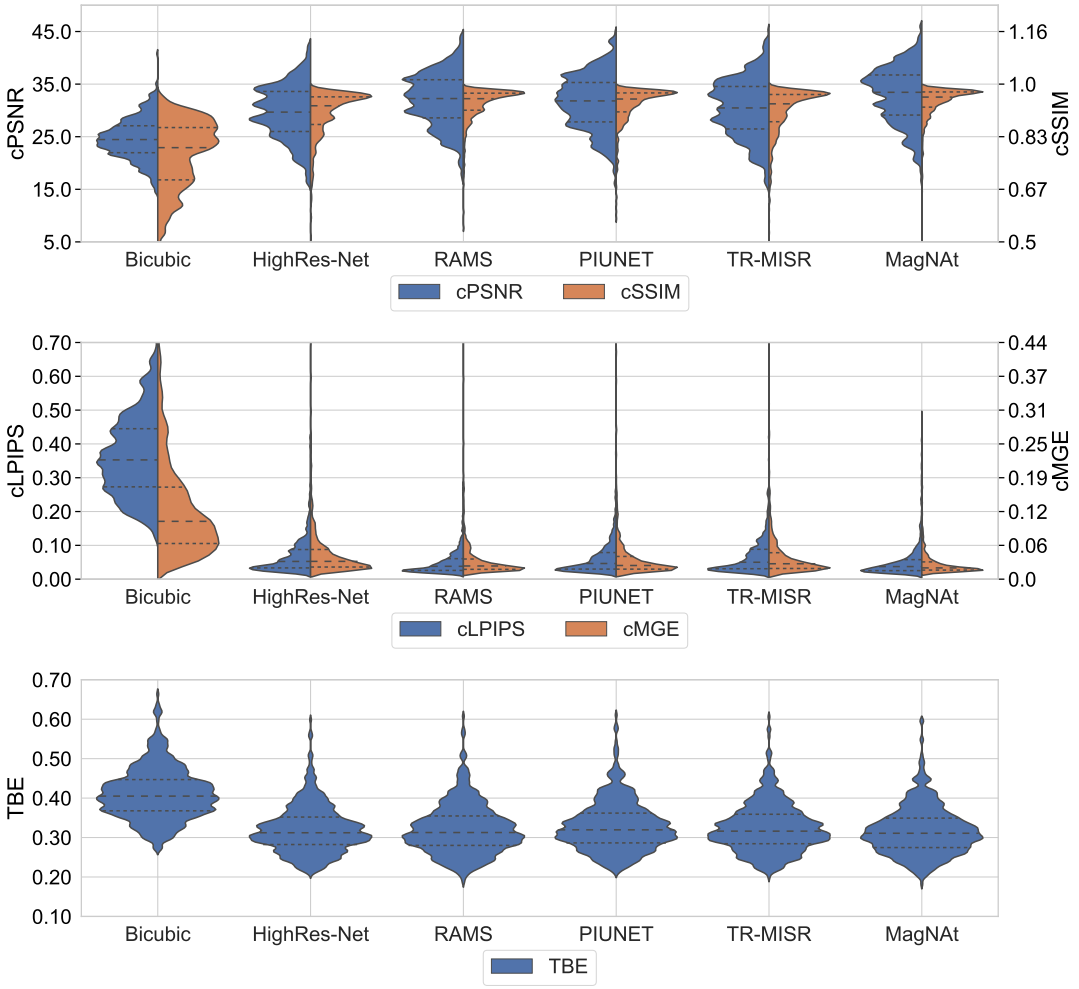


FIGURE 6.1: Half-violin plots of performance metrics for each model on the SRRB dataset, with quartiles (dotted lines) and median (dashed line) indicated.

models. This spread emphasizes its less consistent performance in comparison to more specialized super-resolution models. Moreover, among the contemporary models, MagNAt and RAMS distinctly manifest as the most consistent in their performance, closely trailed by PIUNET. Their score distributions are not just compact but also tend towards the better end of the scale, reinforcing their superior performance.

6.1.2 Assessing Statistical Significance of Model Performance

In experimental research, mere observation of performance differences between various models may not provide a comprehensive understanding. It

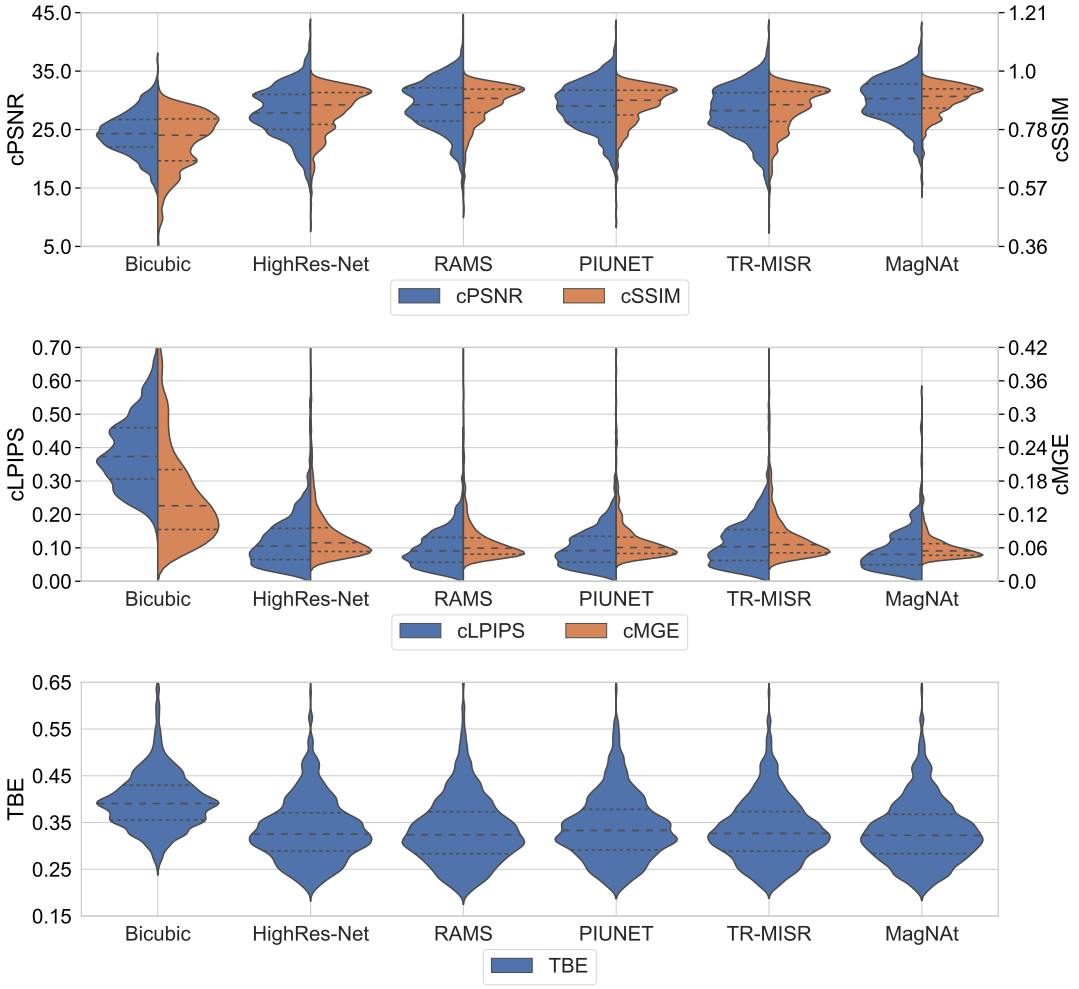


FIGURE 6.2: Half-violin plots of performance metrics for each model on the SRRB_{enh} dataset, with quartiles (dotted lines) and median (dashed line) indicated.

is essential to discern whether these observed variations are genuine or simply due to chance. Basing conclusions purely on observed means or distributions can be misleading, as they do not capture the variability within the data. Therefore, statistical tests are vital in validating the true significance of observed performance differences.

For this study, the *two-tailed Wilcoxon signed-rank test* [148] was employed to determine if there were statistically significant differences in scores between MagNAt and other models. The test's null hypothesis posits no significant performance discrepancy between the two models being compared. Rejecting this hypothesis indicates a statistically significant difference in performance. A conventional significance level is set at 0.05. If a p-value is less than this level, it suggests statistically significant performance differences for the evaluated metric.

The results of the two-tailed Wilcoxon signed-rank tests bolstered the performance evaluation. For the SRRB dataset, only in the instance of the cLPIPS metric when compared with RAMS, did the p-value (0.12) exceed the 0.05 significance threshold, indicating no statistically significant difference between the two. However, on the SRRB_{enh} dataset, MagNAt was statistically distinct from all other models across every metric, highlighting its superior performance with more complex simulated data.

6.1.3 Qualitative Analysis on Simulated Dataset

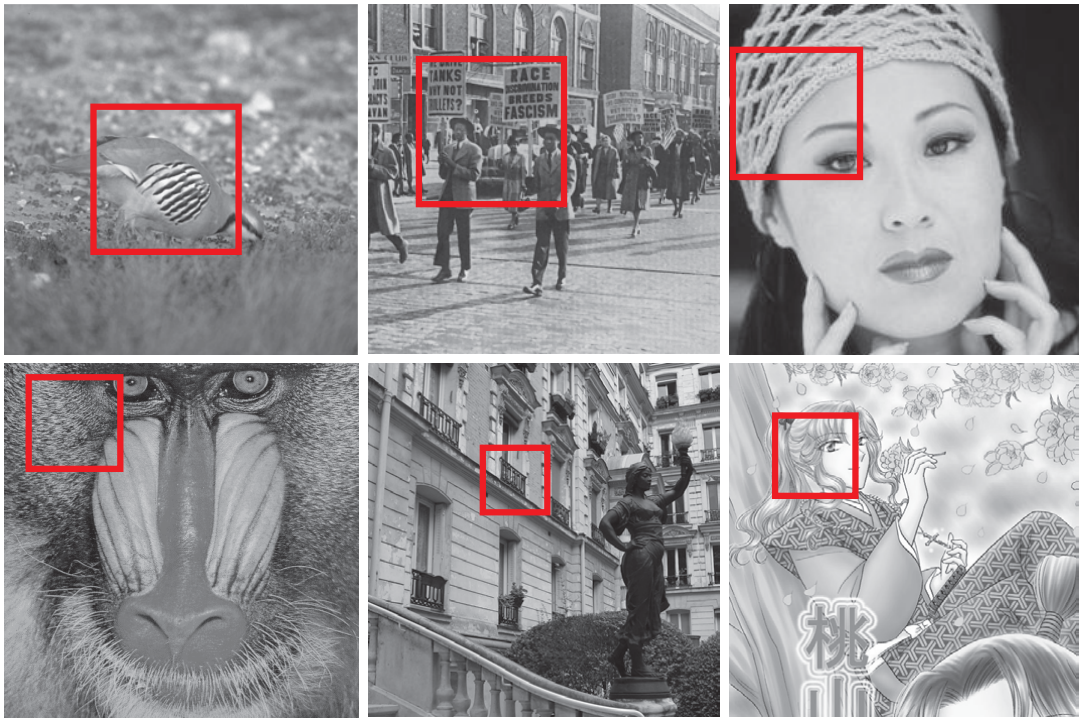


FIGURE 6.3: High-resolution target images utilized in this qualitative assessment. Each image originates from a distinct benchmark dataset, respectively (from top-left to bottom-right): BSDS100, Historical, Set5, Set14, Urban100, Manga109. The red rectangles highlight the regions of interest examined in the qualitative analysis.

Visual examinations of super-resolution outputs serve as critical adjuncts to quantitative metrics, offering intuitive insights into model performance. The high-resolution target images utilized in this qualitative assessment are depicted in Figure 6.3 with red rectangles denoting their specific regions on which this visual analysis is conducted. These super-resolved regions are shown in Figures 6.4-6.6, providing a better understanding of each model's strengths and weaknesses.

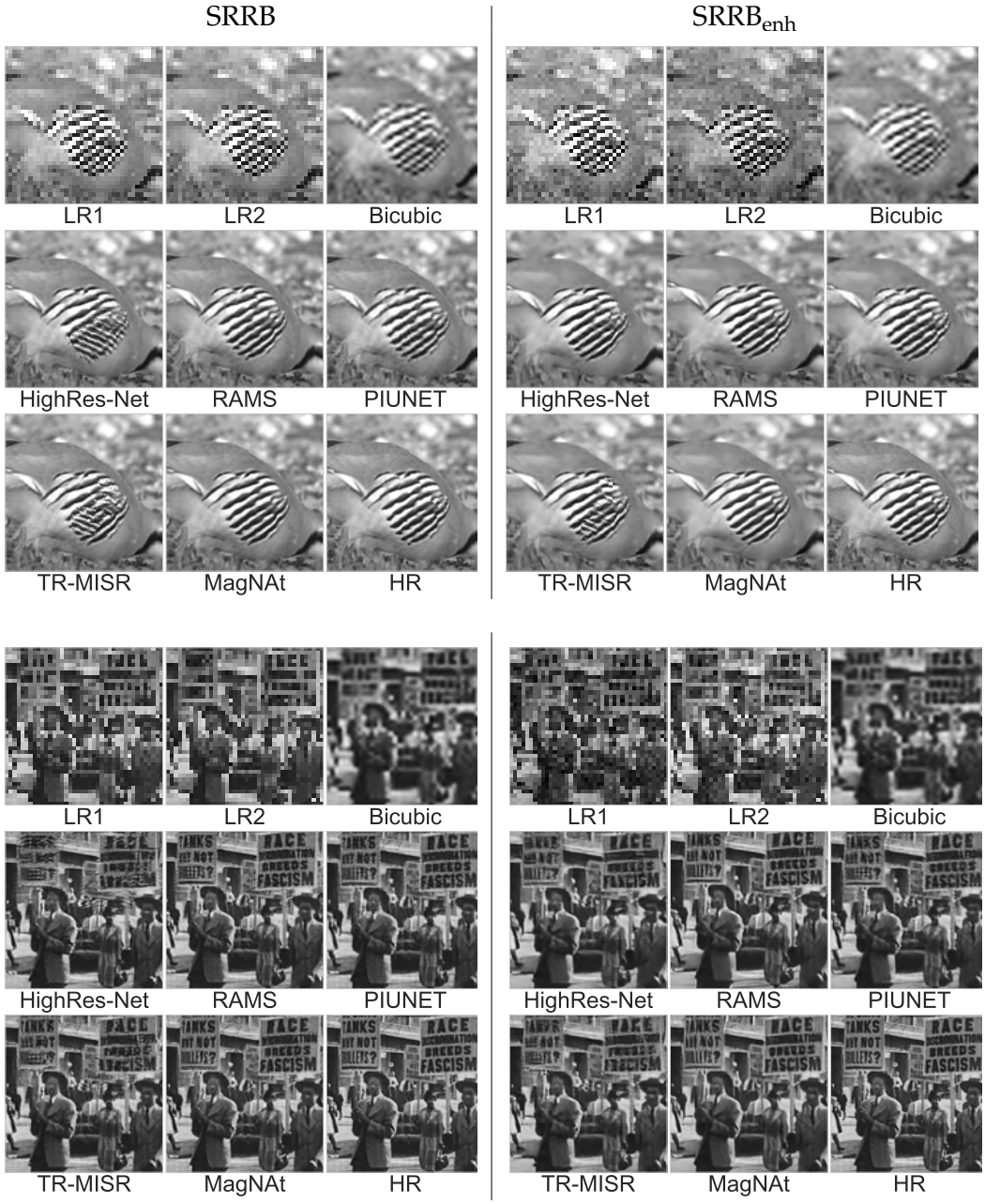


FIGURE 6.4: Examples of super-resolved simulated images from BSDS100 (top) and historical (bottom) datasets.

In the analyzed super-resolution outputs, high-frequency details, initially absent in each LR image, were effectively reintroduced by the models. However, subtle discrepancies in their outputs became apparent, especially in regions featuring complex textures. These variations, often emerging as disruptions in texture or irregular pixel transitions, compromised the visual fidelity of the super-resolved images.

This was particularly evident in high-frequency repeatable patterns such as the parallel lines on the bird's wings in Figure 6.4 or the fur in Figure 6.5. In

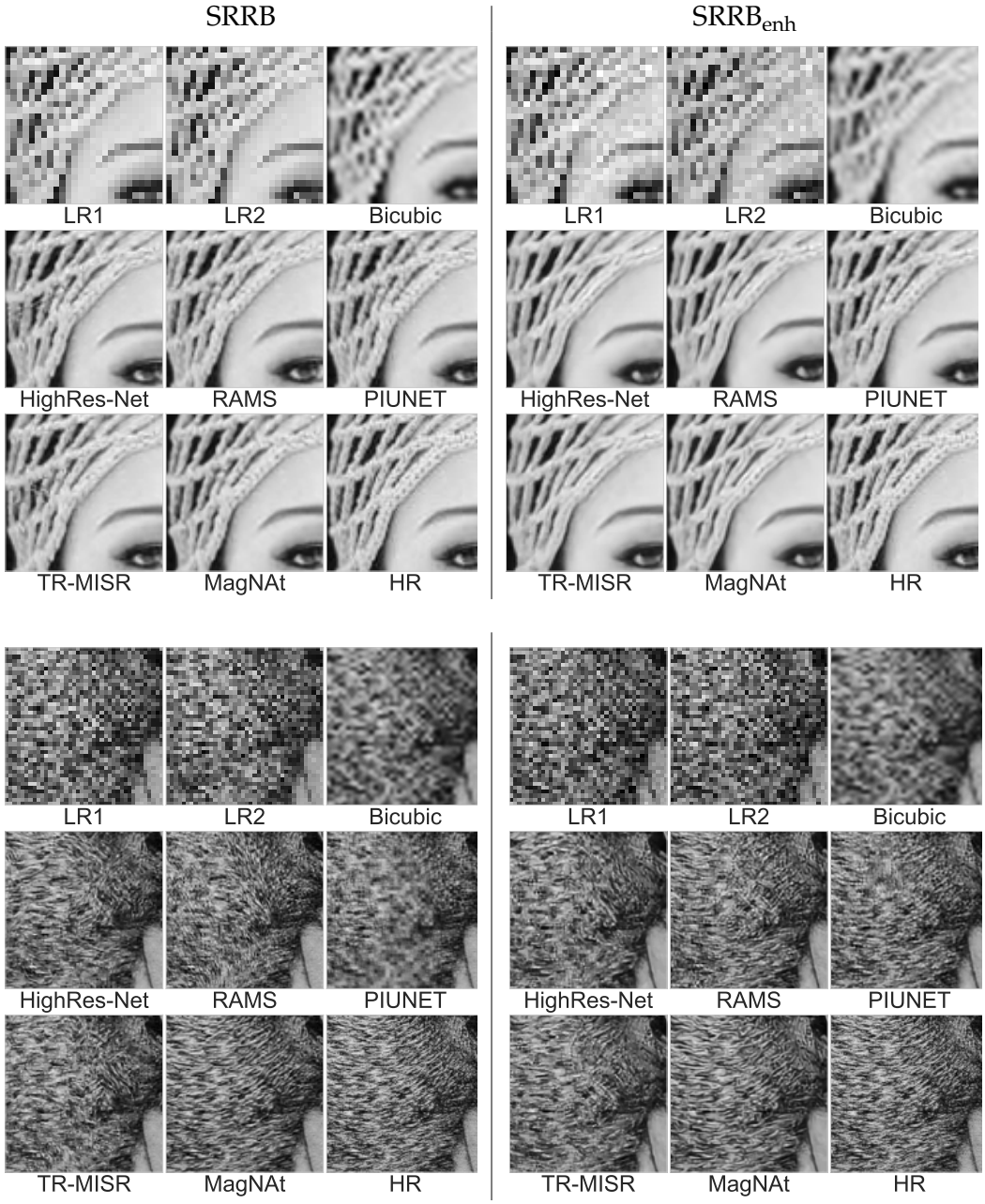


FIGURE 6.5: Examples of super-resolved simulated images from Set5 (top) and Set14 (bottom) datasets.

these examples, while the MagNAt model effectively captured the textures, other models, particularly HighRes-Net and TR-MISR, struggled, often misrepresenting the true direction and nuance of these patterns.

6.1.4 Considerations on Simulated Data

The simulated datasets SRRB and SRRB$_{enh}$ used in this study possess characteristics that might influence model performance. They notably lack local temporal variations seen in real-world scenarios, potentially simplifying the

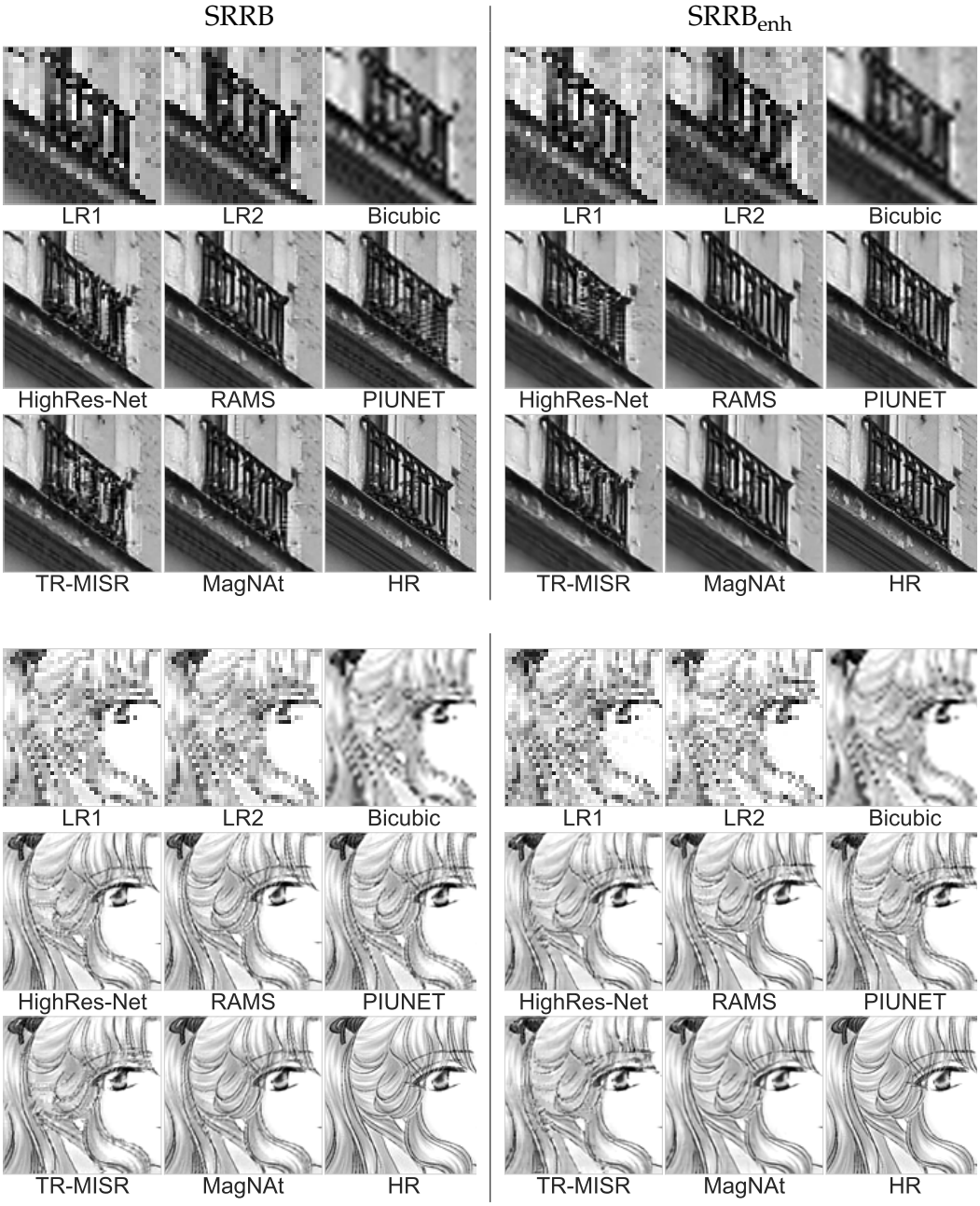


FIGURE 6.6: Examples of super-resolved simulated images from Urban100 (top) and Manga109 (bottom) datasets.

super-resolution task due to the absence of these complexities and the likely reduced registration errors during LR image alignment. Given this context, the strong performance of MagNAt on these datasets might be influenced by its unique attributes. While most models rely on full-pixel registration, the MagNAt model differentiates itself by demanding precise subpixel registration. This trait, combined with its adaptive registration method, might

give MagNAt an advantage in handling minor registration errors, contributing to its enhanced performance in these scenarios. However, despite MagNAt's positive results on the simulated datasets, extending validations with real-world data remains crucial. Engaging with challenging datasets like the Proba-V will shed light on the models' capabilities under more realistic conditions, providing a comprehensive view of their actual strengths and limitations. Nonetheless, these experiments on simulated datasets can be seen as a step forward in proving the first thesis of this dissertation, stating that when a set of LR images with sub-pixel shifts is represented as a graph, GNNs can process this graph to achieve super-resolution results comparable or even superior to those obtained by leading MISR architectures reliant on convolutional networks.

6.2 Real-World Evaluation: The Proba-V Dataset

In this section, a detailed analysis of the performance of the models on both spectral subsets of the Proba-V datasets is presented, along with a discussion on the visual quality of super-resolved images, and the results of the statistical tests. The aim is to ascertain whether the conclusions drawn from the simulated datasets are upheld in a real-world, remote-sensing scenario and to potentially uncover new insights into the performance of these models under more complex conditions.

6.2.1 Performance Dynamics with Varying Number of Input Images

The analysis started by inspecting how the number of LR input images (N) affects the performance of super-resolution models. The goal was to find out how many input images each model needs for the best performance, ranging from a single image up to 32. To keep the input quality consistent across different N values, images were chosen based on clearance maps, with preference given to the least obscured images. The clearest images were included in every set, while more obscured ones were added gradually and only fully included at the highest N.

Because the performance metrics are multifaceted, a combined metric was created to simplify the process of finding the optimal N. This metric mixed standardized versions of five different performance metrics, summing the cPSNR and cSSIM values and subtracting the cLPIPS, cMGE, and TBE values,

for which lower scores indicate better performance. Using this combined metric, the N value that scored the highest was chosen as the optimal N for each model. This assessment was done on the validation subsets, which helped fine-tune the N parameter to evaluate each model's performance on the test subsets.

The data was visualized using line plots to explore the relationship between the number of LR input images and the performance metrics for each model. This is shown in Figure 6.7 and Figure 6.8 for NIR and RED bands, respectively. The best metric scores for each model are marked as dots. Although these plots revealed how the models' performance changed with varying N values, it is worth noting that the RAMS model can only use nine images due to its architectural design; thus, its results appear as single points.

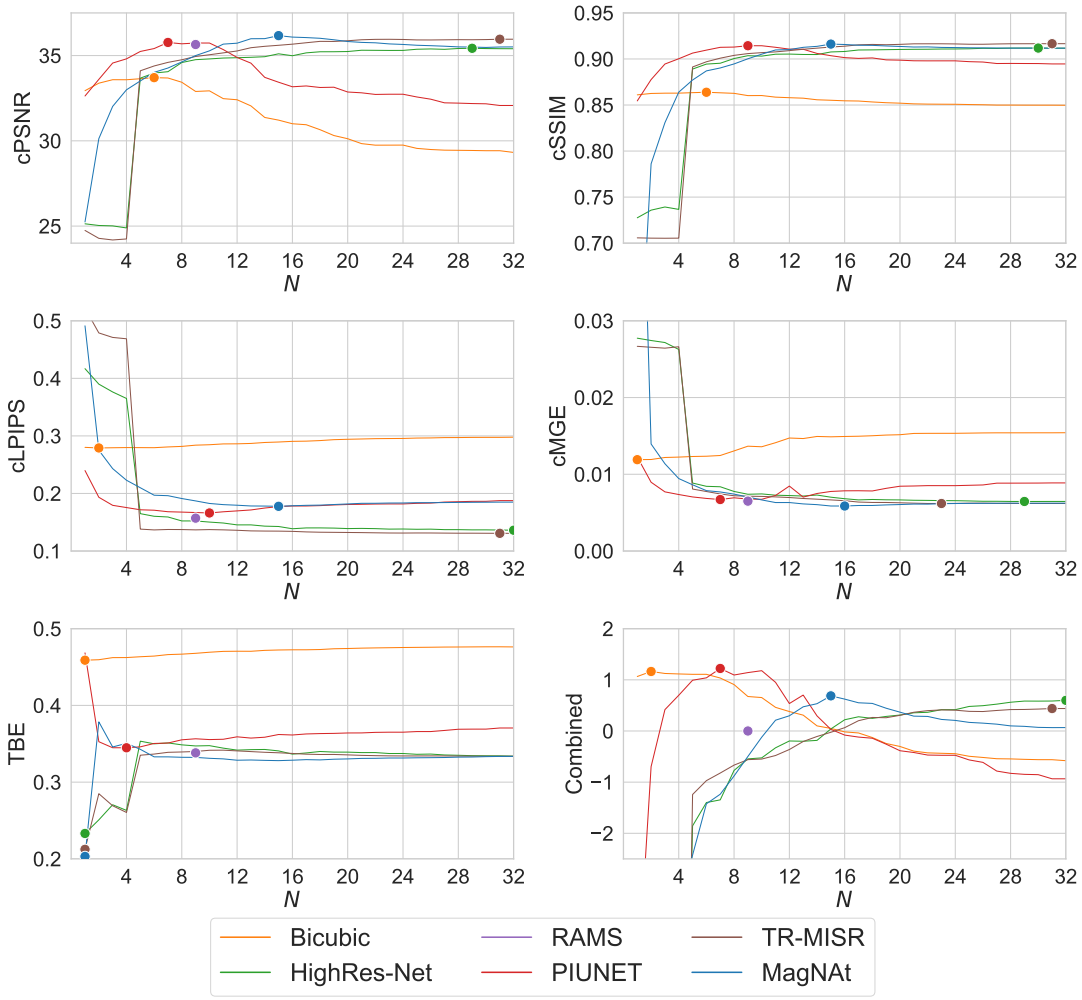


FIGURE 6.7: Performance dynamics of the models on the NIR subset of the Proba-V dataset against varying LR input images (N).

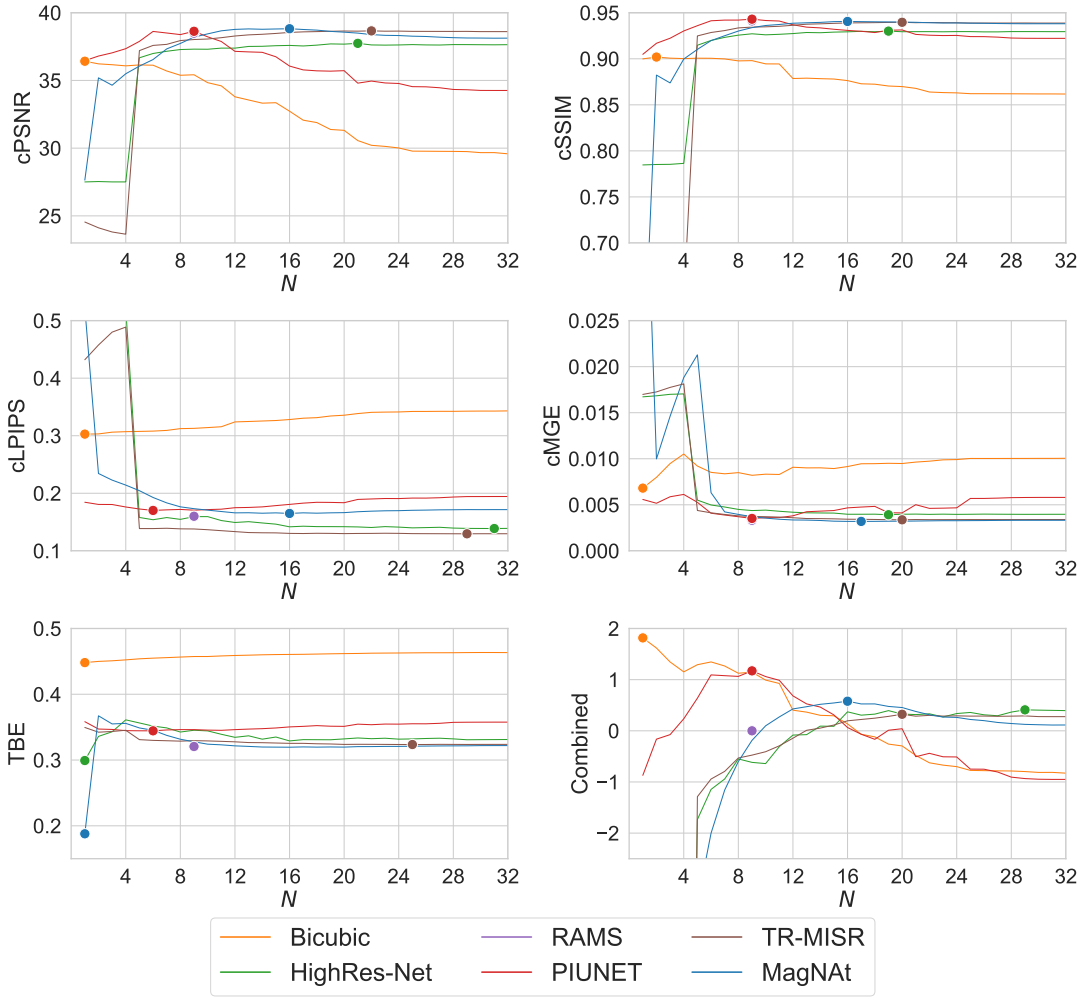


FIGURE 6.8: Performance dynamics of the models on the RED subset of the Proba-V dataset against varying LR input images (N).

The trends for every method in both spectral bands are observed to be very similar, underscoring the consistency in each model's response to varying N. The bicubic interpolation, simply averaging the upsampled LR images, demonstrates a preference for a smaller number of images. As their number increases, the reconstruction quality tends to decline, which suggests bicubic's limited capacity to utilize multiple LR inputs for enhanced super-resolution.

The performance of TR-MISR and HighRes-Net enhances with the increase in N, indicating a positive relationship between the number of input images and improved reconstruction quality. Interestingly, with a higher number of input images, their results begin to plateau. This convergence hints at a saturation point in performance, suggesting that even if more images were available, the models might not achieve significantly better results,

or the improvement might be marginal at best. Yet, this observation requires further validation to conclusively determine the models' behaviour with a higher number of LR image

Distinct from others, PIUNET and MagNAt exhibit a performance peak at specific N values, after which their efficiency tends to decline. MagNAt necessitates a higher number of images compared to PIUNET to achieve its optimal performance. Interestingly, the performance of PIUNET begins to decline post the nine-image mark, which is the exact number of images it was initially trained on. This trend underscores a nuanced interaction between the number of input images and the architectural design of these models, pointing out a threshold beyond which additional images fail to contribute to better super-resolution results.

The optimal N values were identified by studying the line plots for the combined metric, as shown in Figures 6.7 and Figure 6.8. The information from these plots helped determine the right number of LR input images for each model, improving their performance. The optimal N values, for both the NIR and RED bands separately, are listed in Table 6.4.

TABLE 6.4: Optimal number of LR input images (N) for each method across NIR and RED bands, calculated on the validation subset.

Model	**NIR**	**RED**
Bicubic	2	1
HighRes-Net	32	29
RAMS	9	9
PIUNET	7	9
TR-MISR	31	24
MagNAt	15	16

6.2.2 Metric Scores for Optimal Number of LR Images

Having determined the most optimal number of images for each model from the validation subset, as showcased in Table 6.4, the models were subsequently evaluated on the test subsets of the Proba-V dataset. This quantitative comparison is presented in Table 6.5. As anticipated, bicubic interpolation fell behind other methods in every metric, underscoring its limitations in this complex task. However, for the rest of the models, a distinct performance pattern emerged.

TABLE 6.5: Performance metrics obtained by all tested methods on the Proba-V dataset. The best scores are highlighted in bold, while the second-best scores are underlined for each spectral band independently.

Band	Model	cPSNR	cSSIM	cLPIPS	cMGE	TBE
NIR	Bicubic	33.380	.8625	.2791	.0119	.4595
	HighRes-Net	35.401	.9117	.1361	.0065	.3338
	RAMS	35.648	.9148	.1571	.0065	.3382
	PIUNET	35.769	.9127	.1683	.0067	.3510
	TR-MISR	35.958	**.9166**	**.1307**	.0062	.3337
	MagNAt	**36.169**	.9161	.1777	**.0059**	**.3280**
RED	Bicubic	36.419	.9000	.3028	.0068	.4481
	HighRes-Net	37.743	.9337	.1393	.0037	.3245
	RAMS	38.492	.9411	.1601	.0033	.3206
	PIUNET	38.629	**.9430**	.1706	.0035	.3462
	TR-MISR	38.650	.9396	**.1299**	.0033	.3238
	MagNAt	**38.819**	.9406	.1649	**.0032**	**.3195**

The MagNAt model, mirroring its previous performance on simulated datasets, excelled in cPSNR, cMGE, and TBE metrics, affirming its capacity to uphold high fidelity and sharpness akin to the HR reference images. Nevertheless, a deviation was observed in the cSSIM metric, where MagNAt now ranked second and third places for NIR and RED bands, respectively. Additionally, both MagNAt and PIUNET experienced a significant incline in cLPIPS scores, showing challenges in creating images perceptually akin to their HR counterparts.

Among other models, a rearrangement in performance rankings was noticeable. The RAMS model, despite its strong performance in simulated data, demonstrated mixed outcomes, shining solely in cSSIM, cMGE, and TBE metrics but only on the RED band data. On the other hand, the TR-MISR model manifested a notable improvement, especially in the NIR band, where it consistently secured first or second places, showcasing its relative adeptness in navigating real-world data complexities.

Distribution of Metric Scores

The performance score distributions, illustrated in Figures 6.9 and 6.10 for NIR and RED spectral subsets, respectively, reveal the nuanced behaviours of models on the Proba-V dataset at their optimal N. A recurring observation across both subsets is the consistent underperformance of the bicubic model, aligning with trends noticed in simulated datasets.

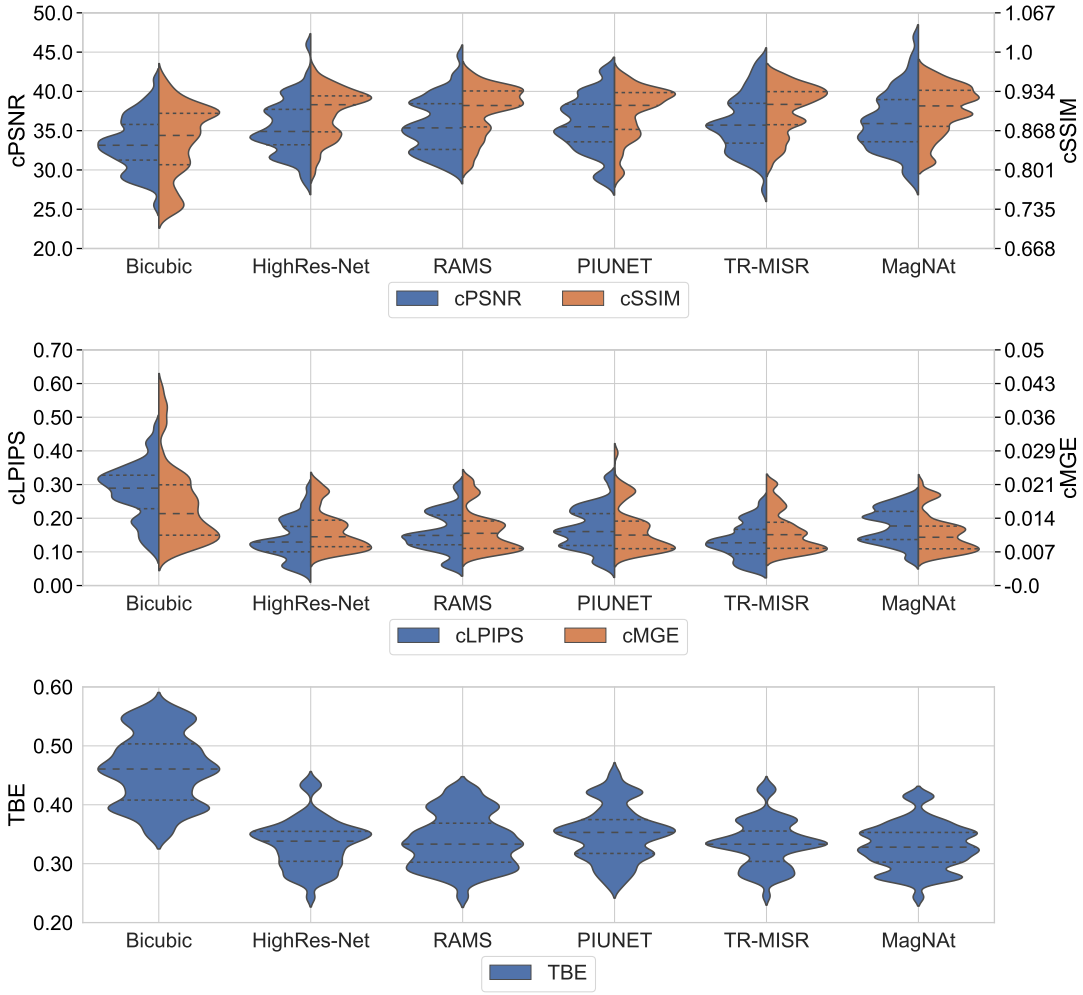


FIGURE 6.9: Violin plots showing the distribution of metric scores for each model on the NIR subset of the Proba-V dataset. The scores are calculated for the optimal number of LR input images for each model.

Identifying a clear leader among the remaining models is challenging due to the irregular behaviour reflected in the metrics. Instead of simply adhering to a Gaussian distribution, the metrics exhibit varied and complex patterns, reflecting the models' different strategies in handling the challenges posed by the real-world scenarios of the Proba-V dataset. This consistency in observations across both spectral subsets highlights the multifaceted performance characteristics exhibited by the models, which, in turn, underscores the complexities involved in transitioning from simulated to real-world datasets.

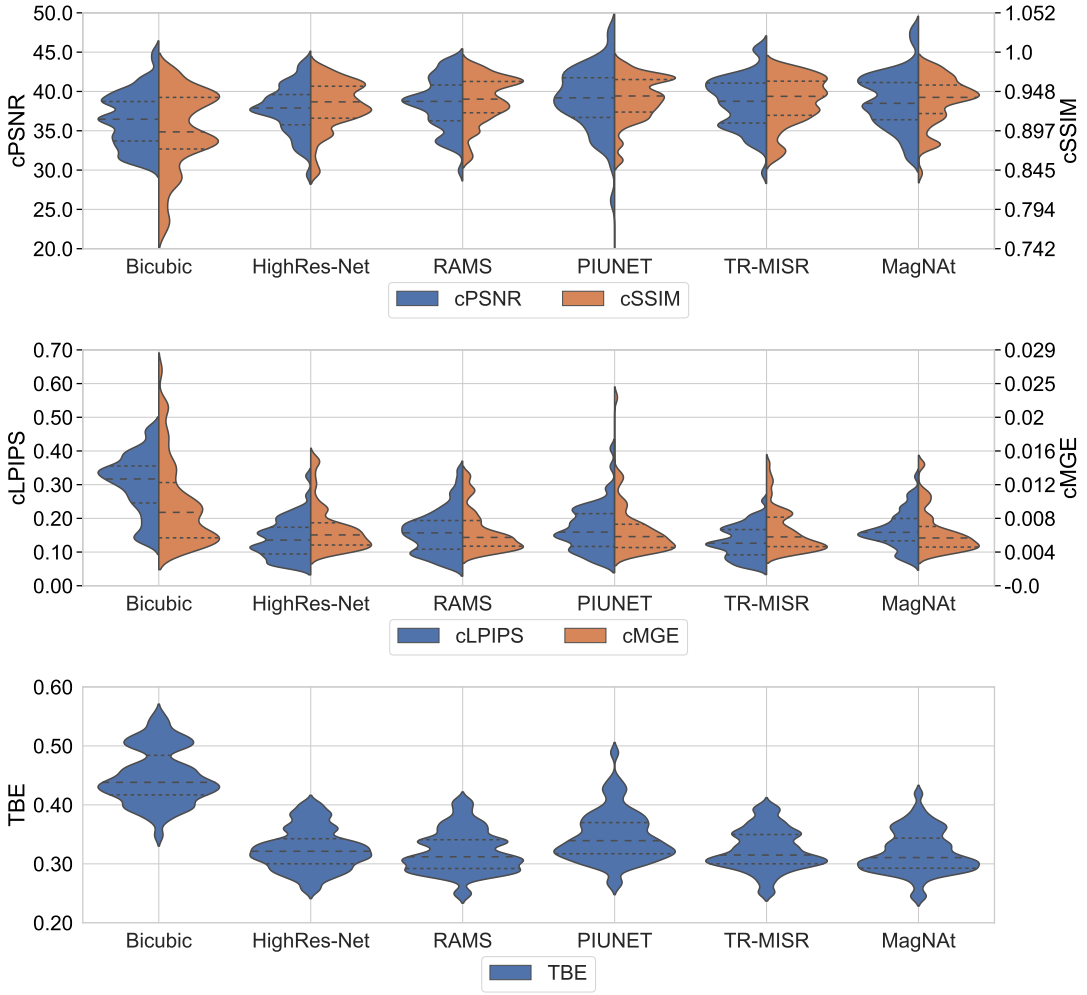


FIGURE 6.10: Violin plots showing the distribution of metric scores for each model on the RED subset of the Proba-V dataset. The scores are calculated for the optimal number of LR input images (N) for each model.

6.2.3 Statistical Significance Testing

The statistical validation was conducted using the Wilcoxon signed-rank test as detailed in the section on simulated datasets, extending the analysis to the performance metrics of MagNAt and other super-resolution models on both spectral subsets of the Proba-V dataset. The derived p-values are depicted in correlation matrices shown in Figures 6.11 and 6.12 for the NIR and RED subsets, respectively.

Upon comparison, MagNAt either outperformed or showcased no significant difference against other models across most metrics, attesting to its competitive performance. Specifically, MagNAt consistently displayed statistical significance in its favour in the TBE metric on NIR, whereas on the

RED subset, it showed statistical superiority only over Bicubic, HighRes-Net, and PIUNET in this metric. Additionally, MagNAt excelled in the MGE metric on NIR. However, in the cLPIPS metric, both HighRes-Net and TR-MISR manifested statistically significant advantages over MagNAt across both subsets, as substantiated by low p-values and superior mean scores, indicating their enhanced perceptual quality outcomes.

In summary, the analysis underscores MagNAt as a solid performer in real-world MISR applications, especially exhibiting statistical superiority in rendering sharper images, as evidenced by the MGE and TBE metrics. However, its weakness surfaces in the cLPIPS metric, highlighting a perceptual similarity gap when compared to HighRes-Net and particularly TR-MISR. Among the models compared, TR-MISR emerges as MagNAt's strongest competitor, showcasing its lead in cLPIPS scores for both subsets while only falling behind in TBE on the NIR subset. This analysis, both in statistical validation and mean score evaluations, not only reinforces MagNAt's strengths and areas of improvement but also sheds light on the nuanced competitive landscape of MISR.

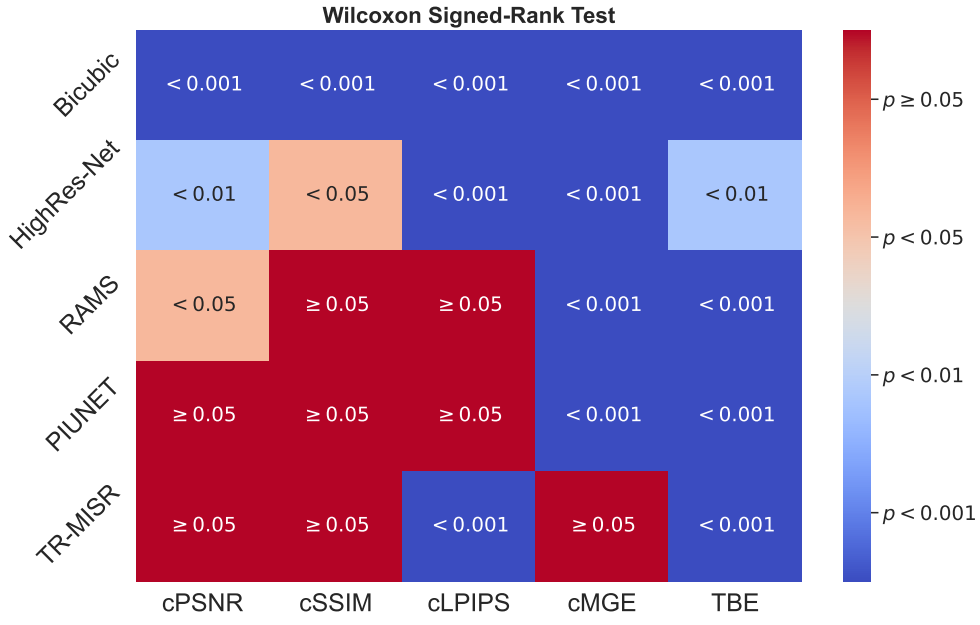


FIGURE 6.11: Correlation matrix of p-values from the two-tailed Wilcoxon signed-tank test comparing MagNAt against other models for various metrics on the NIR subset.

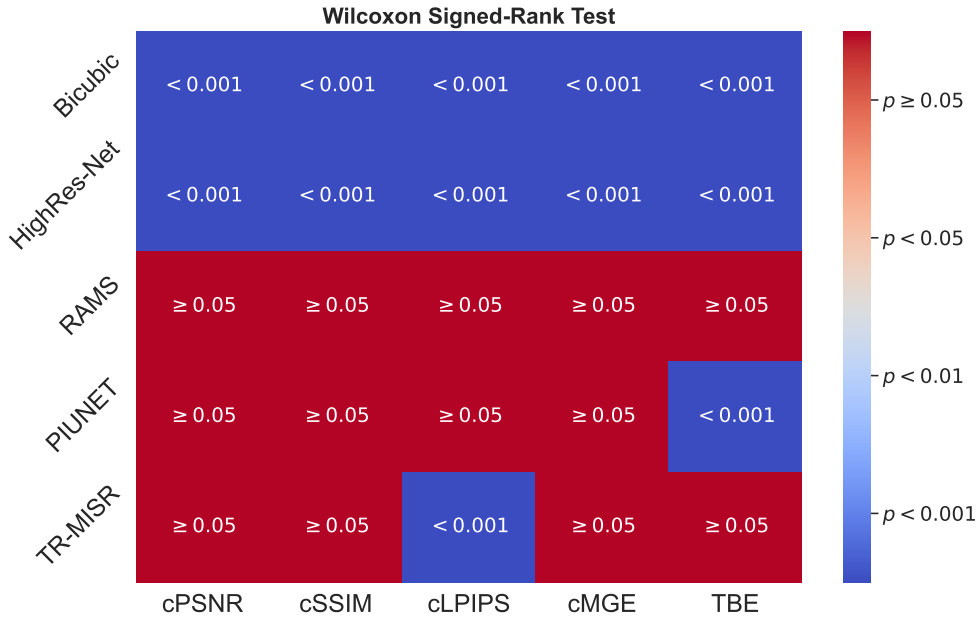
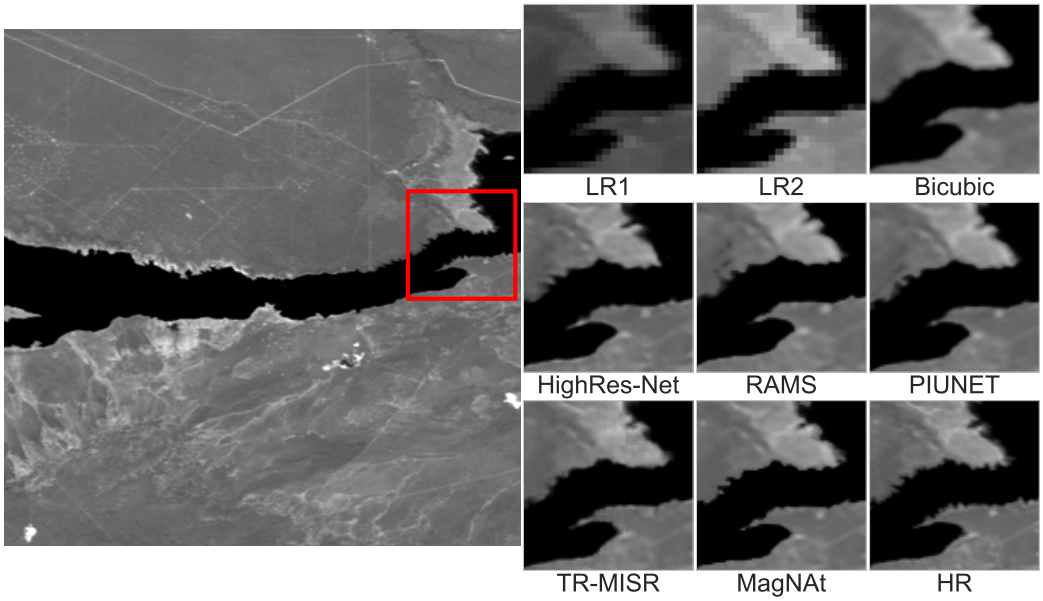


FIGURE 6.12: Correlation matrix of p-values from the two-tailed Wilcoxon signed-rank test comparing MagNAt against other models for various metrics on the RED subset.

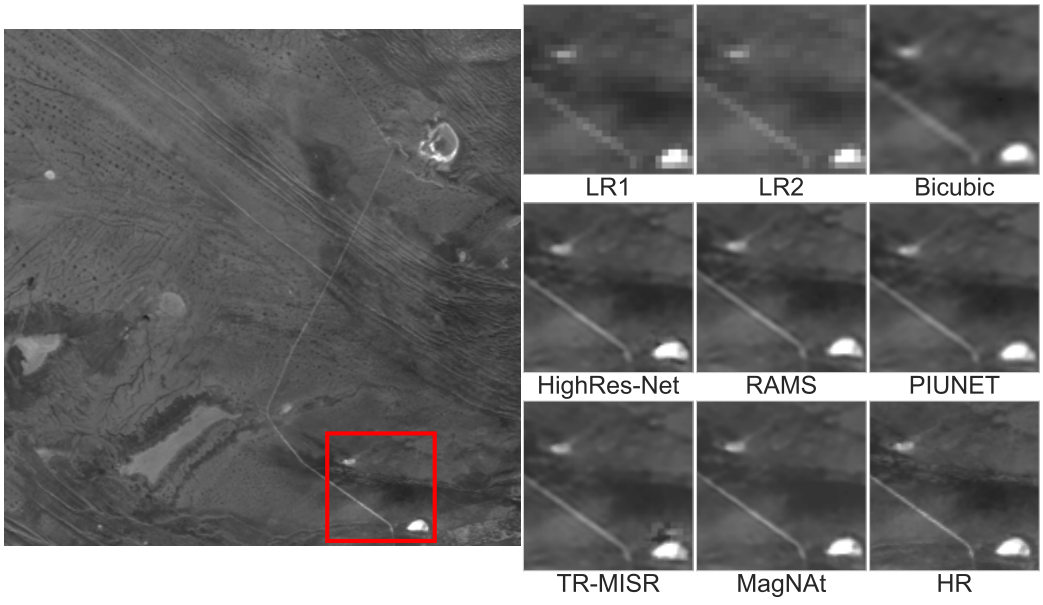
6.2.4 Qualitative Analysis on NIR and RED Subsets

Figure 6.13 visually compares models' performance on the NIR and RED subsets of the Proba-V dataset. The same 80x80 region was used for a direct comparison, allowing for an immediate evaluation of each model's super-resolution capabilities. Bicubic interpolation, as expected, yields images that are less sharp and lack the fine details present in the HR images, showcasing its limitation as a traditional interpolation method.

In the NIR example, the shoreline of a river depicted in the HR image exhibits a level of roughness and irregularity, which is not fully captured by any of the models, including MagNAt. However, MagNAt does render these shoreline edges with sharper contrast compared to other models, yet they still are not as detailed as in the reference image. Similarly, in the RED example, MagNAt delineates the road with slightly better sharpness compared to other models, albeit the difference is not as pronounced as in the NIR example. A notable observation in this scene is the visual artefact above the white object in the output of TR-MISR, likely resulting from high variability between input images in this specific region—an observation discussed further in Section 6.3.



(A) Comparison on NIR subset.



(B) Comparison on RED subset.

FIGURE 6.13: Visual comparisons of different models on the NIR (A) and RED (B) subsets of the Proba-V dataset. The images are cropped to a 150×150 region centred at the same location for all models. The red rectangle on the HR image (leftmost) indicates the cropped region. The results are presented for the optimal number of LR input images (N) for each model.

These visual assessments suggest a potential edge for MagNAt in handling certain scene details and producing sharp images, laying a groundwork for further refinement in MISR tasks on real-world datasets like Proba-V.

The accumulated insights from experiments on both simulated and real-world data serve as compelling evidence in support of the primary thesis of this dissertation. This hypothesis explored the potential of employing GNNs to process a graph representation of LR images with sub-pixel shifts, aiming to achieve super-resolution outcomes that could match or exceed the performance of prevailing MISR architectures based on convolutional networks. The promising results affirm the notion that adapting such an approach not only holds merit but opens up avenues for further exploration and optimization in the domain of MISR.

6.2.5 Benchmark Performance on the Proba-V Challenge

At the time of composing this dissertation, the MagNAt model is ranked sixth in the post-mortem Proba-V super-resolution challenge. The challenge evaluates models' performance on 144 NIR and 146 RED scenes and is based solely on the cPSNR metric. It should be noted that the test subset utilized by the Proba-V challenge is distinct from the subsets used in previous experiments of this study, primarily due to the absence of HR references, which prevents local model evaluations.

The leaderboard of the challenge is presently led by the TR-MISR model, with PIUNET positioned second. There are three other models ranked between MagNAt and PIUNET, which, to the best of the author's knowledge, have not been reported in available literature or peer-reviewed publications. This lack of transparency makes direct comparisons with MagNAt challenging. Evaluations for this challenge are executed on servers maintained by the organizers, ensuring a consistent evaluation environment for all submitted models. The noteworthy ranking of the MagNAt model in this esteemed benchmark highlights its robust capabilities in the realm of MISR tasks. As advancements in the field of MISR persist, further refinements in both MagNAt and competing models can be expected.

6.3 Temporal Variations and Super-Resolution

Temporal variations in the Proba-V dataset provide a unique challenge for super-resolution tasks. Given the evident discrepancies between LR images

captured at different time frames, super-resolution models are presented with the task of fusing contradictory information. To visually illustrate the complexities and the performances of the different models in handling these challenges, two distinct scenes from the dataset are presented.

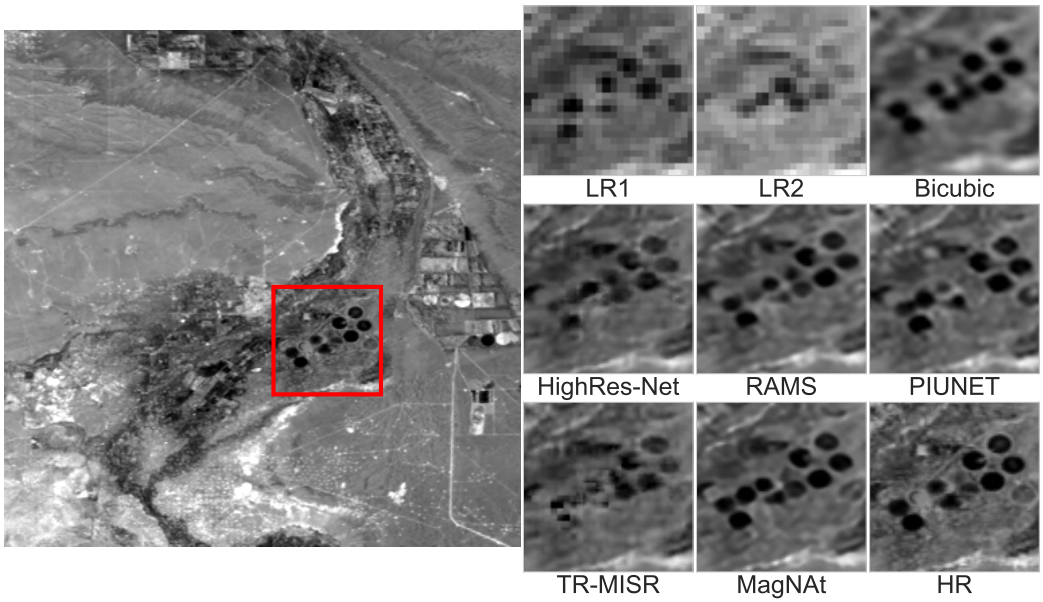


FIGURE 6.14: Visual comparison of super-resolution outputs for a scene containing centre-pivot irrigation fields with significant temporal differences between the LR images.

In the scene depicted in Figure 6.14, the centre-pivot irrigation fields stand out in the cropped region of interest. The vast temporal difference between the two presented LR images poses a challenge for the models, as they have to make decisions on whether to combine or select specific features. This decision-making process is reflected in the varied outputs of the models, with none being a clear match to the HR image. Notably, TR-MISR's output exhibits visual artefacts, suggesting difficulties in handling regions with high temporal variance.

Interestingly, the TR-MISR model displays peculiar behaviour in areas of pronounced temporal variance. In these regions, the model seems to produce artefacts that resemble sections of the LR image that have been merely enlarged without enhancement. It appears as though TR-MISR avoids the complex task of super-resolving these challenging areas, leading to regions in the output that appear as if they were directly taken from the LR image without any enhancement or detail addition. This results in the appearance of

larger, coarse pixels, giving the impression of a lack of true super-resolution in those specific regions. It is surprising that despite these evident shortcomings in handling highly temporally-variant regions, TR-MISR stands as one of the most competitive models to MagNAt in terms of quantitative assessment, especially given the inherent challenges and temporal variances present in the Proba-V dataset.

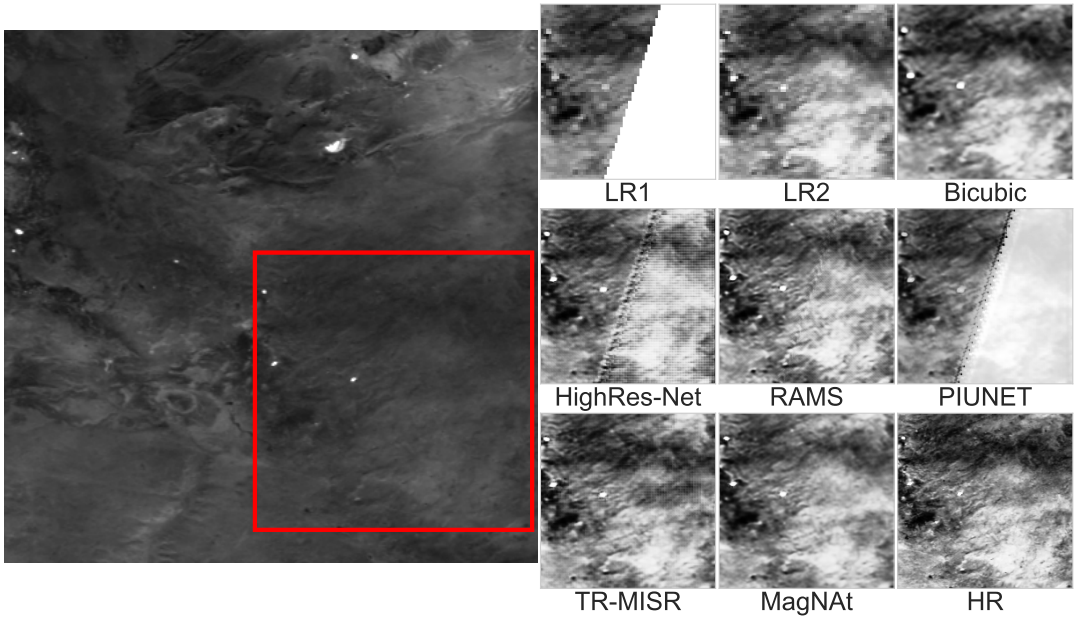


FIGURE 6.15: Visual comparison of super-resolution outputs for a scene with uncaptured pixels in one of the nine LR images, resulting in a white patch.

The next depicted scene in Figure 6.15 offers a different challenge. Some pixels in this scene inherit the maximum possible value due to certain problems in the acquisition process. It becomes apparent that most models struggle to handle these white patches appropriately, leading to visible artefacts. The use of histogram equalization reveals these artefacts more distinctly, especially in the outputs of RAMS and TR-MISR, which might be less noticeable without this adjustment. In contrast, MagNAt stands out by producing an image devoid of such artefacts, emphasizing its resilience and adaptability in handling the challenges posed by temporal variance and uncaptured pixels. A possible explanation for such behaviour is the use of the multi-level attention mechanism in MagNAt model, setting the importance of nodes with respect to each of their neighbours. It might have been learned by the model that such purely white pixels should not be taken into consideration when

reconstructing the scene. If this is true, such nodes would send none to minimal information about their features to their neighbours, thus reducing their influence on the reconstruction output. However, at the moment, this observation remains only on a speculative level and requires further investigation.

The visual examinations presented in this section underscore the nuanced challenges that temporal variations introduce to the super-resolution process. The varied responses of models to these challenges, from managing significant temporal differences to addressing uncaptured pixels, shed light on the balance required between data interpretation and model design. While certain models face difficulties, introducing artefacts or blending details, the performance of the MagNAt model stands as a beacon of potential. It showcases that, with a tailored approach, one can traverse the intricacies of real-world datasets, producing super-resolution outcomes that accentuate image details while retaining the temporal and structural essence of scenes.

6.3.1 Leveraging a Leading Image

From the prior evaluations, it becomes clear that managing the challenges presented by temporal variations is crucial for super-resolution tasks on satellite images. Super-resolution models, in their quest for high-fidelity outputs, often face challenges due to inconsistent data from different time frames. Addressing this inconsistency necessitates a guiding reference that would steer the models to reconstruct the features related to a specific point in time. Taking inspiration from the work of Nguyen et al. [99], where the authors identified the LR image most similar to the downsampled HR for each scene, a similar approach was adopted in this research. Recognizing that every scene in the Proba-V dataset has an LR captured concurrently with its HR counterpart, they tailored existing models to accommodate this knowledge. This approach resulted in increased quantitative performance of each tested model and a more pronounced resemblance of outputs to specific leading LR images. In line with their strategy, for each scene, a specific LR image was designated as the leading one for the $\text{MagNAt}_{\text{lead}}$ model (Section 3.6) in this analysis. By focusing on reproducing the attributes of this leading LR, the $\text{MagNAt}_{\text{lead}}$ can navigate the ambiguities more effectively, harnessing its inherent details and nuances. This approach offers a clearer, more defined task for the model, assisting in the mitigation of challenges introduced by pronounced temporal variations in the dataset.

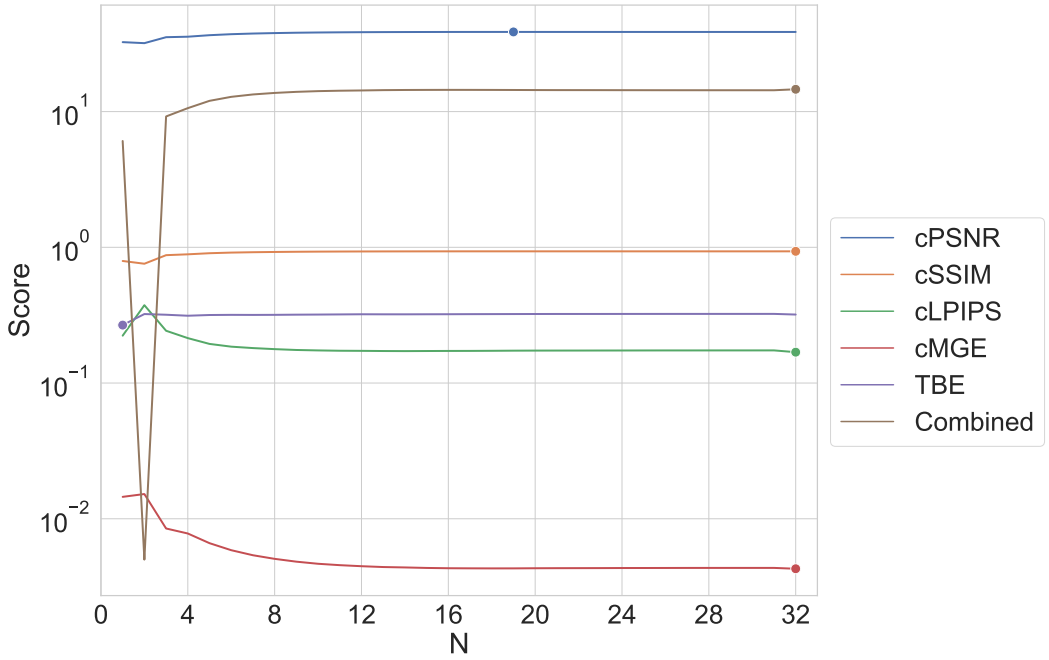


FIGURE 6.16: Performance trends of the MagNAt$_{\text{lead}}$ model across different metrics as the number of input images increases, emphasizing the impact of the leading LR strategy.

To assess the potency of this strategy, MagNAt$_{\text{lead}}$ was trained and evaluated on the Proba-V NIR subset using the leading LR image paradigm, maintaining all the hyperparameters originally set for the MagNAt model. Figure 6.16 provides a line plot that visually charts the performance of the model across varied metrics as the number of input images (N) increases. Intriguingly, for the MagNAt$_{\text{lead}}$, the pinnacle of performance remarkably manifests at N=32, a stark deviation from the traditional MagNAt model, which peaked at N=16 and thereafter witnessed a decline in output quality. Furthermore, the MagNAt$_{\text{lead}}$ exhibits a notable enhancement in results for N<5, a domain where MagNAt's results were unsatisfactory. This underscores the transformative potential and adaptability of the leading LR approach in managing disparate input conditions.

Delving deeper into the performance metrics of the MagNAt$_{\text{lead}}$ model across the test subset of Proba-V NIR, the mean scores and their changes with respect to the MagNAt model were observed to be:

- cPSNR: 38.620 (+2.451)
- cSSIM: 0.9342 (+0.0181)
- cLPIPS: 0.1688 (-0.0089)

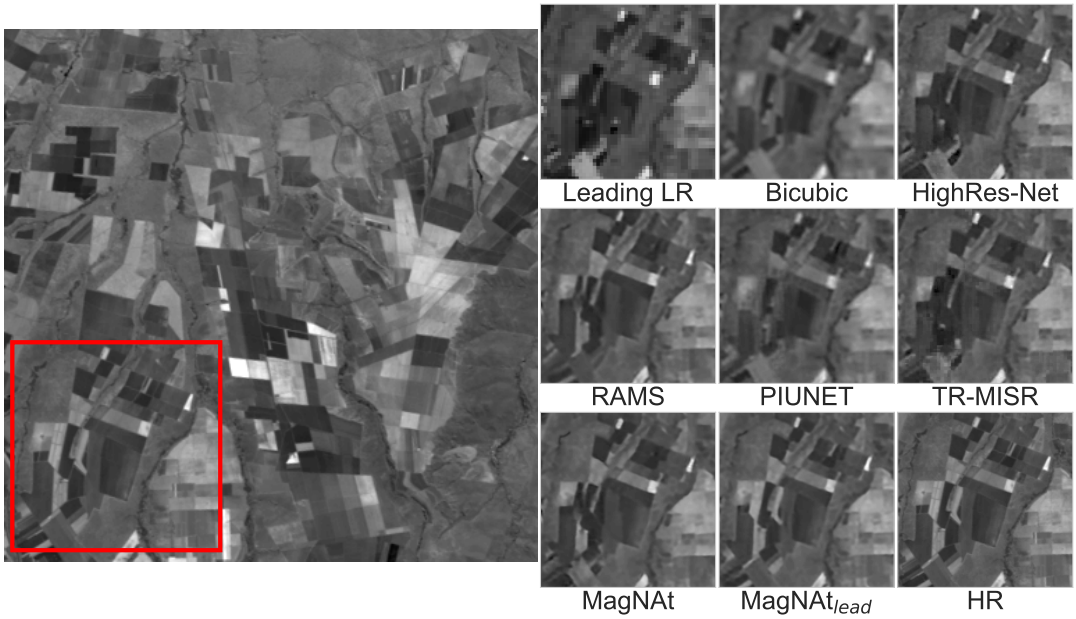
- cMGE: 0.0043 (-0.0016)
- TBE: 0.2673 (-0.0607)

Upon further analysis, the $\text{MagNAt}_{\text{lead}}$ model exhibits marked improvements across all examined metrics on the Proba-V NIR data when compared to the MagNAt model. This includes a noticeable enhancement in cPSNR and cSSIM, as well as a commendable reduction in values for cLPIPS, cMGE, and TBE, indicating a robust performance of the $\text{MagNAt}_{\text{lead}}$ model in addressing the MISR challenge.

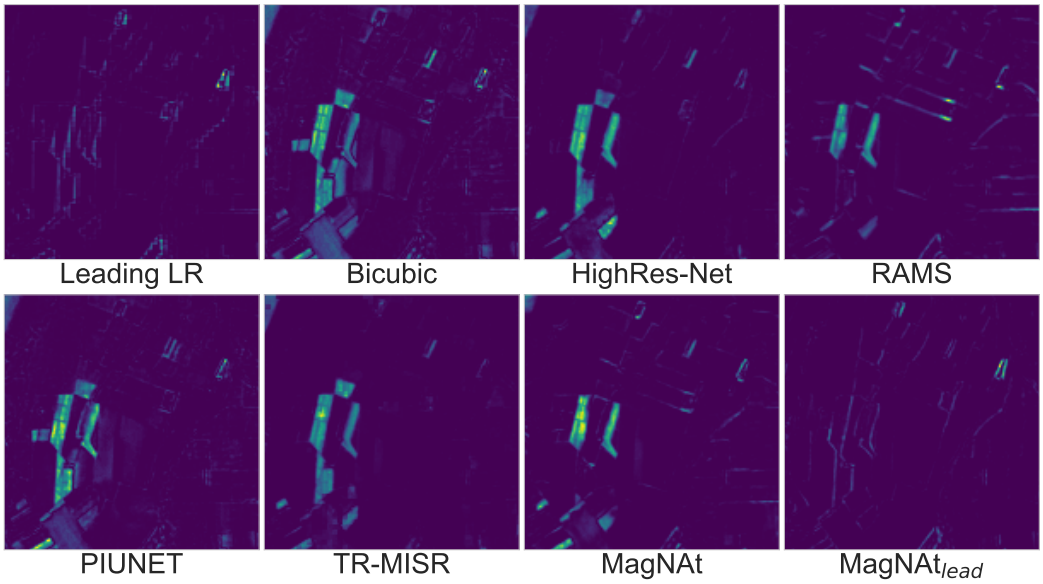
Visual Assessment of Temporal Consistency

For a clearer understanding of the $\text{MagNAt}_{\text{lead}}$ model's performance, especially in terms of image fidelity and temporal accuracy, a direct visual comparison is presented. Figure 6.17 offers a side-by-side view of super-resolution outputs from the tested models, emphasizing the distinct advantages of the leading LR approach. The figure also includes a visual display of difference maps, providing a detailed examination of the errors in reconstructing regions of high temporal variance by different methods with respect to the HR image reference. The showcased scene, representing a cultivated area filled with diverse fields, highlights the pronounced temporal variance inherent in the Proba-V dataset.

Upon examining the super-resolved outputs from various methods, it becomes clear that models like RAMS, HighRes-Net, and MagNAt produced detailed images, yet there are still obvious content differences in their outputs. Additionally, PIUNET's results appear blurred, challenging the clear identification of individual fields. TR-MISR continues to display its characteristic behaviour of introducing artefacts that resemble sections of the LR image that seem merely enlarged without any true enhancement. On the other hand, the output of the $\text{MagNAt}_{\text{lead}}$ model excels in terms of fidelity. The super-resolved image aligns accurately with each cultivated field and maintains its brightness and contrast in harmony with both the leading LR and the HR image. This precise alignment with the scene's temporal and structural characteristics underscores the effectiveness of the leading LR image approach. The provided error maps show well the areas each model failed to reconstruct properly when compared to the leading LR and HR images. From these maps, it is evident that $\text{MagNAt}_{\text{lead}}$ produced the least noticeable errors while maintaining temporal consistency with respect to the leading and HR images, showcasing how $\text{MagNAt}_{\text{lead}}$ skillfully utilizes this



(A) Super-resolved images produced by different methods with inputs of high temporal variance.



(B) Maps depicting a squared error of each pixel between the super-resolved images and the HR reference.

FIGURE 6.17: Visual comparison of super-resolution outputs from different models (A) and their corresponding difference maps with respect to HR reference (B).

approach to produce super-resolved images that authentically capture the original scene's temporal context.

Evaluating Temporal Reconstruction with Varied Leading Images

To further inspect the $\text{MagNAt}_{\text{lead}}$ model's performance, the effect of different leading LR images on reconstructing a specific scene was examined. This approach aims to verify whether $\text{MagNAt}_{\text{lead}}$ can accurately reproduce a scene at different time points dictated by the chosen leading LR images. Figure 6.18 demonstrates the reconstruction results for the same scene using different leading LR images, with one of them (d) taken at the same time as the HR image. As expected, the best quantitative score was obtained for the leading LR image captured concurrently with the HR image. A close observation shows that characteristic features from the leading images, such as uncaptured pixels or clouds, are preserved in the output. Notably, these features are present only in the corresponding leading image and not in the other LR inputs, indicating the model's ability to retain distinct temporal information from the leading image.

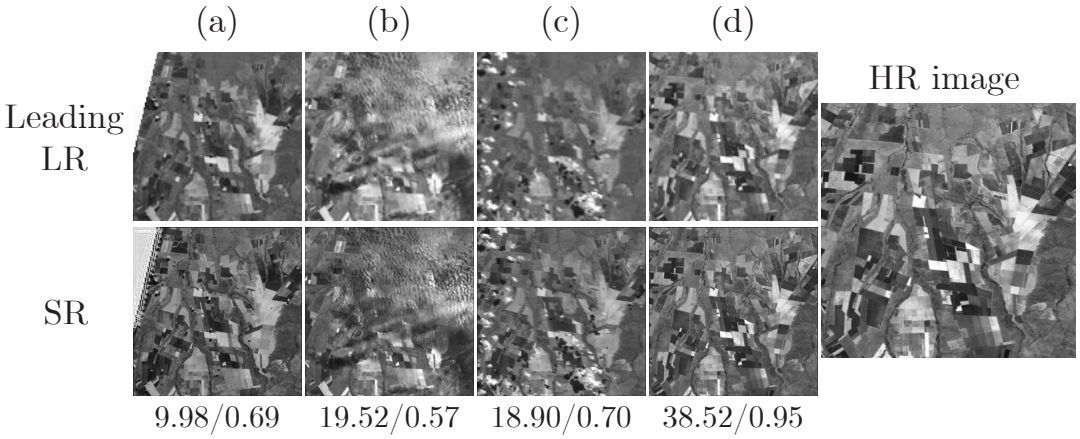


FIGURE 6.18: Reconstruction results from $\text{MagNAt}_{\text{lead}}$ for the same scene using different leading LR images, with one of them (d) captured at the same time as the HR image. Corresponding cPSNR and cSSIM scores are provided below each super-resolved image.

The observations from this segment of the analysis support a part of the third thesis of the dissertation, indicating that by choosing a specific reference image from the input LR image set, GNNs can effectively reconstruct a scene from a particular point in time. This is illustrated by the visual fidelity and temporal accuracy of the super-resolved outputs. On the other hand, the results from Figure 6.16 provide evidence for another segment of the third

thesis, showcasing that other images in the leading LR approach indeed serve as supplementary information sources which contribute to enhanced super-resolution accuracy.

6.4 Comparative Analysis of Architectural Progression

This section outlines the evolution of models developed to substantiate the second thesis of this dissertation: GNNs can improve their MISR performance by integrating techniques inspired by existing state-of-the-art MISR models based on CNNs, such as individual feature extraction for each LR image, the employment of attention mechanisms, and dynamic and trainable input registration. The following list provides a brief overview of the models created in this pursuit, discussed in detail in Chapter 3, each representing a distinct approach or enhancement towards leveraging the aforementioned techniques for MISR:

1. **MagNet**: The pioneering model in this series, MagNet, was the first reported attempt at employing GNNs for MISR. By integrating the power of graph-based data representation with the FSRCNN architecture, MagNet successfully transforms the MISR task into a graph processing challenge.

2. **MagNet++**: Building on MagNet, MagNet++ aims to enhance the upscaling operation. It introduces a new architectural choice that optimises the use of input nodes, minimizing information loss during upsampling by performing convolutional operations on bipartite graphs without the need for max-pooling.

3. **MagNet$_{\mathbf{enc}}$**: This model focused on refining the embedding phase. By treating each LR image individually during feature extraction, it aims to ensure that every input image, irrespective of its inherent characteristics, contributes equally to the super-resolved output.

4. **MagNAt**: This model introduced a multi-level attention mechanism alongside dynamic registration in the context of GNNs for MISR. This combination allows for the evaluation of the relative importance of each neighbouring node and the recalibration of positional shifts between LR images during each forward pass, aiming to optimize the model's adaptability and performance.

5. **MagNAt$_{\text{no_reg}}$**: As a modification of MagNAt, MagNAt$_{\text{no_reg}}$ retains the multi-level attention mechanism but omits the dynamic registration feature. This variation provides a controlled setup to assess the impact of the dynamic registration on the SRR performance of MagNAt.

While each of the above models brought forth some advancements for the GNNs in MISR, it is the MagNAt that stands as the most refined and comprehensive one. This section seeks to provide a comparative analysis of these models, highlighting their strengths, limitations, and incremental improvements leading up to the final MagNAt model.

6.4.1 Performance Analysis for Optimal Number of Inputs

An experiment was conducted on the validation subset of the Proba-V NIR dataset with each model delineated in the preceding discussion, with the intention of determining the optimal number of input images (N) that would enable a fair comparison of performance among the models. Utilizing a similar procedure of varying N as depicted in earlier analyses, this step was crucial to ensure that each model was evaluated at its peak performance for a more accurate comparative analysis.

In Figure 6.19, the performance trends of each model as a function of the number of input images, N, are showcased. An initial performance improvement with increased N, reaching an optimal point, followed by a decline and then a plateau at higher N values, is a shared trend observed across all the models for most metrics. This pattern, consistent across most metrics, elucidates the universal challenge of efficiently managing a larger number of input images. A clear advancement from MagNet to MagNAt, with each subsequent model exhibiting enhanced performance, is depicted. This progression underscores the iterative nature of model refinement and the effectiveness of the architectural modifications implemented at each step.

A notable observation is the TBE performance of MagNet for $N > 8$. Unlike its successors, MagNet produces markedly sharper results as N increases. This enhanced sharpness could potentially be attributed to the utilization of node max-pooling and pixel shuffle operations. By eschewing convolutions during internal supersampling, where spline-based convolutions typically introduce a slight blur, MagNet achieves a sharper output. However, while sharper, it is imperative to approach these findings with caution as TBE, a non-reference metric, assesses image sharpness without a reference image for comparison, implying that while the outputs might be sharper,

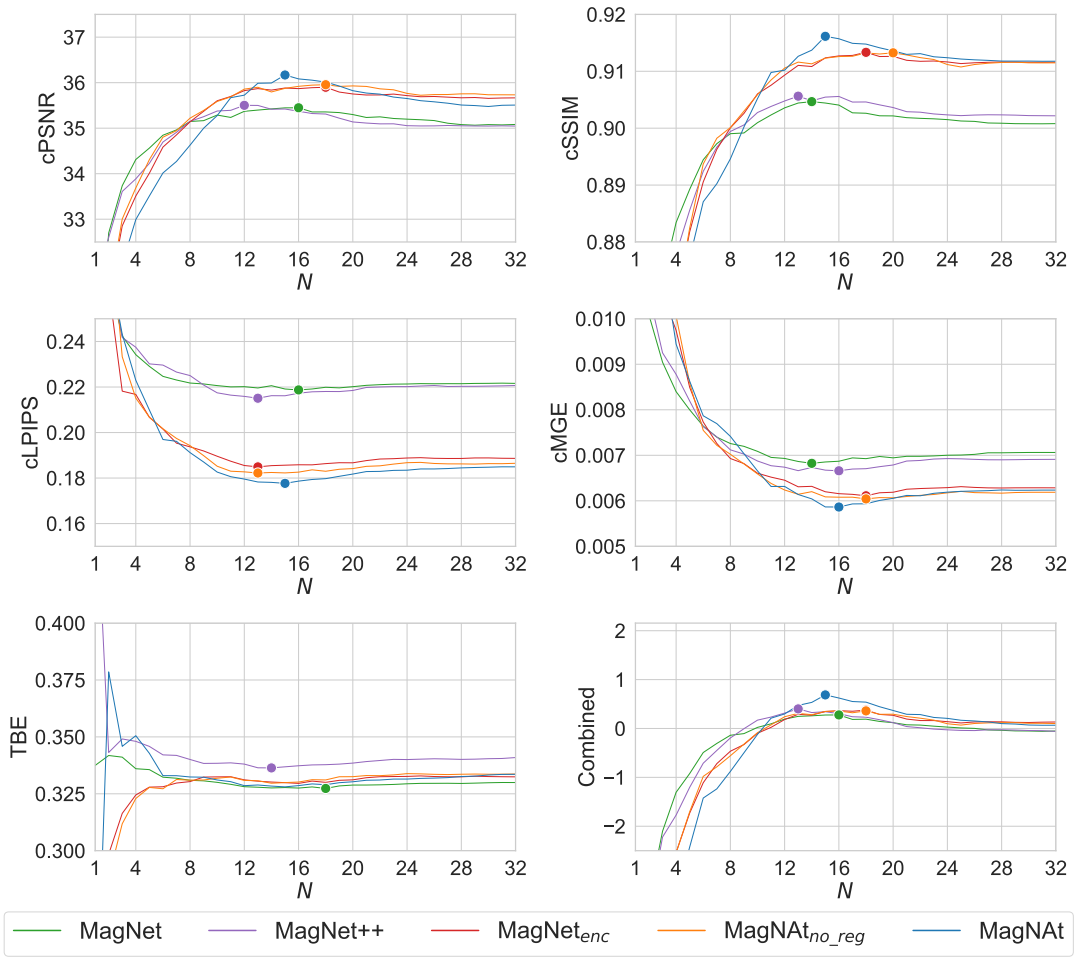


FIGURE 6.19: Performance metrics for the proposed models as a function of the number of input images (N). The graph delineates how each model's performance trends with an increase in N, pinpointing the optimal N for peak performance.

they may not necessarily offer accuracy or fidelity to the desired HR image, as suggested by other metrics, particularly cMGE.

The optimal number of input images, N, at which each model version reaches peak performance, is ascertained by examining the combined metric plot in Figure 6.19. The performance metrics calculated on the test subset for each model at these optimal N values are compiled in Table 6.6.

The transition from MagNet to MagNAt is evident in the results, with each additional component aiding in performance improvement. Notably, MagNAt consistently outperforms its predecessors across most indicators, marking a significant advancement. However, in the TBE metric, MagNet holds a slight lead, demonstrating its ability to yield sharper results, a trend also noticed on the validation subset. Yet, this sharpness does not always

TABLE 6.6: Mean metrics for each model at their optimal N.

Model	N	cPSNR	cSSIM	cLPIPS	cMGE	TBE
MagNet	16	35.448	.9041	.2187	.0069	**.3275**
MagNet++	13	35.500	.9056	.2151	.0067	.3364
$\text{MagNet}_{\text{enc}}$	18	35.900	.9133	.1863	.0061	.3300
$\text{MagNAt}_{\text{no\_reg}}$	18	35.955	.9132	.1830	.0060	.3311
MagNAt	15	**36.169**	**.9161**	**.1777**	**.0059**	.3280

translate to accurate super-resolution, as highlighted by other metrics. Moreover, a significant performance leap is observed with $\text{MagNet}_{\text{enc}}$, which is directly attributed to the incorporation of the encoding block. By treating each LR image individually during the feature extraction phase, this addition ensures a more nuanced representation of the input, facilitating a richer, more detailed super-resolved output. This observation is in line with insights gained during the development of the DeepSent model [126]. The comprehensive advancements and architectural choices in MagNAt affirm its standing as a promising model in MISR, portraying the incremental enhancements and the importance of well-informed design decisions. Furthermore, these findings substantiate the second thesis of this dissertation, that integrating techniques from existing state-of-the-art CNN-based MISR models can enhance GNNs' MISR performance.

6.5 Time and Memory Analysis

Computational efficiency is a paramount criterion when assessing the utility of super-resolution methods. This is especially crucial in situations requiring real-time or rapid image processing, which can be seen in applications such as medical diagnostics during emergencies, autonomous vehicle operations, or security monitoring in high-risk areas. Environments with limited computational resources further underscore the importance of efficiency. The outcomes across various super-resolution methods, given changes in the number of input LR images denoted as N, are depicted in Figure 6.20.

Bicubic interpolation, predictably, emerges as the fastest technique, reaching times from 0.01 ms (at N=1) to 0.3 ms (at N=32). Among the deep learning methodologies, RAMS stands out for its computational efficiency, followed closely by HighRes-Net. The latter's resilience to increasing N values is commendable, signifying its adaptability to a growing number of input images. TR-MISR, though following a similar performance trajectory, exhibits a

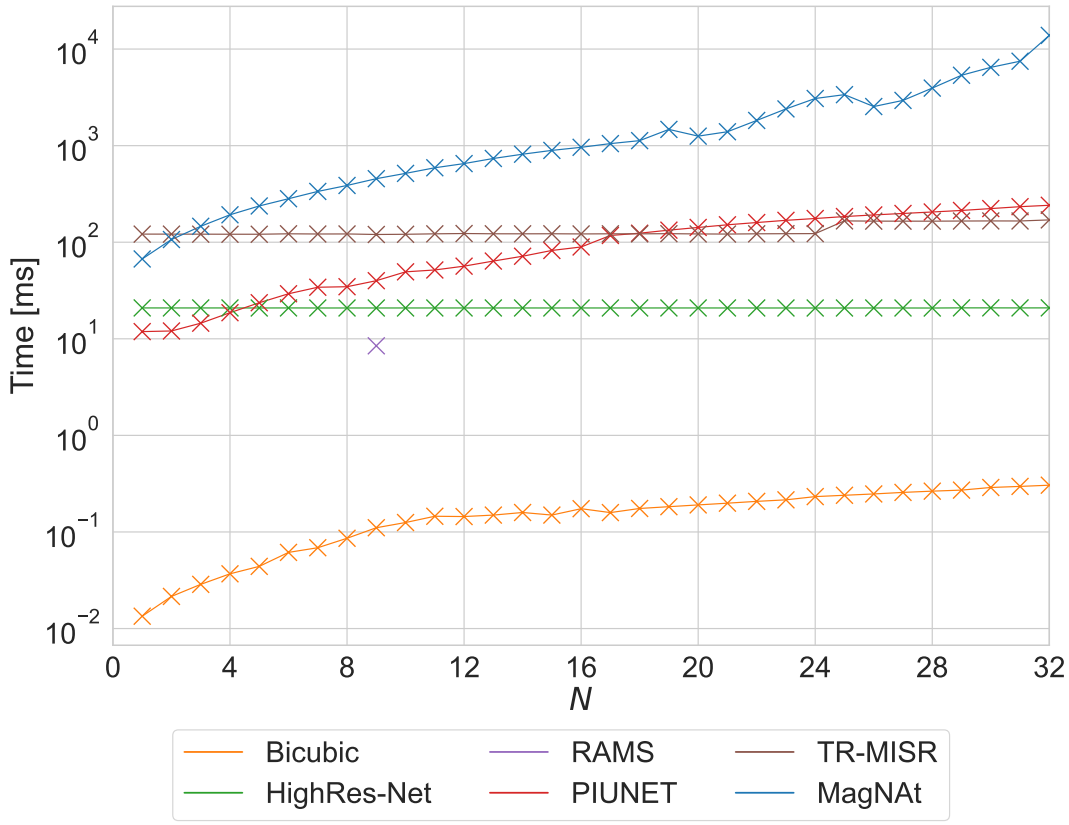


FIGURE 6.20: Computational times (in milliseconds) against the number of LR images (N). The y-axis is presented in a logarithmic scale.

slightly elongated computational time, especially for values of N exceeding 24.

PIUNET presents an interesting pattern, starting competitively for smaller N but demonstrating an exponential rise in computational time as the number of input images grows, making it one of the more time-intensive models for larger image sets.

The MagNAt model, being the most time-intensive among the tested methods, demands a closer inspection. Its performance, with respect to computational time, reveals a distinct exponential increase, more accentuated than that observed with PIUNET. An intriguing aspect of MagNAt's trend is its fluctuating time measurements for higher values of N. In certain instances, and rather counterintuitively, MagNAt achieved shorter computational times for larger N than for some smaller values. Such behaviour suggests the influence of underlying system operations, possibly related to memory management, allocation, and optimization routines. The substantial memory demands of MagNAt might trigger these system responses. Interestingly,

repeated runs of the experiment confirmed that this fluctuation in MagNAt's performance is not consistent across trials, further underscoring the potential impact of background memory-related processes on its performance.

This temporal inconsistency poses questions about MagNAt's predictability in real-world scenarios, while its pronounced time overhead also challenges its viability as a time-efficient solution in the SRR domain. A primary factor behind this behaviour is its graph-based data representation approach. Unlike models that rely on a direct stack of LR images, MagNAt thoroughly processes intricate inter-node relationships and their attributes. While this complexity provides it with the ability to capture richer information, it simultaneously worsens its memory requirements.

For a standard input stack of LR images with the shape $[1, 32, 128, 128]$, represented in a format $[batch, N, H, W]$, and stored as 32-bit floating numbers, the memory consumption is just 2MB. However, the graph-based representation of the same data in MagNAt necessitates considerably more memory, which can be attributed to its low-level components used to describe a graph. In addition to node features, it requires:

- The batch vector, indicating which pixels belong to specific examples in a batch, uses a 64-bit integer format, a requirement of the PyTorch Geometric library, to cater to the extensive number of nodes, thus utilizing 4MB of memory in this specific example.
- The edge indices, dictating node connections, are also maintained in a 64-bit integer format, occupying a significant 256MB.
- The edge attributes, which store the relative position between pairs of nodes and are integral for spline-based convolutions and attention mechanisms in MagNAt, are represented as 32-bit floating numbers, consuming 128MB.

Cumulatively, these components lead to a memory requirement of 392MB for MagNAt's graph-based data representation. This is approximately 200 times the memory footprint of the original input tensor, which is the only input for other tested models. Such an extensive memory allocation restricted the training of MagNAt to a maximum of N=15 images for the Proba-V dataset. It is worth recalling that other models were trained using their originally reported hyperparameters to ensure a balanced comparison, even though they are able to manage a full stack N=32 input images.

In summation, while the advanced graph-based approach of MagNAt offers the enticing prospect of capturing richer, more nuanced information, it

comes at the cost of increased computational time. The model's engagement with detailed inter-node relationships and the intricacies of their attributes necessitates deeper, more complex calculations. This is in stark contrast to more conventional models that process straightforward image data. Consequently, MagNAt's heightened memory requirements and the intricacies of its computational processes make it notably less time-efficient than its counterparts. The trade-off between richer information capture and computational efficiency is evident and remains a pivotal consideration in the broader context of super-resolution research.

Chapter 7

Summary and Conclusions

The primary aim of this dissertation was to explore the potential of GNNs in the realm of MISR. Initially, a thorough introduction to the domain of super-resolution was provided, distinguishing between its main sub-fields: SISR and MISR. The discussion on MISR covered its real-world applications and inherent challenges, with a particular focus on the temporal variations existing between different input images. A comprehensive literature review followed, evaluating various MISR and SISR methodologies, and exploring those which significantly impacted this research. The fundamental concepts of graphs and GNNs were explained next, highlighting specific models like GCN, GAT, and SplineCNN, which heavily influenced this research.

The methodological discussion began with a detailed description of representing multiple LR images as a graph, alongside discussing the rationale behind such representation. The conversation proceeded with the introduction of proposed models, initiating with MagNet and its subsequent enhancements, each one incorporating new elements aiming to increase their super-resolution potential. The motivation for utilizing simulated data to test the models in a controlled environment was articulated, followed by the curation of two simulated datasets, SRRB and SRRB_{enh}, each of different levels of difficulty. The choice of the Proba-V dataset, representing real-world data for MISR, was also justified.

The following chapter delineated the training methodology employed for the models and the image similarity metrics utilized for evaluating the proposed method against existing state-of-the-art MISR models. The experiments were presented on both simulated and real-world data, comparing the proposed method to established solutions in this domain. Detailed discussions on both quantitative and qualitative findings were provided, with qualitative analyses being supported by statistical significance testing of the model against its competitors.

The challenge posed by temporal variability in input data was also discussed, alongside a solution to mitigate it by guiding the reconstruction process to focus on a specific timeframe dictated by the leading input image. Further, experiments were conducted to compare the proposed models, and examine how the modifications introduced by each of them affected their super-resolution performance. Lastly, a time and memory analysis of the MagNAt model was performed, marking the culmination of the discussions on the methodologies and experiments conducted in this dissertation.

7.1 Discussion on Theses

This section explores the theses proposed at the onset of this dissertation, validating them with the empirical evidence gathered from the conducted experiments.

- **Thesis 1:** The primary thesis proposed that by representing a set of LR images with sub-pixel shifts as a graph, GNNs are capable of processing this graph to yield super-resolution results that are comparable or superior to those achieved by leading MISR architectures based on convolutional networks. This thesis was strongly substantiated both quantitatively and qualitatively on simulated and real-world datasets. In the simulated datasets, MagNAt consistently outperformed its state-of-the-art counterparts across most metrics, a claim further supported by statistically significant testing of these results. On the real-world Proba-V dataset, MagNAt achieved better or comparable results in most metrics and in terms of visuals. Furthermore, it was also most robust against temporal variations between input images, which are the main cause for the visual artefacts produced by other models. This comprehensive evaluation supports the viability and potential superiority of GNNs for MISR tasks when employing a graph-based representation of LR images with sub-pixel shifts.

- **Thesis 2:** The second thesis stated that the performance of GNNs in MISR can be elevated by assimilating techniques inspired by existing state-of-the-art MISR models based on CNNs. These techniques encompass individual feature extraction for each LR image, the application of attention mechanisms, and dynamic and trainable input registration. Through a comparative analysis amongst all the proposed

models—each model incorporating at least one new component derived from state-of-the-art CNN-based MISR models—a clear performance enhancement trajectory was established. It was observed that while independent feature extraction for each LR image emerged as the most impactful methodology, other incorporated techniques also significantly contributed to the increase of the models' performance, what quantitatively substantiated this thesis.

- **Thesis 3:** The third thesis proposed that GNNs can reconstruct a scene from a specific point in time by designating a particular reference image from the input LR image set, with the remaining images serving as supplementary information sources to improve super-resolution accuracy. This methodology was anticipated to mitigate visual inconsistencies in regions of high temporal variability and produce a temporally consistent image. MagNAt$_{\text{lead}}$ showcased an experiment where a specific input image was identified as the leading image, guiding the model to reconstruct the scene at that specific point in time. Quantitative analysis revealed that the additional LR images indeed acted as an auxiliary source of information, with MagNAt$_{\text{lead}}$ demonstrating optimal performance with the maximum available number of input images. The qualitative analysis further supported this thesis, revealing a consistent reconstruction of time-specific features present exclusively in each specific leading image.

This systematic discussion manifests the substantial fulfilment of the proposed theses, thereby achieving the primary objectives outlined for this dissertation.

7.2 Future Work

7.2.1 Potential Enhancements

A crucial aspect of improving the proposed method revolves around its time and memory demands, as highlighted in the time and memory analysis section. It was observed that MagNAt is the slowest and the most memory-demanding model among the tested models, with the extensive number of individual connections between nodes being the primary cause. Tackling this challenge is essential to enhance the practicability and scalability of the

model for more comprehensive or real-time applications. The possible enhancements are:

- Refine a new procedure for establishing connections, which could reduce the number of graph edges. This modification could significantly reduce MagNAt's memory demands, leading to more efficient computations.
- There is room for optimization steps that could potentially augment the model's time efficiency. However, these require a thorough investigation to ensure that the model's super-resolution performance is not compromised. Exploring optimized graph construction methods or investigating more efficient GNN architectures may yield valuable insights into reducing the computational burden, without sacrificing the quality of super-resolution.
- Leveraging the advancements in hardware and parallel computing may provide means to accelerate the computations and manage memory usage more effectively. Utilizing distributed computing resources or tailored hardware accelerators for graph-based computations could significantly boost MagNAt's efficiency, making it a more feasible solution for a wider array of MISR scenarios.

7.2.2 Prospective Research

The primary trajectory of the subsequent research revolves around adapting the $\text{MagNAt}_{\text{lead}}$ model for the multispectral MISR domain, specifically targeting the remotely sensed images from the Sentinel-2 satellite. This adaptation could require some conceptual changes in the model to efficiently handle the multispectral data. Additionally, the spectral images from Sentinel-2 vary in size, which presents a new challenge for MagNAt and the graph creation process.

Another promising research direction involves the refinement of the new node-connecting procedure discussed in the previous section. The aim is to reduce the number of edges, thereby decreasing the computational load of MagNAt. This refinement could be crucial, especially in the multispectral scenario, where the data naturally contains a significantly larger volume of information compared to the scenarios that MagNAt has been subjected to thus far. This dual approach not only provides opportunities for optimizing

the existing model but also extends the model's applicability to a broader range of MISR scenarios.

7.2.3 Interesting Avenues

This subsection sheds light on captivating directions for further exploration that could contribute to the refinement or broadening of the methodologies deployed in this research.

- **Handling Rotations and Shifts:** One engaging avenue for MagNAt lies in the ability to manage LR images that are not solely shifted, but also rotated concerning each other. This would demand modifications in the node positioning procedure to ensure the accurate placement of such images on a shared 2D plane. Investigating a procedure to properly register rotations along with translations between images could augment the versatility and applicability of MagNAt in dealing with more complex real-world MISR scenarios.

- **Non-Rigid Transformations:** Extending the scope to accommodate scenarios wherein LR images are influenced by non-rigid transformations presents another interesting avenue. This extension is particularly appealing for super-resolving raw satellite imagery that has not undergone the orthorectification process—a process dedicated to removing image distortions or displacements brought about by sensor tilt and terrain variations to ensure a geometrically correct image. Adapting MagNAt to handle such transformations could challenge the registration process significantly, necessitating careful and more nuanced alignment of each pixel in relation to other images, thus adding a new dimension of complexity and capability to the model.

- **Processing Spatially Irregular Data:** Stepping into the area of spatially irregular data processing, such as handling images with missing pixels or managing point clouds, reveals an entirely new set of challenges and opportunities. This direction would require a significant change in both the model architecture and the corresponding graph creation procedure. Exploring methods to adapt MagNAt to these unconventional data types could discover novel insights and extend the model's effectiveness across a broader spectrum of MISR scenarios.

Bibliography

[1] Eirikur Agustsson and Radu Timofte. "NTIRE 2017 challenge on single image super-resolution: Dataset and study". In: *Proc. IEEE CVPR Workshops*. 2017, pp. 126–135.

[2] Tai An et al. "TR-MISR: Multiimage Super-Resolution Based on Feature Fusion With Transformers". In: *IEEE J-STARS* 15 (2022), pp. 1373–1388.

[3] Pablo Arbelaez et al. "Contour detection and hierarchical image segmentation". In: *IEEE TPAMI* 33.5 (2010), pp. 898–916.

[4] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. *Neural Machine Translation by Jointly Learning to Align and Translate*. 2016. arXiv: 1409.0473 [cs.CL].

[5] Liqiang Bao et al. "Masked Graph Attention Network for Person Re-Identification". In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2019, pp. 1496–1505. DOI: 10.1109/CVPRW.2019.00191.

[6] Albert-Laszlo Barabasi and Zoltan Oltvai. "Network Biology: Understanding The Cell's Functional Organization". In: *Nature reviews. Genetics* 5 (Mar. 2004), pp. 101–13. DOI: 10.1038/nrg1272.

[7] Peter W. Battaglia et al. *Relational inductive biases, deep learning, and graph networks*. 2018. arXiv: 1806.01261 [cs.LG].

[8] Pawel Benecki et al. "Evaluating super-resolution reconstruction of satellite images". In: *Acta Astronautica* 153 (2018), pp. 15–25. ISSN: 0094-5765. DOI: https://doi.org/10.1016/j.actaastro.2018.07.035. URL: https://www.sciencedirect.com/science/article/pii/S0094576518300109.

[9] Marco Bevilacqua et al. "Low-Complexity Single-Image Super-Resolution based on Nonnegative Neighbor Embedding". In: *Proc. BMVC*. 2012.

[10] Goutam Bhat et al. "Deep burst super-resolution". In: *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021, pp. 9209–9218.

[11] Alan C. Bovik. *Handbook of Image and Video Processing (Communications, Networking and Multimedia)*. USA: Academic Press, Inc., 2005. ISBN: 0121197921.

[12] Leo Breiman. "Random Forests". English. In: *Machine Learning* 45.1 (2001), pp. 5–32. ISSN: 0885-6125. DOI: 10.1023/A:1010933404324. URL: http://dx.doi.org/10.1023/A%3A1010933404324.

[13] Sergey Brin and Lawrence Page. "The Anatomy of a Large-Scale Hypertextual Web Search Engine". In: *Computer Networks* 30 (1998), pp. 107–117. URL: http://www-db.stanford.edu/~backrub/google.html.

[14] Shaked Brody, Uri Alon, and Eran Yahav. *How Attentive are Graph Attention Networks?* 2022. arXiv: 2105.14491 [cs.LG].

[15] Michael M. Bronstein et al. "Geometric Deep Learning: Going beyond Euclidean data". In: *IEEE Signal Processing Magazine* 34.4 (July 2017), pp. 18–42. DOI: 10.1109/msp.2017.2693418. URL: https://doi.org/10.1109%2Fmsp.2017.2693418.

[16] Lisa Gottesfeld Brown. "A Survey of Image Registration Techniques". In: *ACM Comput. Surv.* 24.4 (Dec. 1992), pp. 325–376. ISSN: 0360-0300. DOI: 10.1145/146370.146374. URL: https://doi.org/10.1145/146370.146374.

[17] Benoit Brummer and Christophe De Vleeschouwer. "On the Importance of Denoising when Learning to Compress Images". In: *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, Jan. 2023. DOI: 10.1109/wacv56688.2023.00247. URL: https://doi.org/10.1109%2Fwacv56688.2023.00247.

[18] Joan Bruna et al. "Spectral networks and locally connected networks on graphs". In: *International conference on learning representations*. 2014.

[19] Chi Chen et al. "A Review of Hyperspectral Image Super-Resolution Based on Deep Learning". In: *Remote Sensing* 15.11 (2023). ISSN: 2072-4292. DOI: 10.3390/rs15112853. URL: https://www.mdpi.com/2072-4292/15/11/2853.

[20] Frédérique Crété-Roffet et al. "The Blur Effect: Perception and Estimation with a New No-Reference Perceptual Blur Metric". In: *Human Vision and Electronic Imaging* 12 (Mar. 2007). DOI: 10.1117/12.702790.

[21] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. "Deep Image Homography Estimation". In: *CoRR* abs/1606.03798 (2016). arXiv: 1606.03798. URL: http://arxiv.org/abs/1606.03798.

[22] Michel Deudon, Alfredo Kalaitzis, et al. "HighRes-net: Recursive Fusion for Multi-Frame Super-Resolution of Satellite Imagery". In: *arXiv preprint arXiv:2002.06460* (2020).

[23] Kunio Doi. "Computer-aided diagnosis in medical imaging: Historical review, current status and future potential". In: *Computerized Medical Imaging and Graphics* 31.4 (2007). Computer-aided Diagnosis (CAD) and Image-guided Decision Support, pp. 198–211. ISSN: 0895-6111. DOI: https://doi.org/10.1016/j.compmedimag.2007.02.002. URL: https://www.sciencedirect.com/science/article/pii/S0895611107000262.

[24] Chao Dong, Chen Change Loy, and Xiaoou Tang. "Accelerating the Super-Resolution Convolutional Neural Network". In: *CoRR* abs/1608.00367 (2016). arXiv: 1608.00367. URL: http://arxiv.org/abs/1608.00367.

[25] Chao Dong, Chen Change Loy, and Xiaoou Tang. "Accelerating the super-resolution convolutional neural network". In: *Proc. ECCV*. 2016, pp. 391–407.

[26] Chao Dong et al. *Image Super-Resolution Using Deep Convolutional Networks*. 2015. arXiv: 1501.00092 [cs.CV].

[27] Chao Dong et al. "Image super-resolution using deep convolutional networks". In: *IEEE TPAMI* 38.2 (2016), pp. 295–307.

[28] Chao Dong et al. "Learning a deep convolutional network for image super-resolution". In: *Proc. ECCV*. Springer. 2014, pp. 184–199.

[29] Vincent Dumoulin and Francesco Visin. *A guide to convolution arithmetic for deep learning*. 2018. arXiv: 1603.07285 [stat.ML].

[30] Michael Elad. "A Fast Super-Resolution Reconstruction Algorithm for Pure Translation Motion and Common Space-Invariant Blur". In: (Apr. 2002).

[31] M. Elgohary et al. "A Proposed Video Super-Resolution Strategy using Wavelet Multi-Scale Convolutional Neural Networks". In: *MEJ-Mansoura Engineering Journal* 47.4 (2022), pp. 1–10. ISSN: 1110-0923. DOI: 10.21608/bfemu.2022.258300. eprint: https://bfemu.journals.ekb.eg/article_258300_aa1c9f4d2853301b44bdd9bbf79e6a58.pdf. URL: https://bfemu.journals.ekb.eg/article_258300.html.

[32] Sina Farsiu et al. "Fast and robust multiframe super resolution". In: *IEEE Trans. on Image Process.* 13.10 (2004), pp. 1327–1344.

[33] Matthias Fey and Jan E. Lenssen. "Fast Graph Representation Learning with PyTorch Geometric". In: *ICLR Workshop on Representation Learning on Graphs and Manifolds*. 2019.

[34] Matthias Fey et al. "SplineCNN: Fast geometric deep learning with continuous B-spline kernels". In: *Proc. IEEE CVPR*. 2018, pp. 869–877.

[35] Keith Fife, Abbas El Gamal, and H.-S. Philip Wong. "A 3MPixel Multi-Aperture Image Sensor with 0.7m Pixels in 0.11m CMOS". In: *2008 IEEE International Solid-State Circuits Conference - Digest of Technical Papers*. 2008, pp. 48–594. DOI: 10.1109/ISSCC.2008.4523050.

[36] Justin Gilmer et al. *Neural Message Passing for Quantum Chemistry*. 2017. arXiv: 1704.01212 [cs.LG].

[37] Arushi Goel, Keng Teck Ma, and Cheston Tan. *An End-to-End Network for Generating Social Relationship Graphs*. 2019. arXiv: 1903.09784 [cs.CV].

[38] Rafael C. Gonzalez and Richard E. Woods. *Digital image processing*. Upper Saddle River, N.J.: Prentice Hall, 2008. ISBN: 9780131687288 013168728X 9780135052679 013505267X. URL: http://www.amazon.com/Digital-Image-Processing-3rd-Edition/dp/013168728X.

[39] Ian J. Goodfellow et al. *Generative Adversarial Networks*. 2014. arXiv: 1406.2661 [stat.ML].

[40] Aditya Grover and Jure Leskovec. *node2vec: Scalable Feature Learning for Networks*. 2016. arXiv: 1607.00653 [cs.SI].

[41] Shuhang Gu et al. "Convolutional Sparse Coding for Image Super-Resolution". In: *2015 IEEE International Conference on Computer Vision (ICCV)*. 2015, pp. 1823–1831. DOI: 10.1109/ICCV.2015.212.

[42] Manuel Guizar-Sicairos, Samuel T. Thurman, and James R. Fienup. "Efficient subpixel image registration algorithms". In: *Opt. Lett.* 33.2 (Jan. 2008), pp. 156–158.

[43] Rui GUO et al. "Super-resolution reconstruction of astronomical images using time-scale adaptive normalized convolution". In: *Chinese Journal of Aeronautics* 31.8 (2018), pp. 1752–1763. ISSN: 1000-9361. DOI: https://doi.org/10.1016/j.cja.2018.06.002. URL: https://www.sciencedirect.com/science/article/pii/S1000936118301882.

[44] Rohit Gupta, Anurag Sharma, and Anupam Kumar. "Super-Resolution using GANs for Medical Imaging". In: *Procedia Computer Science* 173 (2020). International Conference on Smart Sustainable Intelligent Computing and Applications under ICITETM2020, pp. 28–35. ISSN: 1877-0509. DOI: https://doi.org/10.1016/j.procs.2020.06.005. URL: https://www.sciencedirect.com/science/article/pii/S1877050920315076.

[45] William L. Hamilton, Rex Ying, and Jure Leskovec. *Inductive Representation Learning on Large Graphs*. 2018. arXiv: 1706.02216 [cs.SI].

[46] William L. Hamilton, Rex Ying, and Jure Leskovec. *Representation Learning on Graphs: Methods and Applications*. 2018. arXiv: 1709.05584 [cs.SI].

[47] F. Harary. *Graph Theory*. Reading, MA: Addison-Wesley, 1969.

[48] Kaiming He et al. *Deep Residual Learning for Image Recognition*. 2015. arXiv: 1512.03385 [cs.CV].

[49] Kaiming He et al. "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification". In: *CoRR* abs/1502.01852 (2015). arXiv: 1502.01852. URL: http://arxiv.org/abs/1502.01852.

[50] Zhi He et al. "Multiframe Video Satellite Image Super-Resolution via Attention-Based Residual Learning". In: *IEEE Transactions on Geoscience and Remote Sensing* 60 (2022), pp. 1–15. DOI: 10.1109/TGRS.2021.3072381.

[51] Mohammad Hijji et al. "Intelligent Image Super-Resolution for Vehicle License Plate in Surveillance Applications". In: *Mathematics* 11.4 (2023). ISSN: 2227-7390. DOI: 10.3390/math11040892. URL: https://www.mdpi.com/2227-7390/11/4/892.

[52] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. "Single Image Super-resolution from Transformed Self-Exemplars". In: *Proc. IEEE CVPR*. 2015, pp. 5197–5206.

[53] Jun-Jie Huang and Wan-Chi Siu. "Practical application of random forests for super-resolution imaging". In: *2015 IEEE International Symposium on Circuits and Systems (ISCAS)*. 2015, pp. 2161–2164. DOI: 10.1109/ISCAS.2015.7169108.

[54] Wei Huang et al. "A new pan-sharpening method with deep neural networks". In: *IEEE Geoscience and Remote Sensing Letters* 12.5 (2015), pp. 1037–1041.

[55] J.R. Jensen. *Remote Sensing of the Environment: An Earth Resource Perspective*. Prentice Hall series in geographic information science. Pearson Prentice Hall, 2007. ISBN: 9780131889507. URL: https://books.google.pl/books?id=A6YsAQAAMAAJ.

[56] Junjun Jiang et al. "Noise Robust Face Image Super-Resolution Through Smooth Sparse Representation". In: *IEEE Transactions on Cybernetics* 47.11 (2017), pp. 3991–4002. DOI: 10.1109/TCYB.2016.2594184.

[57] N. Kanopoulos, N. Vasanthavada, and R.L. Baker. "Design of an image edge detection filter using the Sobel operator". In: *IEEE Journal of Solid-State Circuits* 23.2 (1988), pp. 358–367. DOI: 10.1109/4.996.

[58] Armin Kappeler et al. "Video super-resolution with convolutional neural networks". In: *IEEE TCI* 2.2 (2016), pp. 109–122.

[59] Jayashree Karmakar et al. "A Novel Super-Resolution Reconstruction from Multiple Frames via Sparse Representation". In: https://link.springer.com/chapter/10.1007/978-981-15-2854-5_4. Germany: Springer Singapore, Apr. 2, 2020, pp. 33–45. DOI: 10.1007/978-981-15-2854-5_4. URL: https://lens.org/048-290-652-634-19X.

[60] Rajandeep Kaur and. "A Review of Image Compression Techniques". In: *International Journal of Computer Applications* 142 (May 2016), pp. 8–11. DOI: 10.5120/ijca2016909658.

[61] Michal Kawulok et al. "Deep learning for multiple-image super-resolution". In: *IEEE GRSL* 17.6 (2020), pp. 1062–1066.

[62] Michal Kawulok et al. "Deep Learning for Multiple-Image Super-Resolution of Sentinel-2 Data". In: *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*. 2021, pp. 3885–3888. DOI: 10.1109/IGARSS47720.2021.9553243.

[63] Michal Kawulok et al. "Evolving Imaging Model for Super-resolution Reconstruction". In: *Proc GECCO*. Kyoto, Japan: ACM, 2018, pp. 284–285. ISBN: 978-1-4503-5764-7.

[64] R. Keys. "Cubic convolution interpolation for digital image processing". In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 29.6 (1981), pp. 1153–1160. DOI: 10.1109/TASSP.1981.1163711.

[65] Dong-Wook Kim et al. "Constrained adversarial loss for generative adversarial network-based faithful image restoration". In: *ETRI Journal* 41.4 (2019), pp. 415–425.

[66] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. "Accurate image super-resolution using very deep convolutional networks". In: *Proc. IEEE CVPR*. 2016, pp. 1646–1654.

[67] Soo Ye Kim et al. "Video super-resolution based on 3D-CNNs with consideration of scene change". In: *Proc. IEEE International Conference on Image Processing (ICIP)*. IEEE. 2019, pp. 2831–2835.

[68] Thomas N. Kipf and Max Welling. *Semi-Supervised Classification with Graph Convolutional Networks*. 2017. arXiv: 1609.02907 [cs.LG].

[69] Pawel Kowaleczko et al. "A Real-World Benchmark for Sentinel-2 Multi-Image Super-Resolution". In: *Scientific Data* 10.1 (Sept. 2023), p. 644. ISSN: 2052-4463. DOI: 10.1038/s41597-023-02538-9. URL: https://doi.org/10.1038/s41597-023-02538-9.

[70] Maciej Krzywda, Szymon Lukasik, and Amir H. Gandomi. "Graph Neural Networks in Computer Vision - Architectures, Datasets and Common Approaches". In: *2022 International Joint Conference on Neural Networks (IJCNN)*. IEEE, July 2022. DOI: 10.1109/ijcnn55064.2022.9892658. URL: https://doi.org/10.1109%2Fijcnn55064.2022.9892658.

[71] C.D. Kuglin. "The phase correlation image alignment method," in: *Proc.International Conference on Cybernetics Society* (1975), pp. 163–165.

[72] Wei-Sheng Lai et al. *Deep Laplacian Pyramid Networks for Fast and Accurate Super-Resolution*. 2017. arXiv: 1704.03915 [cs.CV].

[73] Wei-Sheng Lai et al. "Fast and accurate image super-resolution with deep Laplacian pyramid networks". In: *IEEE TPAMI* 41.11 (2018), pp. 2599–2613.

[74] Charis Lanaras et al. "Super-resolution of Sentinel-2 images: Learning a globally applicable deep neural network". In: *ISPRS Journal of Photogrammetry and Remote Sensing* 146 (2018), pp. 305–319.

[75] Y. Lecun et al. "Gradient-based learning applied to document recognition". In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324. DOI: 10.1109/5.726791.

[76] Christian Ledig, Lucas Theis, Ferenc Huszár, et al. "Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network." In: *Proc. CVPR*. Vol. 2. 3. 2017, p. 4.

[77] Qiang Li, Qi Wang, and Xuelong Li. "Exploring the relationship between 2D/3D convolution for hyperspectral image super-resolution". In: *IEEE Transactions on Geoscience and Remote Sensing* 59.10 (2021), pp. 8693–8703.

[78] Xiaofang Li et al. "Deep learning methods in real-time image super-resolution: a survey". In: *Journal of Real-Time Image Processing* 17 (Dec. 2020). DOI: 10.1007/s11554-019-00925-3.

[79] T.M. Lillesand. *Remote Sensing and Image Interpretation*. John Wiley, 2004. ISBN: 9780470088272. URL: https://books.google.pl/books?id=8fiIPwAACAAJ.

[80] Bee Lim et al. "Enhanced deep residual networks for single image super-resolution". In: *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 2017, pp. 136–144.

[81] Huan Ling et al. *Fast Interactive Object Annotation with Curve-GCN*. 2019. arXiv: 1903.06874 [cs.CV].

[82] Hongying Liu et al. "Video super-resolution based on deep learning: a comprehensive survey". In: *Artificial Intelligence Review* (2022), pp. 1–55.

[83] David G. Lowe. "Distinctive Image Features from Scale-Invariant Keypoints". In: *Int. J. Comput. Vision* 60.2 (Nov. 2004), pp. 91–110. ISSN: 0920-5691. DOI: 10.1023/B:VISI.0000029664.99615.94. URL: https://doi.org/10.1023/B:VISI.0000029664.99615.94.

[84] Zhengyang Lu and Ying Chen. *Dense U-net for super-resolution with shuffle pooling layer*. 2021. arXiv: 2011.05490 [eess.IV].

[85] Zhengyang Lu and Ying Chen. *Single Image Super Resolution based on a Modified U-net with Mixed Gradient Loss*. 2019. arXiv: 1911.09428 [eess.IV].

[86] Zhisheng Lu et al. "Transformer for single image super-resolution". In: *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 457–466.

[87] Ziyang Ma et al. "Handling motion blur in multi-frame super-resolution". In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015, pp. 5224–5232. DOI: 10.1109/CVPR.2015.7299159.

[88] Andrew L. Maas. "Rectifier Nonlinearities Improve Neural Network Acoustic Models". In: 2013.

[89] Robert Maier, Jörg Stückler, and Daniel Cremers. "Super-resolution Keyframe Fusion for 3D Modeling with High-Quality Textures". In: *2015 International Conference on 3D Vision*. 2015, pp. 536–544. DOI: 10.1109/3DV.2015.66.

[90] K. Malczewski and R. Stasiński. "Super Resolution for Multimedia, Image, and Video Processing Applications". In: *Recent Advances in Multimedia Signal Processing and Communications*. Ed. by Mislav Grgic, Kresimir Delac, and Mohammed Ghanbari. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 171–208. ISBN: 978-3-642-02900-4. DOI: 10.1007/978-3-642-02900-4_8. URL: https://doi.org/10.1007/978-3-642-02900-4_8.

[91] Marcus Märtens et al. "Super-resolution of PROBA-V images using convolutional neural networks". In: *Astrodynamics* 3.4 (2019), pp. 387–402.

[92] Yusuke Matsui et al. "Sketch-based manga retrieval using Manga109 dataset". In: *Multimedia Tools and Applications* 76.20 (2017), pp. 21811–21838.

[93] Henry Meißner, Michael Cramer, and Ralf Reulke. "Evaluation of Structures and Methods for Resolution Determination of Remote Sensing Sensors". In: (Jan. 2020). DOI: 10.13140/RG.2.2.14188.92805.

[94] Alan Mislove et al. "Measurement and Analysis of Online Social Networks". In: *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement*. IMC '07. San Diego, California, USA: Association for Computing Machinery, 2007, pp. 29–42. ISBN: 9781595939081. DOI: 10.1145/1298306.1298311. URL: https://doi.org/10.1145/1298306.1298311.

[95] Dr Mistry and Asim Banerjee. "Review: Image Registration". In: *International Journal of Graphics and Image Processing(ISSN 2249 – 5452)*, II (Feb. 2012), pp. 18–22.

[96] Andrea Bordone Molini et al. "DeepSUM: Deep neural network for super-resolution of unregistered multitemporal images". In: *IEEE TGRS* 58.5 (2020), pp. 3644–3656.

[97] Kamal Nasrollahi and Thomas B Moeslund. "Super-resolution: a comprehensive survey". In: *Machine vision and applications* 25.6 (2014), pp. 1423–1468.

[98] M. E. J. Newman. *Networks: an introduction*. Oxford; New York: Oxford University Press, 2010. ISBN: 9780199206650 0199206651. URL: http://www.amazon.com/Networks-An-Introduction-Mark-Newman/dp/0199206651/ref=sr_1_5?ie=UTF8&qid=1352896678&sr=8-5&keywords=complex+networks.

[99] Ngoc Long Nguyen et al. *Proba-V-ref: Repurposing the Proba-V challenge for reference-aware super resolution*. 2021. arXiv: 2101.10200 [cs.CV].

[100] Ngoc Long Nguyen et al. "Self-supervised multi-image super-resolution for push-frame satellite images". In: *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021, pp. 1121–1131.

[101] Bruno A. Olshausen and David J. Field. "Sparse coding with an overcomplete basis set: A strategy employed by V1?" In: *Vision Research* 37.23 (1997), pp. 3311–3325. ISSN: 0042-6989. DOI: https://doi.org/10.1016/S0042-6989(97)00169-7. URL: https://www.sciencedirect.com/science/article/pii/S0042698997001697.

[102] Zaid Omar and Tania Stathaki. "Image Fusion: An Overview". In: Jan. 2014, pp. 306–310. DOI: 10.1109/ISMS.2014.58.

[103] Ram Krishna Pandey et al. *Binary Document Image Super Resolution for Improved Readability and OCR Performance*. 2018. arXiv: 1812.02475 [cs.CV].

[104] Dyah R. Panuju, David J. Paull, and Amy L. Griffin. "Change Detection Techniques Based on Multispectral Images for Investigating Land Cover Dynamics". In: *Remote Sensing* 12.11 (2020). ISSN: 2072-4292. DOI: 10.3390/rs12111781. URL: https://www.mdpi.com/2072-4292/12/11/1781.

[105] Sung Cheol Park, Min Kyu Park, and Moon Gi Kang. "Super-resolution image reconstruction: a technical overview". In: *IEEE signal processing magazine* 20.3 (2003), pp. 21–36.

[106] Les Piegl and Wayne Tiller. *The NURBS Book*. second. New York, NY, USA: Springer-Verlag, 1996.

[107] Bartłomiej Pogodziński, Tomasz Tarasiewicz, and Michal Kawulok. "Transformer-based spectro-temporal fusion for Sentinel-2 super-resolution". In: *2023 30th International Conference on Systems, Signals and Image Processing (IWSSIP)*. 2023, pp. 1–5. DOI: 10.1109/IWSSIP58668.2023.10180305.

[108] Nikolay Ponomarenko et al. "Modified image visual quality metrics for contrast change and mean shift accounting". In: *2011 11th International Conference The Experience of Designing and Application of CAD Systems in Microelectronics (CADSM)*. 2011, pp. 305–311.

[109] Md Rifat Arefin et al. "Multi-Image Super-Resolution for Remote Sensing Using Deep Recurrent Networks". In: *Proc. IEEE CVPR Workshops*. 2020, pp. 206–207.

[110] David E. Rumelhart and James L. McClelland. "Learning Internal Representations by Error Propagation". In: *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations*. 1987, pp. 318–362.

[111] Francesco Salvetti et al. "Multi-Image Super Resolution of Remotely Sensed Images Using Residual Attention Deep Neural Networks". In: *Remote Sensing* 12.14 (2020), p. 2207.

[112] Franco Scarselli et al. "The Graph Neural Network Model". In: *IEEE Transactions on Neural Networks* 20.1 (2009), pp. 61–80. DOI: 10.1109/TNN.2008.2005605.

[113] Michael Schlichtkrull et al. *Modeling Relational Data with Graph Convolutional Networks*. 2017. arXiv: 1703.06103 [stat.ML].

[114] Huanfeng Shen et al. "Deep-Learning-Based Super-Resolution of Video Satellite Imagery by the Coupling of Multiframe and Single-Frame Models". In: *IEEE Transactions on Geoscience and Remote Sensing* 60 (2022), pp. 1–14. DOI: 10.1109/TGRS.2021.3121303.

[115] Wenzhe Shi et al. *Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network*. 2016. arXiv: 1609.05158 [cs.CV].

[116] Jarrett Ethan Singian et al. "Ghosting Effect Removal for Multi-Frame Super-Resolution on CCTV Videos with Moving Objects". In: *2022 the 5th International Conference on Machine Vision and Applications (ICMVA)*. 2022, pp. 43–49.

[117] Khushboo Singla, Rajoo Pandey, and Umesh Ghanekar. "A review on Single Image Super Resolution techniques using generative adversarial network". In: *Optik* 266 (2022), p. 169607. ISSN: 0030-4026. DOI: https://doi.org/10.1016/j.ijleo.2022.169607. URL: https://www.sciencedirect.com/science/article/pii/S0030402622009032.

[118] J.P. Snyder. *Flattening the Earth: Two Thousand Years of Map Projections*. University of Chicago Press, 1997. ISBN: 9780226767475. URL: https://books.google.pl/books?id=0UzjTJ4w9yEC.

[119] Kaicong Sun et al. *FL-MISR: Fast Large-Scale Multi-Image Super-Resolution for Computed Tomography Based on Multi-GPU Acceleration*. 2021. arXiv: 2108.04315 [eess.IV].

[120] Shenggui Tang et al. "Graph Neural Networks with Interlayer Feature Representation for Image Super-Resolution". In: *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*. WSDM '23. Singapore, Singapore: Association for Computing Machinery, 2023, pp. 652–660. ISBN: 9781450394079. DOI: 10.1145/3539597.3570436. URL: https://doi.org/10.1145/3539597.3570436.

[121] Yunqing Tang, Yin Xiang, and Guangfeng Chen. "A Nighttime and Daytime Single-Image Dehazing Method". In: *Applied Sciences* 13.1 (2023). ISSN: 2076-3417. DOI: 10.3390/app13010255. URL: https://www.mdpi.com/2076-3417/13/1/255.

[122] Tomasz Tarasiewicz and Michal Kawulok. "Graph-based representation for multi-image super-resolution". In: *13th IAPR-TC15 International Workshop on Graph-Based Representations in Pattern Recognition (GBR)*. in press.

[123] Tomasz Tarasiewicz, Jakub Nalepa, and Michal Kawulok. "A Graph Neural Network For Multiple-Image Super-Resolution". In: *Proc. IEEE ICIP*. 2021, pp. 1824–1828.

[124] Tomasz Tarasiewicz, Jakub Nalepa, and Michal Kawulok. "Semi-Simulated Training Data for Multi-Image Super-Resolution". In: *IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium*. 2022, pp. 481–484. DOI: 10.1109/IGARSS46834.2022.9884565.

[125] Tomasz Tarasiewicz et al. "Extracting High-Resolution Cultivated Land Maps from Sentinel-2 Image Series". In: *IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium*. 2022, pp. 175–178. DOI: 10.1109/IGARSS46834.2022.9883919.

[126] Tomasz Tarasiewicz et al. "Multitemporal and multispectral data fusion for super-resolution of Sentinel-2 images". In: *IEEE Transactions on Geoscience and Remote Sensing (TGRS)* (in press).

[127] P. Thevenaz, U.E. Ruttimann, and M. Unser. "A pyramid approach to subpixel registration based on intensity". In: *IEEE Transactions on Image Processing* 7.1 (1998), pp. 27–41. DOI: 10.1109/83.650848.

[128] Qi Tian and Michael N. Huhns. "Algorithms for subpixel registration". In: *Computer Vision, Graphics, and Image Processing* 35.2 (1986), pp. 220–233. ISSN: 0734-189X. DOI: https://doi.org/10.1016/0734-189X(86)90028-9. URL: https://www.sciencedirect.com/science/article/pii/0734189X86900289.

[129] Radu Timofte, Vincent De, and Luc Van Gool. "Anchored Neighborhood Regression for Fast Example-Based Super-Resolution". In: *2013 IEEE International Conference on Computer Vision*. 2013, pp. 1920–1927. DOI: 10.1109/ICCV.2013.241.

[130] Ms. Priyanka S. Tondewad and Ms. Manisha P. Dale. "Remote Sensing Image Registration Methodology: Review and Discussion". In: *Procedia Computer Science* 171 (2020). Third International Conference on Computing and Network Communications (CoCoNet'19), pp. 2390–2399. ISSN: 1877-0509. DOI: https://doi.org/10.1016/j.procs.2020.04.259. URL: https://www.sciencedirect.com/science/article/pii/S1877050920312515.

[131] H Topan et al. "Information content of optical satellite images for topographic mapping". In: *International Journal of Remote Sensing* 30.7 (2009), pp. 1819–1827.

[132] Roger Y. Tsai and Thomas S. Huang. "Multiframe image restoration and registration". In: 1984. URL: https://api.semanticscholar.org/CorpusID:59796060.

[133] Ken Turkowski. "Filters for common resampling tasks". In: *Graphics gems* (1990), pp. 147–165. URL: https://api.semanticscholar.org/CorpusID:117128712.

[134] Diego Valsesia and Enrico Magli. "Permutation Invariance and Uncertainty in Multitemporal Image Super-Resolution". In: *IEEE TGRS* 60 (2022), pp. 1–12.

[135] Patrick Vandewalle, Sabine Süsstrunk, and Martin Vetterli. "A frequency domain approach to registration of aliased images with application to super-resolution". In: *EURASIP journal on advances in signal processing* 2006 (2006), pp. 1–14.

[136] Ashish Vaswani et al. *Attention Is All You Need*. 2023. arXiv: 1706.03762 [cs.CL].

[137] Andrea Vedaldi and Karel Lenc. *MatConvNet - Convolutional Neural Networks for MATLAB*. 2016. arXiv: 1412.4564 [cs.CV].

[138] Petar Veličković et al. "Graph Attention Networks". In: *6th International Conference on Learning Representations* (2017).

[139] Peng Wang et al. *Neighbourhood Watch: Referring Expression Comprehension via Language-guided Graph Attention Networks*. 2018. arXiv: 1812.04794 [cs.CV].

[140] Xintao Wang et al. "ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks". In: *CoRR* abs/1809.00219 (2018). arXiv: 1809.00219. URL: http://arxiv.org/abs/1809.00219.

[141] Xuan Wang et al. "A Review of Image Super-Resolution Approaches Based on Deep Learning and Applications in Remote Sensing". In: *Remote Sensing* 14.21 (2022). ISSN: 2072-4292. DOI: 10.3390/rs14215423. URL: https://www.mdpi.com/2072-4292/14/21/5423.

[142] Zhihao Wang, Jian Chen, and Steven C. H. Hoi. "Deep Learning for Image Super-Resolution: A Survey". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43.10 (2021), pp. 3365–3387. DOI: 10.1109/TPAMI.2020.2982166.

[143] Zhongyuan Wang et al. "Ultra-dense GAN for satellite imagery super-resolution". In: *Neurocomputing* 398 (2020), pp. 328–337.

[144] Zhou Wang et al. "Image quality assessment: from error visibility to structural similarity". In: *IEEE Trans. on Image Processing* (2004), pp. 600–612.

[145] Zhou Wang et al. "Image quality assessment: from error visibility to structural similarity". In: *IEEE Transactions on Image Processing* 13.4 (2004), pp. 600–612. DOI: 10.1109/TIP.2003.819861.

[146] Stephen T. Welstead. "Fractal and Wavelet Image Compression Techniques". In: 1999. URL: https://api.semanticscholar.org/CorpusID:118961255.

[147] Douglas B. West. *Introduction to Graph Theory*. 2nd ed. Prentice Hall, Sept. 2000. ISBN: 0130144002.

[148] Frank Wilcoxon. "Individual Comparisons by Ranking Methods". In: *Biometrics Bulletin* 1.6 (1945), pp. 80–83. ISSN: 00994987.

[149] Bartlomiej Wronski et al. "Handheld multi-frame super-resolution". In: *ACM Transactions on Graphics* 38.4 (July 2019), pp. 1–18. DOI: 10.1145/3306346.3323024. URL: https://doi.org/10.1145%2F3306346.3323024.

[150] Bin Wu et al. "Cross-Scale Internal Graph Neural Network for Image Super-Resolution". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 15738–15747.

[151] Zonghan Wu et al. "A Comprehensive Survey on Graph Neural Networks". In: *IEEE Transactions on Neural Networks and Learning Systems* 32.1 (Jan. 2021), pp. 4–24. DOI: 10.1109/tnnls.2020.2978386. URL: https://doi.org/10.1109%2Ftnnls.2020.2978386.

[152] Liangbin Xie et al. *Mitigating Artifacts in Real-World Video Super-Resolution Models*. 2022. arXiv: 2212.07339 [cs.CV].

[153] Keyulu Xu et al. *How Powerful are Graph Neural Networks?* 2019. arXiv: 1810.00826 [cs.LG].

[154] Jize Xue et al. "When Laplacian Scale Mixture Meets Three-Layer Transform: A Parametric Tensor Sparsity for Tensor Completion". In: *IEEE Transactions on Cybernetics* 52.12 (2022), pp. 13887–13901. DOI: 10.1109/TCYB.2021.3140148.

[155] Sijie Yan, Yuanjun Xiong, and Dahua Lin. *Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition*. 2018. arXiv: 1801.07455 [cs.CV].

[156] Jianchao Yang et al. "Image Super-Resolution Via Sparse Representation". In: *IEEE Transactions on Image Processing* 19.11 (2010), pp. 2861–2873. DOI: 10.1109/TIP.2010.2050625.

[157] Rex Ying et al. "Graph Convolutional Neural Networks for Web-Scale Recommender Systems". In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery &amp Data Mining*. ACM, July 2018. DOI: 10.1145/3219819.3219890. URL: https://doi.org/10.1145%2F3219819.3219890.

[158] Linwei Yue et al. "Image super-resolution: The techniques, applications, and future". In: *Signal Processing* 128 (2016), pp. 389–408.

[159] Matthew D. Zeiler, Graham W. Taylor, and Rob Fergus. "Adaptive deconvolutional networks for mid and high level feature learning". In: *2011 International Conference on Computer Vision*. 2011, pp. 2018–2025. DOI: 10.1109/ICCV.2011.6126474.

[160] Matthew D. Zeiler et al. "Deconvolutional networks". In: *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 2010, pp. 2528–2535. DOI: 10.1109/CVPR.2010.5539957.

[161] Roman Zeyde, Michael Elad, and Matan Protter. "On Single Image Scale-Up Using Sparse-Representations". In: vol. 6920. June 2010, pp. 711–730. ISBN: 978-3-642-27412-1. DOI: 10.1007/978-3-642-27413-8_47.

[162] Roman Zeyde, Michael Elad, and Matan Protter. "On single image scale-up using sparse-representations". In: *Proc. Int. Conf. on Curves and Surfaces*. 2010, pp. 711–730.

[163] Mingliang Zhai et al. "Optical flow and scene flow estimation: A survey". In: *Pattern Recognition* 114 (2021), p. 107861. ISSN: 0031-3203. DOI: https://doi.org/10.1016/j.patcog.2021.107861. URL: https://www.sciencedirect.com/science/article/pii/S0031320321000480.

[164] Muhan Zhang et al. "An End-to-End Deep Learning Architecture for Graph Classification". In: *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*. Ed. by Sheila A. McIlraith and Kilian Q. Weinberger. AAAI Press, 2018, pp. 4438–4445. URL: https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17146.

[165] Richard Zhang et al. "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric". In: *Proc. IEEE/CVF CVPR*. 2018.

[166] Yulun Zhang et al. "Image Super-Resolution Using Very Deep Residual Channel Attention Networks". In: *Proc. ECCV*. Ed. by Vittorio Ferrari et al. Cham: Springer, 2018, pp. 294–310.

[167] WenYi Zhao and Harpreet S. Sawhney. "Is Super-Resolution with Optical Flow Feasible?" In: *Computer Vision — ECCV 2002*. Ed. by Anders Heyden et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2002, pp. 599–613. ISBN: 978-3-540-47969-7.

[168] Jie Zhou et al. *Graph Neural Networks: A Review of Methods and Applications*. 2021. arXiv: 1812.08434 [cs.LG].

[169] Barbara Zitova and Jan Flusser. "Image registration methods: a survey". In: *Image and vision computing* 21.11 (2003), pp. 977–1000.

[170] Assaf Zomet and Shmuel Peleg. “Super-Resolution from Multiple Images Having Arbitrary Mutual Motion”. In: *Super-Resolution Imaging*. Ed. by Subhasis Chaudhuri. Boston, MA: Springer US, 2000, pp. 195–209. ISBN: 978-0-306-47004-2. DOI: 10.1007/0-306-47004-7_8. URL: https://doi.org/10.1007/0-306-47004-7_8.

Acknowledgements

Firstly, to my family: Thank you for always believing I was the smart one, even when my experiments and research told a very different story. Your unwavering faith (or well-practised act of it) has been the wind beneath my sometimes very weary wings.

To my wife: Thank you for enduring those nights when you'd head to bed and I'd still be at my desk, lost in research until sunrise. For the times I mumbled about my findings in my sleep and for those moments when our cosy dinners at home turned into brainstorming sessions. Your patience might either be a testament to your boundless love or a remarkable ability to tune out my academic ramblings. Regardless, I am deeply grateful.

Lastly, and most crucially, to my supervisor: Without your guidance, this thesis might have turned into a novel, a cookbook, or an overly ambitious board game. Your ability to steer me back on course, with a mix of wisdom, just the right amount of sarcasm, and those unforgettable dad jokes, has been invaluable. I promise the number of times I considered turning my research into interpretative dance was minimal, thanks to you.

This work has been financially supported by the European Union through the European Social Fund (grant POWR.03.05.00-00-Z305) and the National Science Centre (grant 2019/35/B/ST6/03006).

List of Figures

List of Tables

List of Abbreviations

CNN	**C**onvolutional **N**eural **N**etwork
GAN	**G**enerative **A**dversarial **N**etwork
GAT	**G**raph **A**ttention **N**etwork
GCN	**G**raph **C**onvolutional **N**etwork
GDC	**G**lobal **D**ynamic **C**onvolution
GNN	**G**raph **N**eural **N**etwork
GSD	**G**round **S**ampling **D**istance
HR	**H**igh **R**esolution
LPIPS	**L**earned **P**erceptual **I**mage **P**atch **S**imilarity
LR	**L**ow **R**esolution
MGE	**M**ean **G**radient **E**rror
MISR	**M**ulti-**I**mage **S**uper-**R**esolution
MRI	**M**agnetic **R**esonance **I**maging
MSE	**M**ean **S**quared **E**rror
NIR	**N**ear **I**nfra**r**ed
PSNR	**P**eak **S**ignal-to-**N**oise **R**atio
RGN	**R**ecurrent **G**raph **N**etwork
RNN	**R**ecurrent **N**eural **N**etwork
SISR	**S**ingle-**I**mage **S**uper-**R**esolution
SSIM	**S**tructural **S**imilarity **I**ndex **M**easure
SRR	**S**uper-**R**esolution **R**econstruction
TBE	**T**he **B**lur **E**ffect

List of Symbols

Symbol	Description
N	Number of input images
W	Width of an image
H	Height of an image
F_{in}	Number of input channels
$\mathcal{S}$	Upsampling factor
F	Number of features
LR_i	i^{th} LR image
LR_{lead}	Leading LR image
$\mathcal{G}$	Graph
n	Number of nodes
$\mathcal{V}$	Set of vertices
v_i	i^{th} node of a graph
$\mathbf{H}$	Feature matrix
$\boldsymbol{h_i}$	Feature vector of node v_i
$\mathcal{E}$	Set of graph's edges
e_{ij}	Edge connecting node v_i with v_j
$\mathcal{U}$	Set of edge's attributes
$\mathbf{A}$	Adjacency matrix of a graph
$\boldsymbol{u_{ij}}$	Attributes of an edge e_{ij}
$\mathcal{N}(i)$	Neighbourhood of node v_i
$\vec{u}_i$	Shift vector of the i^{th} image
loc	Function returning a discrete position of a node
loc'	Function returning a shifted position of a node
x_i	Horizontal position of node v_i before shift
y_i	Vertical position of node v_i before shift
x'_i	Horizontal position of node v_i after shift
y'_i	Vertical position of node v_i after shift
d	Euclidian distance function
ϕ	Message function

γ	Update function
$\boldsymbol{m_i}$	Message vector collected from neighbours of node v_i
$\boldsymbol{h_i'}$	Output feature vector of node v_i
$\mathbf{W}$	Learnable weight matrix
$\mathbf{a}$	Learnable weight vector
$\tilde{d}_i$	Degree of node v_i
α_{ij}	Attention coefficient for edge e_{ij}
$\mathbf{k}$	Set defining a dimensionality of a spline-based kernel
μ	Mean
σ	Standard deviation
D	Number of dimensions of $\mathbf{k}$
$\bigoplus$	Differentiable and permutation invariant aggregation
$\|\|$	Concatenation
$\cdot^T$	Transposition