Chapter 2

Review of Literature

2.1 Introduction

The bottleneck for many remote sensing (RS) applications is due to the fact that satellite sensors cannot provide information at the actual spatial resolution of a scene. Therefore, increasing the spatial resolution of RS images using SR is one of the most cost-effective software-based solutions that can be implemented at reasonable ease. There are many fundamental assumptions to be taken into consideration, while formulating the SR problem. This results into different SR algorithms, and, therefore, different approaches have been reported in the literature. This chapter provides a detailed analysis of paradigm shifts across different SR algorithms in order to demonstrate their significance for producing visually pleasing and meaningful HR images. A detailed discussion on RS image SR using learning-based approaches is also provided, with an analysis of their merits and demerits.

The rest of the chapter is organized as follows: Section 2.2 discusses the categorization of SR methods. Section 2.3 explains learning-based SISR techniques and their related works on remote sensing images. Section 2.4 presents benchmarking criteria of remote sensing SISR algorithms. Section 2.5 explains the role of parallel computing in SR. Section 2.6 gives an overview of commonly used remote sensing datasets for SR. Section 2.7 discusses various metrics employed to evaluate the quality of SR remote sensing images. Section 2.8 concludes with a summary of key findings and research gaps, emphasizing the significance of ongoing exploration in remote sensing SISR.

2.2 Classification of SR methods

SR algorithms can be categorized according to various criteria, such as their operating domains, the number of LR images used, and the reconstruction approaches used. The detailed taxonomy of image SR is depicted in Fig. 2.1. Image SR approaches are broadly divided into two categories: *frequency or transform* and *spatial domain*. Although the initial developments of SR algorithms originated from signal processing techniques in the frequency domain, later on, most of the SR algorithms were formulated in the spatial domain. With regard to the number of LR images utilized, SR algorithms can be divided into two categories: SISR or MISR. We begin off by discussing domain-specific approaches as follows:

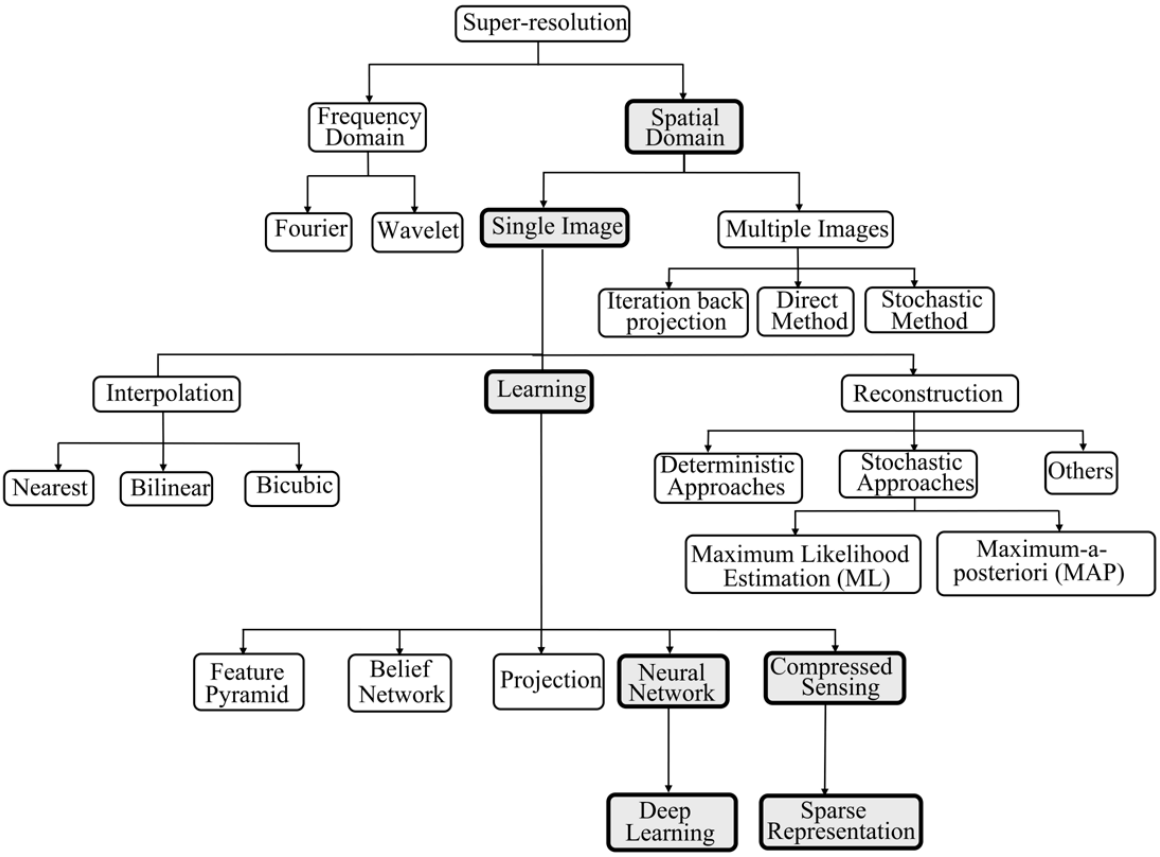

Figure 2.1: Categorization of SR techniques.

2.2.1 Frequency domain

In the frequency domain-based SR, input LR images are first transformed from the spatial domain to the frequency domain and then the HR image is estimated. After obtaining the HR reconstructed image, it is converted back to the spatial domain. The core of frequency-domain analysis is a non-linear mapping that smoothes the image by altering the frequency components (coefficients). Clearly, the method neglects the time-related information (in terms of information retrieval) during the inversion process [8]. These techniques are divided into two categories based on the transformation used: Fourier transform- and wavelet transform-based methods. SR algorithms were introduced by Gerchberg [32] and then improved by Santis and Gori [83]. These were iterative frequency-domain techniques based on the Fourier transform (FT) that could enhance resolution by expanding a signal's spectrum over the diffraction limit. Tsai and Huang [104] developed the first frequency-domain MISR algorithm. This technique compares the continuous FT of the original scene to the discrete FT of the observed LR images by measuring the sampling periods in the horizontal- and vertical directions of the digital image. Tom and Katsaggelos [100] presented a two-stage SR method; first, it processes the LR images by registering, deblurring, and de-noising them, and then project the processed images onto a target grid of HR image(s).

Wavelet transform-based approaches are alternatives to FT that are often used in frequency domain-based SR algorithms. In this approach, multi-scale representation of images is used for effective HR restoration. Nguyen and Milanfar [71] proposed an efficient method that used wavelet coefficients to represent the input LR images and then map the coefficients to that of the target SR image. Bose et al. [9] presented a second-generation wavelet-based approach that performs well in the presence of noise. However, the resultant SR image(s) have high-frequency artifacts that resemble a wavelet. It has been observed that many works are [15, 82, 136] based on the discrete wavelet transform (DWT). Frequency domain has some disadvantages for real-world applications, such as improper modeling of motion, sensitivity

to errors, inadequate prior models, and difficult mathematical formulation, etc. In order to overcome these constraints, researchers have shifted their attention towards spatial-domain SR methods.

2.2.2 Spatial domain

Spatial domain approaches address complex problems using simple pixel shift operations, overcoming the limitations of frequency domain SR methods. These techniques try to estimate how each LR pixel corresponds to its corresponding HR pixel by analysing the relationship between LR and HR images. This modelling enables direct estimate of HR pixel values from LR inputs. The spatial domain methods allows unconstrained motion between frames and makes it easier to incorporate prior knowledge. The spatial domain methods can be again classified into two parts: multiple image SR and single image SR.

2.2.2.1 Multiple image super-resolution

In multiple image SR (MISR), multiple LR images are considered to generate the HR image. Images are acquired using the same camera at various time intervals, or from various angles, or using different cameras at various locations [75]. Such images are helpful for estimating the motion in the imaging system, both controlled and uncontrolled. The objective of MISR is for predicting motion information that are subsequently employed for incorporating the sub-pixel shift of LR images onto an HR grid in order to perform SR reconstruction. Fig. 2.2 depicts the process of image registration and sub-pixel shifting in the MISR process. The artifacts resulting from aliasing, which exist in the observed LR images caused by the undersampling process can be eliminated with the help of MISR algorithms. Iterative back projection (IBP) [42] was one of the first techniques in MISR. Here, HR estimation is obtained iteratively, where the initial estimate for the HR image is started with $f^{(0)}$. The aim of this IBP technique is to improve the initial estimate $f^{(0)}$ by back-

projecting the difference between the simulated LR images $\{g_k^{(0)}\}$, $k = 1.....K$ and the observed LR images $\{g_k\}$ onto its receptive field in $f^{(0)}$. This projection is repeated iteratively for minimizing the following error function (e^n):

$$e^n = \sqrt{\frac{1}{K}\sum_{k=1}^{K}\left\|g_k - g_k^{(t)}\right\|_2^2} \tag{2.1}$$

where n is the total number of iterations, and t is the current iteration. The

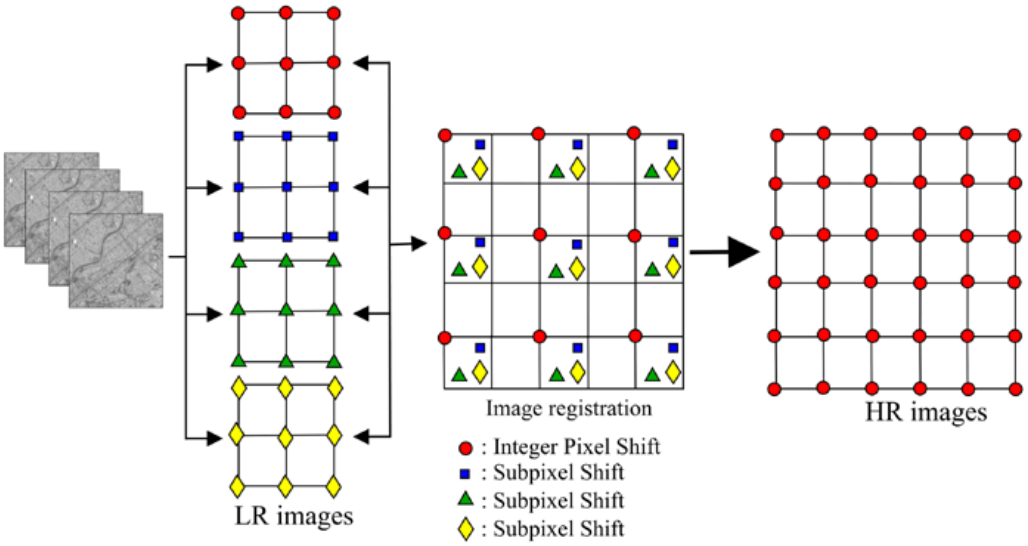

Figure 2.2: The process of registration and sub-pixel shifting in MISR.

limitation of IBP is that it is difficult to include a priori knowledge in their solution. Other methods of MISR are the direct methods [41], where HR image is generated through registration and warping of upscaled versions of LR images. These methods are faster than IBP techniques. Stochastic methods, such as maximum a posteriori (MAP) [36] and maximum likelihood (ML) [11] are some other MISR methods. They offer a powerful theoretical framework for a prior knowledge required for solving the ill-posed SR inverse problem. The ML problem may be transformed into a MAP problem by involving a priori information. In the case when the number of LR images over-determines the SR images (a large number of LR images would lead to a uniform contribution of each of them on the target HR image), the outputs of ML and MAP are similar, and ML is chosen over MAP due to its simple computation. If insufficient number of LR images are available to determine the SR image, a priori information becomes crucial, then MAP performs better than ML.

MISR approaches, however, are time consuming as they need a registration procedure involving sub-pixel alignment of multiple LR input images. The effectiveness of MISR is mostly affected by the registration error i.e. the estimation of motions between various observed LR images. Motion estimate is a very unstable and complicated procedure in MISR because in real-world applications, objects that are present in the same frame can have different motions and directions. In RS, acquiring multiple images of the same scene for MISR is also a difficult challenge due to cloud coverage, moving objects and other atmospheric disturbance, etc.

2.2.2.2 Single image super-resolution

In recent years, single image super-resolution, also known as SISR, has emerged as an important research topic for many computer vision applications. In particular, high-definition image processing requires the reconstruction of an image from a single observation. The SISR method utilizes mainly spatial information to recover the HR image from a single LR image, making it very suitable for RS applications. SISR algorithms are divided into three categories: interpolation-, reconstruction-, and learning-based methods, which are discussed as follows:

(i) **Interpolation-based:** Interpolation method is one of the simplest forms of super-resolution methods, which are typically used for the ease of zooming images. This method can produce an HR image from its LR images by estimating the pixel intensities on an up-sampled grid. The most widely used interpolation methods are the nearest neighbor, the bilinear, and the bicubic interpolation. These methods are based on the direct manipulation of the pixels only. The operation time of the interpolation is reduced because it relies on neighboring pixel values; however, the resulting pixel's accuracy is decreased. The smoothing effect of bicubic interpolation has made it a powerful tool in the field of SR.

(ii) **Reconstruction-based:** To minimize the artifacts produced by interpolation-based methods, different reconstruction-(also known as regularization) based

techniques are applied on the LR image to reconstruct HR image. Statistical restoration approaches, such as deterministic and stochastic approaches are used in SISR that is based on reconstruction. In order to reduce the number of potential solutions, these strategies utilize different prior information as constraints. The deterministic method accurately predicts the relationship between the LR and HR images. Unfortunately, this method of reconstruction is inadequate for solving the SR problem, which is notoriously ill-posed. The Bayesian techniques, such as ML and MAP are used in stochastic approach. ML recovers HR images by maximizing the likelihood of LR images generated from them. While an augmented optimization function and prior term estimate the HR image in MAP. Mathematically, it can be expressed as follows:

$$\mathbf{x} = \underset{\mathbf{x}}{\operatorname{argmin}} \sum_{k=1}^{N} \|SH\mathbf{x}_k - \mathbf{y}_k\|_2^2 + \lambda \Re(\mathbf{x}) . \tag{2.2}$$

where $k = 1, 2, ..., N$ number of patches of LR and HR images. The expression has two parts: the first: a data fidelity term, and the second: a constraint or regularization term. The regularization term consists of a constant λ and $\Re(-)$, the prior. There may be a requirement for more than one prior. Fig. 2.3 depicts various types of priors, which may be classified as local or non-local depending on the intrinsic features of the image. Local priors, such as gradient profile [92] and total-variation (TV) [5, 70], are utilized as constraints for solving inverse problems of SR by considering the statistical and local features of the image. On the other hand, non-local similarities in the image are used for regularization of the inverse in non-local priors [127]. They help in sparse representation required for the SR process. Further, a hybrid model that combines non-local means with TV regularization, called the non-local total variation (NLTV), is developed in the sparse representation-based SR process [128]. These types of hybrid models that utilize both local and non-local priors are often used to maximize their efficacies. However, reconstruction-based SISR methods generally generate HR outputs with better sharpness, but at the cost of undesired edges, as well as ringing artifacts, especially at the salient edges. Also, these methods are often computationally intensive, and their efficacy

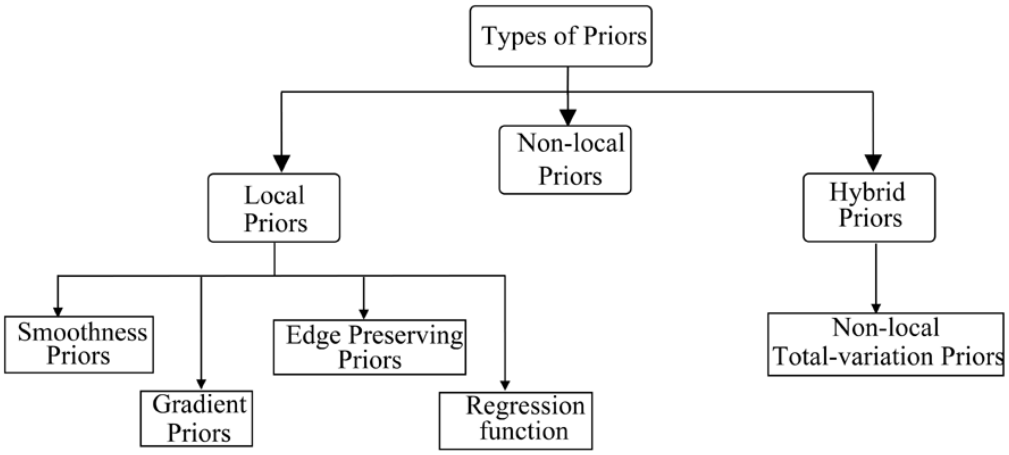

Figure 2.3: Types of priors.

gradually decreases as the upscaling factor increases.

(iii) **Learning-based** In recent years, learning-based approaches have shown excellent performance over the aforementioned methods in terms of reconstruction quality as well as perceptual accuracy. These approaches can effectively recover missing high frequency information by learning the relationship between HR examples and their corresponding LR counterparts through training. This learned information is used as a prior knowledge, which is then incorporated in the reconstruction phase. This approach is discussed in more depth in the following section.

2.3 Learning-based SISR methods

This approach is divided into two subcategories based on the available datasets- external (exemplar-based) and adaptive. In the adaptive learning, the LR input image is used instead of an external database-a collection of high-quality images related to the input LR image. The external example-based learning maps the relationship between HR and LR image patch pairs using an external dataset. Methods, such as the neighbor embedding [12], the random forest [85], the anchored neighborhood regression [98, 99], the sparse coding [25, 118, 122], and the deep learning (DL) [18, 22, 45, 87, 97, 125] are the most popular as they can predict the HR

patch by learning the correspondance between LR and HR patches using an external dataset.

2.3.1 Sparse-representation-based SISR

Sparse coding has the ability to give improved reconstruction results and outperforms traditional SR methods. Sparse coding-based SR method can be divided into two stages: first, training of the sparse overcomplete dictionaries, and second is the reconstruction of the SR image by using the sparse representations. According to the database for dictionary training, it can be either global (external database) or adaptive (internal database). SR methods using global dictionary very much depend upon the quality of example HR images in the training database. Yang *et al.* [118] proposed sparse prior-based SR method (ScSR), which is based on a global dictionary, learned from both LR and HR image patches. Zeyde *et al.* [122] simplified ScSR that makes it computationally less heavy and used the K-singular value decomposition (K-SVD). Timofte *et al.* [99] proposed adjusted anchored (A+) method; an improved variant of anchored neighborhood regression (ANR) [98]. A+ relies on the regressors and features of ANR, it uses the whole training data for learning regressors rather than the dictionary. Zhang *et al.* [131] proposed a simple and fast SR method that combines clustering and collaborative representation (CCR). It first extracts numerous clustered features from LR and corresponding HR images based on their local geometry. Using collaborative representation, several projection matrices are calculated in order to map between LR and HR feature subspaces. Song *et al.* [90] proposed guided SR approach based on coupled dictionary learning-based multimodal image SR method (CDLSR) that uses another HR image modality as a prior to restore the HR image from its LR version. This method also shows remarkable adaptability to noisy LR input. In particular, it incorporates both the low-rank constraint and nonlocal self-similarity within the sparse representation model concurrently, aiming to retain the global similarity information effectively.

In some cases, like images having more local features such as texture, edges in

RS applications, global dictionary might not be able to represent the image patches accurately. More training samples may be employed to learn the database, but the correctness of information yield by the database for any LR test image can not be guaranteed. The reconstructed HR may seem reasonable, but the information acquired from the training database may be irrelevant. In order to overcome the aforementioned limitations of global dictionary, adaptive dictionary has been broadly used to solve the SR problem. In the adaptive dictionary, only the test LR image is used instead of an external database. It is assumed that many similar image patches may exist within the same and across different scales of the test LR image. Zhu *et al.* [141] proposed a fast novel SISR algorithm based on sparse representation and adaptive dictionary learning. This technique used a faster approximation of SVD and an orthogonal matching pursuit (OMP) algorithm for efficient implementation. Chang *et al.* [14] proposed joint-regularization-based SR (JRSR) that includes a group-residual-based regularization (GRR) and a ridge regression-based regularization (3R). GRR utilizes the non-local similarity, while 3R adds HR information from an external training dataset. Zhang *et al.* [123] proposed a joint sparse representation framework that utilizes the nonlocal similarity and the low-rank property of intensity and gradient images, respectively. Li *et al.* [56] proposed a hybrid framework that addresses both the low-rank requirement and nonlocal self-similarity in the sparse representation model to retain global similarity information.

2.3.1.1 Sparse-representation-based remote sensing SISR

Huihui [89] presented RS super-resolution method, which can reconstruct LR Landsat image based on a global dictionary and sparse coding algorithm. Satellite pour l'Observation de la erre-5 (SPOT-5) and simulated Landsat thematic mapper (TM)/enhanced thematic mapper plus (ETM+) are used to acquire HR and LR feature patches to learn the dictionary pair using the K-SVD. Hou *et. al* [37] proposed a novel sparse representation-based SISR for RS imagery by using the global joint dictionary model (GJDM) in order to exploit the local and global features of images. Pan *et al.* [74] presented an SR technique for RS images, where structural

self-similarity (SSSIM) and K-SVD based dictionary is trained by using the test LR image and its interpolated version as an intermediate HR image.

2.3.2 Deep learning-based SISR

DL-based SR techniques can be divided into two main categories: generative adversarial networks (GAN)-based [52, 59, 110] and convolutional neural network (CNN)-based SR [18, 22, 45, 61, 97]. Although GAN-based SR networks approximate the original HR images to produce more realistic and perceptually enhanced HR images, their main limitation is that the reconstructed HR images obtained by these networks have a large structural difference from the original HR images. Furthermore, GAN network training is a challenging task, posing significant practical constraints. CNN-based SR methods have the ability to minimize the structural error between reconstructed HR and original images by improving the objective criteria, resulting in better PSNR and SSIM values. In general, the CNN network consists of convolutional and activation layers that are stacked together to learn feature maps automatically. Dong *et al.* [22] introduced a DL network very first time in SISR by designing a simple three convolutional layered CNN network (SRCNN), which gives better performance as compared to the traditional SISR methods. However, the SRCNN architecture is insufficient for learning high-level deeper features of the image, therefore, Kim *et al.* [45] proposed another CNN-based SR method termed very deep super-resolution (VDSR), which contains 20 convolutional layers. This method has the ability to reconstruct HR efficiently as well as increase the training stability by integrating gradient clipping and residual learning. However, since targeted upscaled versions of LR images are fed to the networks, the computational cost and complexity of SRCNN and VDSR are significant. To overcome this issue, LR images are directly applied to the DL-model and post-processing methods after feature extraction are proposed to upscale the images [3, 23, 87, 134]. Shi *et al.* [87] presented an efficient sub-pixel convolutional neural network (ESPCN), which used a sub-pixel convolutional layer to upscale the LR feature maps to the desired output and significantly increased the speed of the reconstruction process. Dong *et al.* [23]

later introduced fast super-resolution convolutional neural network (FSRCNN) in order to upscale the feature maps directly from the LR images using deconvolutional layers as a post upsampling method. These post-upsampling [3, 23, 87, 134] methods have made them the preferred choice for many DL-based SISR algorithms.

Utilizing a higher depth of the model is an effective approach to analyze the hierarchical feature information from LR images and enhance the quality of the images [108]. However, the complexity of model training and the number of parameters increase as well. Kim *et al.* [46] proposed a deeply-recursive convolutional network (DRCN) that apply same convolutional layers recursively up to 16 times. The recursive application of this network involves repeatedly applying the same set of convolutional layers multiple times. The training complexity of this network is reduced by incorporating skip-connection and recursive supervision. Lim *et al.* [61] proposed an enhanced deep super-resolution (EDSR) by removing redundant modules from conventional residual network. Ahn *et al.* [4] proposed an efficient and precise deep network named Cascading Residual Network (CARN) for SISR by implementing a cascading mechanism into a residual network. The aforementioned approaches use deeper networks, which provide good performance, but they are ignoring factors such as the model size and the inference time. Both EDSR and CARN emphasize the importance of residual learning; however, their ability to capture spatial features and contextual information may be limited due to the absence of attention mechanisms. This presents a potential limitation in their representation capabilities.

2.3.2.1 Attention-based DL for SISR

The objective of the attention mechanism in DL is an effort to mimic the cognitive process a human brain would take up while in a decision making process. To put it another way, it can be considered as a tool to emphasize the most relevant information of an input, while designing the relevant networks and algorithms for their implantations [38]. Typically, a gating function is included in attention mechanisms to produce a feature mask. It has been used for a variety of computer vision

applications, including, image classification [38], image restoration [133] and image captioning [17]. To increase the representational ability and efficiency of DL network during learning, the network inculcates the ability to pay more attention to some particular areas of interests. A novel non-local attention-based DL network for video classification was developed by Wang *et al.* [106] by capturing long-range dependencies using non-local operations. Hu *et al.* [38] proposed a novel architecture named as "Squeeze-and-Excitation (SE)" module for image classification in order to increase the representational ability of the network by recalibrating the feature maps channel-wise. This network may leverage global information to enhance useful features and suppress irrelevant ones. Further, Woo *et al.* [115] introduced an attention mechanism which can recalibrate the feature maps in both spatial and channel-wise.

Attention mechanisms integrated into CNN-based SISR networks have been very effective for extracting informative features. The basic goal of SISR is to restore as much relevant high-frequency information as possible. But, generally CNN-based approaches treat the features from different channels and spatial locations with equal importance; limiting flexibility, when dealing with diverse information. To tackle this problem, numerous methods [58, 62, 64, 132] incorporate the attention mechanism into the DL-based SISR networks to effectively learn the high frequency information. Zhang *et al.* [132] proposed very deep residual channel attention network (RCAN), where the channel-based attention mechanism was first introduced into a residual in residual (RIR) structure for SISR to adaptively enhance the channel-wise feature. Dai *et al.* [18] proposed a second-order channel attention (SAN) module to learn feature interdependencies channel-wise using global covariance pooling in order to capture more discriminative features. Recently, Niu *et al.* [72] presented a holistic attention network (HAN) for SISR to establish the correlations among layers, positions and channels. Further, Lu *et al.* [58] proposed a channel and spatial attention-based SISR model to increase the representational ability of the network.

2.3.2.2 DL-based remote sensing SISR

The recent development in CNN-based networks have offered various new research avenues for remote sensing SR. Lei *et al.* [54] presented a deep CNN-based remote sensing SISR approach that uses local-global combination networks (LGCNet) to learn multi-level representations integrating both local features and global environmental priors. A multiscale residual neural network (MRNN) was developed by Lu *et al.* [65] to learn multiscale features for reconstructing high-frequency information in remote sensing SR images. Zhang *et al.* [125] presented a mixed high-order attention network (MHAN) for learning the hierarchical features to restore missing information in the reconstructed HR remote sensing images. This network combines a feature extraction block with a high-order attention-based feature refinement block. Dong *et al.* [26] introduced a second-order technique that effectively reuses both small- and large-difference features at both the local and global levels in order to maximize the utilization of learned multi-level information. Wang *et al.* [109] developed a lightweight SR network that incorporates context enhancement and contextual feature aggregation modules to boost the representational power of the extracted feature. Dong *et al.* [24] proposed a kernel aware SR network (KANet) for real-world RS images in order to address the degradation and the high-frequency recovery issues. Lei *et al.* [49] presented a hybrid-scale self-similarity exploitation network (HSENet), which enhances feature representations by exploiting single- and cross-scale similarity information in RS images. Similarly, Wang *et al.* [111] proposed a lightweight network termed as feature enhancement network (FeNet) for RS images, consisting of a lightweight latice block (LLB) based channel-wise and an attention module. Finally, a feature enhancement block (FEB) is designed with the LLB in a nested manner in two steps- first, obtains the most relevant features in different layers with varying texture richness. Next, features from different layers, deep to shallow, are sequentially fused to obtain the final feature vector.

2.3.2.3 Joint image SISR and deblurring

There are only a few strategies where CNN networks have been used to solve the image SR and deblurring jointly [29, 116, 124, 130]. Zhang *et al.* [130] proposed an encoder-decoder network to enhance resolution of LR images with uniform gaussian blur. Zhang *et al.* [129] proposed a gated fusion network (GFN) to super-resolve the blurry LR images by exploiting different branches to extract deblurring and spatial feature. Zhang *et al.* [124] proposed attention dual supervised network (ADSR) that performs both SR and deblur tasks jointly by incorporating dual supervised learning. Xi *et al.* [116] presented a pixel-guided dual-branch attention network (PDAN) to jointly solve SR and deblur problem by using hard pixel example mining loss (HPEM). Esmaeilzehi *et al.* [29] proposed a deep light-weight residual CNN for efficiently addressing the problem of image SR and deblurring through a nonlinear end-to-end mapping. The network incorporates a residual block with the ability to generate features in multiple receptive fields, which improves the network's representational abilities. As mentioned earlier, there is a lack of approaches available in the field of joint image SR and deblurring. Furthermore, to the best of our knowledge, no existing literature has specifically addressed the issue of joint image super-resolution (SR) and deblurring in the field of RS. Our thesis aims to fill this gap by proposing a novel approach; introducing a unique network architecture to address SR and deblurring jointly for RS images. Another limitation of the existing methods is that they have used the heavyweight deblurring network, while our proposed network uses a lightweight deblurring network by incorporating the attention mechanism to increase the representational abilities. The existing joint networks have not considered the spatial features effectively in order to preserve high-frequency details during the SR and deblurring process. The proposed network introduces attention mechanisms or other architectural modules that explicitly consider spatial information, leading to improved reconstruction of high-frequency details in the resulting HR and deblurred images.

2.4 Benchmarking criteria of remote sensing SISR algorithms

The process of benchmarking algorithms for remote sensing SISR involves evaluating the performance of the algorithms based on a set of criteria. A list of techniques and criteria that are often used for benchmarking RS SISR algorithms are as follows:

(i) **Evaluation metrics:** Different evaluation metrics are used for evaluating the performance of SISR algorithms on RS images, including the most commonly employed peak signal-to-noise ratio (PSNR), structural similarity index (SSIM), and other RS image-specific metrics for assessment. The detailed explanation will be provided in Section 2.7. In the context of RS imagery, where the preservation of fine details and texture is paramount, PSNR and SSIM serve as invaluable metrics for assessing the performance of SISR algorithms. Numerous studies in the RS field have relied on these metrics to evaluate the effectiveness of various SR techniques. Previous research works such as GJDM [37], [74], MHAN [125], and HSENet [49] have leveraged PSNR and SSIM to quantify the improvement in image quality achieved by their respective SISR methods. In addition to PSNR and SSIM, other RS image-specific metrics play a crucial role in providing a comprehensive evaluation of SISR algorithms. These metrics are tailored to address the unique characteristics by RS imagery. Previous research works such as [89], [44], and [81] have explored the use of these specialized metrics to assess the performance of SISR algorithms in RS applications.

(ii) **Datasets:** Standardised datasets are necessary for benchmarking in order to accurately evaluate the algorithms. Publicly available RS datasets, such as PatternNet, AID, and many more freely available datasets, should be included to evaluate SISR algorithms. Previous studies have extensively utilized datasets like PatternNet and AID for evaluating SISR algorithms. For instance, research works such as [16], [94], MHAN [125], [26], and [86] have

leveraged these datasets to assess algorithm performance and compare results against state-of-the-art techniques. By using these standardized datasets, researchers can ensure the reproducibility and reliability of their findings, thereby advancing the state-of-the-art in SISR for RS applications. In the case of RS, real ground truth data is unavailable for benchmarking SISR algorithms; therefore, simulated datasets are often used. These datasets are generated to mimic realistic RS scenarios, including sensor characteristics, atmospheric effects, etc. Simulated datasets provide an appropriate setting for assessing the performance of algorithms. Simulated data may be used for reliable assessments and insights on algorithmic strengths and shortcomings, despite the fact that it lacks genuine ground truth.

(iii) **Visual quality assessment:** In addition to quantitative measures, visual quality evaluation is crucial. Human observers compare SR images to ground truth images to assess their visual quality. Previous studies, including [74], [89], [37], and [125], have utilized visual quality as a metric for quantitative evaluation. Quantitative metrics may sometimes neglect perceptual differences; they might be captured by subjective evaluations.

(iv) **Computational efficiency:** When benchmarking, the computational efficiency of algorithms is also taken into consideration by measuring the time required to generate a HR image from a LR input. The processing times of different SISR algorithms should be compared to assess their efficiency.

(v) **Comparison with state-of-the-art:** Benchmarking also considers comparative analysis of SISR algorithms by comparing them to state-of-the-art methods. These algorithms are carefully selected to cover a diverse range of approaches, including sparse coding techniques, DL-based methods, and traditional interpolation methods. Comparative analysis highlights improvements in terms of visual quality, quantitative metrics, and computational efficiency. Moreover, it provides valuable insights into the relative performance gains achieved by new methodologies compared to established benchmarks, facilitating the continuous evolution and refinement of SISR techniques. Pre-

vious studies, such as those conducted by GJDM [37], [74], MHAN [125], HSENet [49], and FeNet [111], have employed similar comparative approaches for benchmarking SISR algorithms. These studies have systematically evaluated the performance of various techniques across a range of evaluation criteria, providing a comprehensive assessment of algorithmic capabilities and limitations.

2.5 Parallel computing for SR

One of the most popular parallel computing devices in computational machines is the graphics processing unit (GPU), which is used for engineering and scientific computing since several parallel processes are concurrently executed on hundreds of processor cores and thousands of threads. Since computing power of GPUs is highly intensive, they are much faster than the CPU. GPUs accelerate image/video and computer vision algorithms by exploiting their parallel architecture to handle data simultaneously over thousands of cores, enabling real-time image/video processing. They excel at performing operations such as filtering, transformation, and deep learning computations, making them indispensable for applications, like object detection, image classification, and video analysis. The typical CPU-GPU architecture comparison is shown in Fig. 2.4. The computational power of GPU can be harnessed with the help of many application programming interfaces (APIs) without knowledge of graphics programming, which allow users to boost the performance of any time consuming and computationally heavy algorithms. General-purpose programming on GPU has become very popular since the introduction of advanced programming environments, like compute unified device architecture (CUDA) [91] and open computing language (OpenCL) [69], etc. It can also increase the computing efficiency of many applications by using existing hardware on end-user devices. CUDA is a parallel programming platform introduced by NVIDIA in 2007. It is used to create software for graphics processors and to create a wide range of general purpose applications for GPUs that are extremely parallel in design and run-on hun-

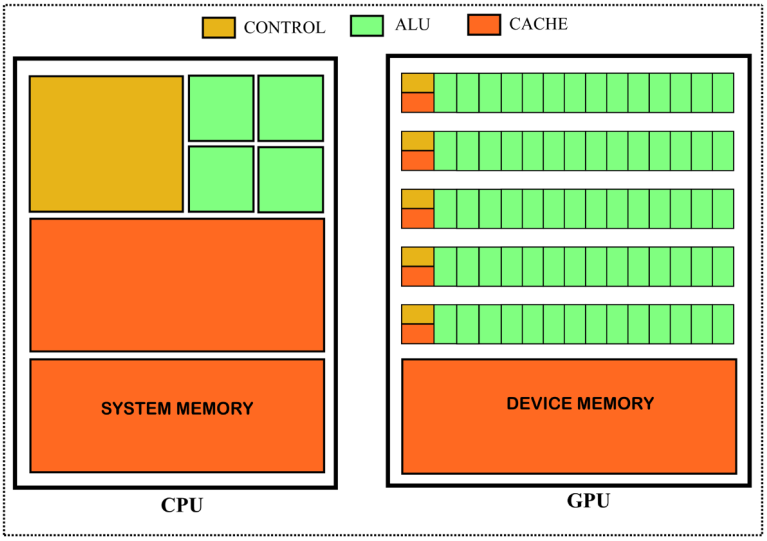

Figure 2.4: Comparison between CPU and GPU architecture.

dreds of GPU processor cores. The CUDA API enables users to build a huge number of threads to run a code on the GPU. A block is composed of several threads, which are indexed in the block using "threadIdx". A grid is arranged in the same way, and each block in a grid is indexed using "blockIdx". "ThreadIdx" and "blockIdx" are both CUDA pre-defined variables. In addition, there are also two pre-defined variables "blockDim" and "gridDim", which are used to specify the size of a block or grid determined by the total number of threads per block or the total number of blocks per grid. As shown in Fig. 2.5, all threads in CUDA are arranged [91] into a hierarchical way: grid and block. Kernels are specific functions used in CUDA that is called by the CPU. It runs N times in parallel on the GPU using N threads. CUDA also supports shared memory and thread synchronization. The CUDA programming model combines serial and concurrent processing. An ordinary CUDA programme consists of three steps: copying data from the CPU/host memory to the device's global memory, execution of CUDA codes in kernels, and restoring data from the device memory to the host memory. CUDA uses the bottom-up approach of parallelism, with a thread serving as an atomic unit of parallelism

MS remote sensing images have a large data volume since they consist of several spectral bands; processing these data in near real time has been a significant computational problem. Also, sparse representation-based SR methods are computationally very intensive due to their inverse ill-posed nature. Additionally, another

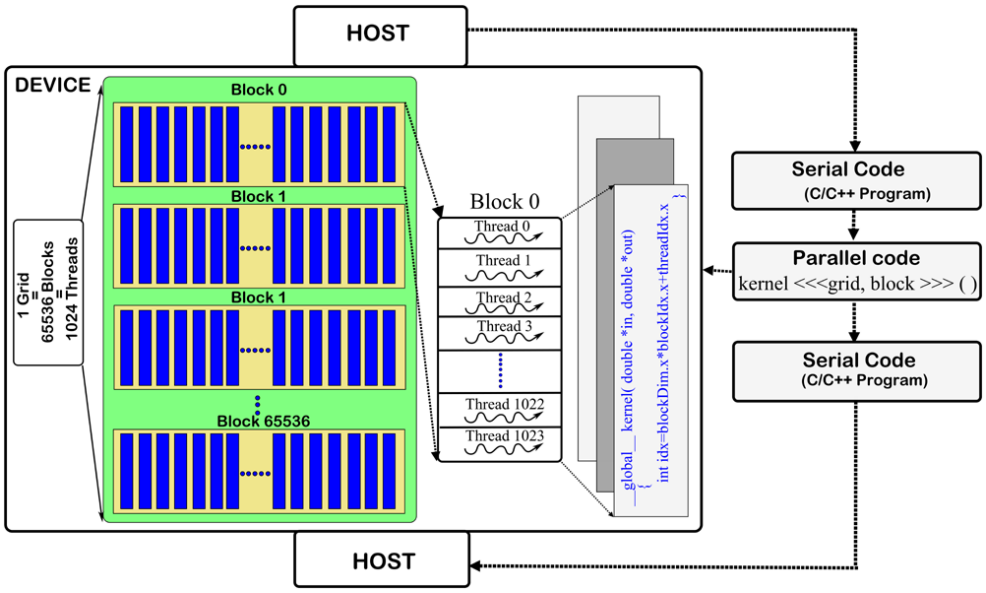

Figure 2.5: Heterogeneous architecture of CUDA and its three-level thread hierarchy.

reason of having slow performance in sparse representation-based SR method is that many image patches are being processed sequentially. So, computationally, it becomes highly exhausted because of the sizes of real RS images. In order to solve the above problems, the parallel computing approach can be exploited to design highly parallelized sparse-representation algorithms and implement on CUDA-enabled GP-GPU for real time SR reconstruction of RS images. Similarly, DL network encounters an intensive computational overhead, especially at the time of network training and re-training. Training SR models demand computationally heavy matrix multiplications and other arithmetic operations. It may take several days for a CPU to complete intensive training of a DL-based SR model on an extremely large dataset. For optimizing the training time of DL, GPU is the most straight-forward, yet highly efficient choice to the researcher.

In recent years, SR techniques employing parallel hardware have been proposed for real-time applications [43, 96, 120]. Yuan *et al.* [120] have proposed a CUDA-based SISR method that employ an internal training dataset to exploit image self-similarity. This method achieves better visual quality and speed-up than the existing methods. Hanlin *et al.* [96] have proposed CUDA-accelerated SISR based on a fast least absolute shrinkage and selection operator (LASSO) that outperforms state-of-the-art methods and exhibits 6.2× speed up. Moustafa *et al.* [68] have reported a

fast CUDA-enabled GPU accelerated MS image SR using morphological component analysis (MCA) and adaptive dictionary. It achieves speed-up of around 20× to 40× for different image sizes.

2.6 Remote sensing datasets

There are many publicly available RS datasets, including the aerial image dataset (AID) [117], PatternNet [138], university of california, Merced (UCmerced) [119], remote sensing Scene classification (RSSCN7) [142] and WHU-RS20 [39], etc. However, PatternNet and AID are chosen to validate the work done in the thesis. These datasets offer a larger number of images, thus enhancing their suitability for training DL-based SISR methods. Furthermore, real MS remote sensing images, like Linear Imaging Self Scanner-3 and -4 (LISS-III and LISS-IV) are collected from National Remote Sensing Centre (NRSC), Hyderabad to carry out the experiments. Table 2.1 elucidates detailed specifications of publicly available RS datasets.

2.6.1 Publicly available datasets

In this thesis, the following publicly available datasets were selected to carry out the experiment:

(i) **PatternNet:** It consists of 30,400 images, making it a large-scale HR remote sensing dataset. These images were acquired from US cities using the Google Map API and Google Earth imagery. It has 38 classes, each with 800 images of size 256×256, with spatial (pixel) resolutions ranging from 0.062 meter to 4.693 meters (m).

(ii) **AID:** A RS high-resolution dataset, collected from Google Earth imagery over different countries: China, the US, England, France, Italy, Japan, Germany, etc. It has 30 classes, each containing 200–420 images of size 600×600. This

dataset has pixel resolutions ranging from 0.5 to 8 m.

Table 2.1: Specifications of publicly available RS datasets.

Dataset	Images per class	Classes	Images	Pixel Resolution (m)	Image Size	Types of classes
PatternNet	800	38	30,400	0.062–4.693	256×256	Airplane, baseball field, basketball court, beach bridge, cemetery, chaparral, Christmas tree, closed road, coastal mansion, crosswalk, dense residential, ferry terminal, football field, forest, freeway, golf course, harbor, intersection, mobile home park, nursing home, oil gas field, oil well, overpass, parking lot, parking space, railway, river, runway, runway marking, shipping yard, solar panel, sparse residential, storage tank, swimming pool, tennis court, transformer station, Wastewater treatment plant
AID	220–420	30	10,000	0.5 -8	600×600	Airport, bare land, baseball field, beach, bridge, center, church, commercial, dense residential, desert, farmland, forest, industrial, meadow, medium residential, mountain, park, parking, playground, pond, port, railway station, resort, river, school, sparse residential, square, stadium, storage tanks, and viaduct.

2.6.2 Self-procured real MS remote sensing images

Real MS remote sensing images collected by Indian satellites: Resourcesat 2A equipped with sensors LISS-III and LISS-IV are procured through the National Remote Sensing Centre (NRSC) data center of Indian Space Research Organization (ISRO), Govt. of India. LISS-III provides four MS bands- two in visible, one in near infrared (NIR) and one in Short Wave Infrared (SWIR). This sensor covers a 140-km orbital swath at a spatial resolution of 23.9 m with a 24-day repeat cycle. in three spectral bands in the Visible and Near Infrared Regions (VNIR) with 5.8 m spatial resolution. On the other hand, LISS-IV provides images in three spectral bands: two in visible and one in NIR (NIR) with a spatial resolution of 5.8 m. These images are downloaded via FTP (`ftp1.nrsc.gov.in`). These land cover images are captured during January 6, 2013 to April 10, 2018, over many places in India. The detailed specifications of these datasets are given in Table 2.2.

2.7 Image quality evaluation metrics

The quality of SR reconstruction can be evaluated using qualitative approaches based on human perception, such as how realistic the reconstructed image is in vi-

Table 2.2: Specifications of LISS-IV and LISSS-III satellite sensor.

Satellite Sensor	LISS-IV	LISS-III
Spatial resolution	5.8 m	23.5 m
Spectral bands	Green (Band 2: 0.52-0.59 μm) Red (Band 3: 0.62-0.68 μm) Near infrared (Band 4: 0.77-0.86 μm)	Green (Band 2: 0.52-0.59 μm) Red (Band 3: 0.62-0.68 μm) NIR (Band 4: 0.77-0.86 μm) SWIR (Band 5: 1.55-1.70 μm)
Swath	23.9 km	141 km
Image size	18000×16000	7700×7000

sual terms besides quantitative metrics. Qualitative methods are simpler and more practical, but these methods suffer the following limitations: (i) personal preferences have a significant impact on the evaluation; (ii) the assessment process is often costly and time taking process. Quantitative methods or objective assessment, on the other hand, is more straightforward to use. There are two types of quantitative methods to validate the SR results: reference-based that carry out evaluation using reference images and no-reference-based metrics, which do not use any reference images. In case of SR, the original HR images (acquired images without applying blurring and downsampling operations) are considered as the reference images or ground truth. The most commonly used reference metrics for quantifying SR images are: PSNR and SSIM. While, reconstruction quality of remote sensing SR is measured in terms of erreur relative globale adimensionnelle de synthese (ERGAS) [105], spectral angle mapper (SAM) [121], universal image quality index (Q-index) [113] and spatial correlation coefficient (sCC) [137]. For no-reference-based quantitative metrics, natural image quality evaluator (NIQE) [67] and entropy (EN) are used to validate the results. Notably, PSNR and SSIM have been extensively used to evaluate SR performance in RS, as seen in studies by GJDM [37], [74], MHAN [125], and HSENet [49]. Similarly, ERGAS, SAM, Q-index, and sCC are employed for assessment, as demonstrated in works such as [89], [44], and [81]. NIQE is utilized for evaluation in studies such as [60], [140]. These metrics are discussed in details in the following sections:

2.7.1 Reference-based quantitative metrics

(i) **Peak signal-to-noise ratio (PSNR):** PSNR is the most commonly used full-reference quantitative metric for measuring the quality of image reconstruction. It can be calculated by using the following formula:

$$\text{PSNR(dB)} = 10 \log \frac{255^2 \times M \times N}{\sum_{i=1}^{M-1} \sum_{j=1}^{N-1} \left[X(i,j) - \hat{X}(i,j) \right]^2} \quad , \tag{2.3}$$

where M and N are row and column dimensions of the image. X and $\hat{X}$ are the ground truth and the reconstructed images, respectively. It is considered as a good reconstruction, if the PSNR of the reconstructed image is high since it may lead to improved visual perception, feature clarity, and edge retention. This means that higher the PSNR better is the reconstruction quality.

(ii) **Mean structural similarity index (SSIM):** SSIM [114] is used to measure structural similarity between the original (X) and the reconstructed image ($\hat{X}$) in terms of luminance, structure and contrast. It is described as follows:

$$\text{SSIM} = \frac{(2\mu_X \mu_{\hat{X}} + C_1)(2\sigma_{X\hat{X}} + C_2)}{(\mu_X^2 + \mu_{\hat{X}}^2 + C_1)(\sigma_X^2 + \sigma_{\hat{X}}^2 + C_2)} \quad , \tag{2.4}$$

where μ_X, $\mu_{\hat{X}}$ and σ_X, $\sigma_{\hat{X}}$ represent the means and standard deviations of X and $\hat{X}$, respectively; C_1 and C_2 are the constants. $\sigma_{X\hat{X}}$ denotes the covariance between X and $\hat{X}$. Ideally, SSIM value is unity for a perfect similarity between X and $\hat{X}$. Therefore , it indicates that the reconstruction quality is better if the SSIM value is higher.

(iii) **Erreur relative globale adimensionnelle de Synthese (ERGAS):** ERGAS [105] is used to measure quality of the reconstructed image $\hat{X}$ by taking into account scaling factor and root-mean-square error (RMSE) values and is

expressed as follows:

$$\text{ERGAS} = \frac{100}{S}\sqrt{\frac{1}{C}\left(\frac{\sqrt{MSE(X,\,\hat{X})}}{\mu_0}\right)} \quad , \tag{2.5}$$

where S is the SR scaling factor, C represents the number of bands of the MS remote sensing image, and μ_0 is the mean value of the ground truth HR image (X). The value of ERGAS is zero for ideal case. The lower the ERGAS values, the higher the quality of the reconstructed image.

(iv) **Spectral angle mapper (SAM):** SAM [121] is used to compare similarity of spectra between the reconstructed ($\hat{X}$) with respect to the ground truth (X). It calculates the average angle between the pixels of X and $\hat{X}$ using each band as a coordinate axis.

$$\text{SAM} = \frac{1}{N}\sum_{i}^{N}\arccos\frac{X_i\hat{X}_i}{\|X_i\|\left\|\hat{X}_i\right\|} \quad , \tag{2.6}$$

where N is total number of pixels. Zero SAM value indicates no spectral distortion. Therefore, the lower the SAM value, the higher the reconstruction quality.

(v) **Universal image quality index (Q-index):** Q-index [113] utilizes three properties such as correlation, luminance and contrast for evaluating the reconstructed image ($\hat{X}$) quality with respect to the original HR image (X).

$$\text{Q} = \frac{1}{C}\sum_{j}^{C}\left(\frac{\sigma_{X\hat{X}}}{\sigma_X\sigma_{\hat{X}}}\frac{2\mu_X\mu_{\hat{X}}}{\mu_X^2\mu_{\hat{X}}^2}\frac{2\sigma_X\sigma_{\hat{X}}}{\sigma_X^2{\sigma_{\hat{X}}}^2}\right) \quad , \tag{2.7}$$

where μ_X and $\mu_{\hat{X}}$ are the mean values of X and $\hat{X}$, respectively. Similarly, σ_X and $\sigma_{\hat{X}}$ are the standard deviations of X and $\hat{X}$, respectively, C is the total number of MS remote sensing bands, and $\sigma_{X\hat{X}}$ represents the covariance between X and $\hat{X}$. Ideally, Q-index should be equal to one.

(vi) **Spatial correlation coefficient (sCC):** sCC [137] measures the normalized

cross-correlation between the reference image (X) and the reconstructed image ($\hat{X}$). It is expressed as

$$\text{sCC} = \frac{\sigma_{X\hat{X}}}{\sigma_X \sigma_{\hat{X}}} \quad , \tag{2.8}$$

where σ_X and $\sigma_{\hat{X}}$ denote the standard deviations of X and $\hat{X}$, respectively.

2.7.2 No-reference-based quantitative metrics

No-reference metrics are used to estimate the image quality independently without employing a ground truth image. The following parameters are used for no-reference metrics:

(i) **Natural image quality evaluator (NIQE):** The NIQE technique involves generating a set of quality-aware features and then fitting them into a multivariate Gaussian (MVG) model. These quality-aware features are generated using a highly regular natural scene statistics (NSS) model [67]. Finally, the quality is determined by computing the distance between the MVGs of high quality images and target distorted images.

(ii) **Entropy:** Entropy is used to evaluate the quality of the reconstructed image based on the probability distribution of pixel intensities in the image. The higher the value of entropy, higher the reconstruction quality. It is calculated by

$$\text{Entropy} = -\sum_i P(\hat{X}_i) \log P(\hat{X}_i) \tag{2.9}$$

where $P(\hat{X}_i)$ is the probability associated with the gray-level of reconstructed image ($\hat{X}_i$).

2.8 Summary and research gaps

From the above literature study, we may summarize the following research issues on RS image SR:

a) The impact of dictionary learning based on sparse representations is very effective on SISR. The reconstruction of LR image highly depends on how good the dictionary is. Since global dictionary-based SR has high computational cost and memory consumption, adaptive dictionary is widely used in sparse representation-based SISR. In the RS domain, it is quite time consuming and expensive to produce the HR training data. On that account, self-example or adaptive dictionary learning based SISR is a promising technique for obtaining better results in MS remote sensing imagery

b) Preservation of textural or structural features, particularly the edges, is crucial in the SISR of MS remote sensing images. In order to obtain edge-enhanced SR, these features may be deployed as a prior information in both dictionary learning and sparse reconstruction. Similarly, feature extraction from LR images plays a crucial role in the sparse representations-based SISR methods. In many sparse coding-based SISR methods, gradient-based feature extraction is used, but it fails to produce sharp edges. Most of the sparse representations-based methods have not enhanced the edge-based high-frequency details in the reconstructed images. An edge-enriched feature extraction scheme along with a joint sparse reconstruction framework may help to preserve high-frequency information as well as smooth structures of the reconstructed image.

c) In the sparse representations-based SISR approach, dictionary learning and sparse regularizations involved in solving the sparse approximation problem are quite time consuming. Moreover, in the above problems, large number of HR and LR patch pairs on which the patch-based sparse coding sub-problems are to be solved individually, making the entire process highly computationally intensive. As a result, near real time SR reconstruction for RS image is quite challeng-

ing. Therefore, in order to solve patch-based sparse representation problems efficiently, parallel computing is done on highly parallelized algorithms by using GP-GPU for near real time SR reconstruction of RS images.

d) Although attention-based DL SISR networks have been gaining popularity on natural images, it is not significantly explored in the RS domain. Therefore, there is a great scope for implementation and improvement of such networks specially designed for RS. It is found from the literature that deep CNN-based SISR methods treat spatial and channel features in the same manner, which results in the lack of discriminative learning ability when dealing with these features. Therefore, it is vital to understand how CNN-based SISR approaches can be applied effectively both channel- and spatial-wise in order to reconstruct HR images, and requires detailed investigation.

e) RS images frequently suffer from a reduction in spatial resolution as well as the effect of blurring due to imperfect luminance, atmospheric propagation, and sensor characterization. Dealing with the blurring problem in LR remote sening images using the CNN-based SISR networks is very challenging because they focus only on improving the resolution of LR images but do not explicitly address problem of blurring separately. Although certain minimum level of blurring can be managed, but they struggle to restore high-frequency information (edges and textures), when LR remote sensing images suffer from uniform blur. There are only a few CNN-based networks available for RS images in the present literature that solve both the problems concurrently. Therefore, it is crucial to design an effective joint deblurring and SR network to super-resolve the blurred LR remote sensing images.