Graphical Abstract

Deep learning for single-image super-resolution in remote sensing: A review

Highlights

Deep learning for single-image super-resolution in remote sensing: A review

- A summary of deep learning architectures in super resolution field.
- A critical review of remote sensing datasets and methods for single image super resolution since 2016.
- Current challenges include training data quality, model generalization and transferability, high scale factor super resolution, lack of application-aware model designs, computational concerns, and limited physical constraints.
- Future directions could focus on constructing application-oriented datasets and models, exploring generative priors, self-supervised and unsupervised learning, multi-task learning, and physically guided methods.

Deep learning for single-image super-resolution in remote sensing: A review

Abstract

Remote sensing image super-resolution has become a critical task to enhance spatial detail for downstream applications such as land cover mapping, environmental monitoring, and precision agriculture. However, the unique characteristics of remote sensing data—including limited training samples, complex sensor degradations, and spectral diversity—pose significant challenges to conventional super-resolution pipelines. In this paper, we present a comprehensive review of recent advances in remote sensing single image super resolution (RSSISR), spanning across supervised, self-supervised, unsupervised, training-free, generative adversarial network-based, diffusion-based, and physically/statistically guided frameworks. We introduce a structured taxonomy that organizes these approaches and analyze their strengths, limitations, and application domains.

In addition, we identify key challenges in the field, including (1) data scarcity, (2) generalization gaps, (3) high computational cost, (4) lack of task-oriented evaluation, and (5) limited physical integration in degradation process modelling. By pinpoiting these challenges, we outline promising research directions including (1) data augmentation and enhancement, (2) generative priors and large foundation models, (3) learning paradigms less reliant on supervision, (4) application-aware frameworks, (5) physics-guided modeling, and (6) multi-task learning. Case studies across real-world tasks such as crop monitoring, spatio-temporal image upscaling, flood mapping, and object detection further illustrate the practical utility of different RSSISR strategies. This review not only summarizes the current landscape but also provides a forward-looking perspective on the future of RSSISR.

Keywords: Single Image, Super Resolution, Remote Sensing, Deep Learning, Diffusion Model

1. Introduction

Coarse spatial resolution is one of the most important factors hindering the application of remote sensing (RS) images [1]. RS utilizes satellite or aerial imagery to reflect the observation, which may have limited spatial and/or spectral resolution due to sensor degradation or atmospheric interference [2]. For instance, Landsat-8's 30 m resolution is often insufficient for smallholder crop fields smaller than 1 ha [3] or urban land cover mapping [4], while Sentinel-2's 10 m bands still struggle to delineate narrow crop strips, small and irregularly shaped fields [5].

Single image super resolution (SISR) relies on one given low-resolution (LR) image to generate the corresponding high-resolution (HR) image, as shown in Figure 1 [6]. SISR can play a crucial role in enhancing the spatial resolution of RS images without the need for additional sensor upgrades. There are three main categories of SISR methods: interpolation-based, reconstruction-based, and learning-based methods [7]. Interpolation-based methods, such as nearest neighbor, bilinear and bicubic interpolation, can estimate new pixels based on weighted calculations of surrounding pixels, but often produce overly smooth results and fail to recover feature edges and textures. Reconstruction-based methods introduce image priors (e.g., image gradients) to guide HR image generation [7]. These methods can better preserve details than interpolation, but often rely on handcrafted features and are sensitive to noise. Last but not least, learning-based methods—especially those based on deep learning (DL)—have recently dominated the SISR field [7]. These methods learn mappings from LR to HR images from large data pairs, leveraging convolutional neural networks (CNNs) such as SRCNN [8], VDSR [9], and EDSR [10], generative adversarial networks (GANs) such as SRGAN [11] and ESRGAN [12], and other model architectures. These methods can effectively capture complex patterns, learn general representations, and generate high-fidelity images.

Recent advances in RS single image super resolution (RSSISR) have been extensively reviewed in previous surveys. For example, there is a study summarizing various families and methods of RSSISR, yet primarily focusing on traditional techniques grounded in physical models and prior knowledge, rather than deep learning approaches [13]. Alternatively, a recent review summarized SISR models from a computer science perspective, covering interpolation methods, reconstruction methods, and learning-based methods [7]. However, this study lacks perspectives of domain-specific applications

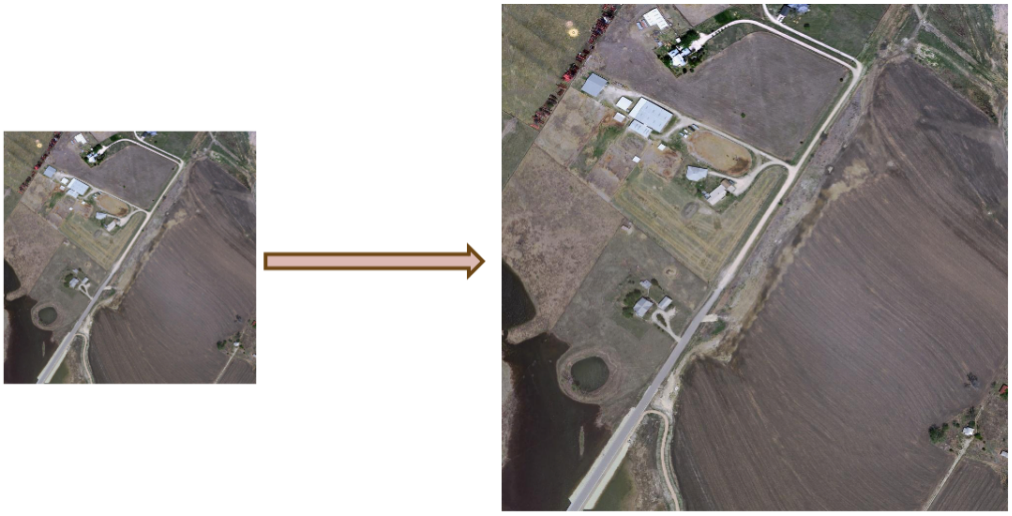

Figure 1: SR aims to reconstruct a high-resolution (HR) image from its degraded low-resolution (LR) counterpart.

such as RS field. More recent literature introduces significant deep learning work on RSSISR, covering aspects including datasets, image quality, model performance evaluation, design principles, and relevant applications [14, 15, 16]. However, given the rapid development in this field, many state-of-the-art (SOTA) models and approaches may not have been covered in these studies. In addition, this review provides a critical analysis of the current literature, identifies remaining challenges, and discusses potential directions for future research. Therefore, it aims to offer valuable insights into emerging models and trends, highlighting the need for continued investigation. Our contributions and highlights can be summarized as:

- We summarize DL architectures in RSSR field. The review highlights that generative priors and large foundation models show promising results across various fields, and are still underexplored in RS field.
- We provide a comprehensive review of the DL-based RSSISR algorithms thoroughly since 2016. Current research predominantly focus on supervised learning methods. However, self-supervised, unsupervised and training-free learning deserve further investigation.

- Despite substantial advancements, challenges such as training data quality, poor model generalization and transferability, high computational cost, oversimplified degradation process, and limited application alignment remain unresolved. Moreover, many current models focus on visual metrics while neglecting performance in downstream geospatial tasks.

- To address these limitations, we identify several promising research directions, including data augmentation and enhancement, integration of generative priors and foundation models, exploring more training paradigms less reliant on supervision, application-aware designs, and physically guided approaches that simulate sensor characteristics and real-world degradations. Additionally, the rise of diffusion models and vision-language architectures opens new opportunities for zero-shot enhancement and multimodal learning.

2. Methodology

2.1. Problem Definition

RSSISR aims to restore HR RS images from the corresponding LR RS images. Typically, LR images could be represented as a degradation of HR images:

$$\mathbf{I}_{LR} = D(\mathbf{I}_{HR}) + \mathbf{n} \tag{1}$$

where: D represents the degradation operation (e.g., blurring and downsampling); $\mathbf{n}$ represents noise.

Thus, the super-resolution process could be represented as:

$$\mathbf{I}_{HR} = D^{-1}(\mathbf{I}_{LR} - \mathbf{n}) \tag{2}$$

where D^{-1} is the inverse of the degradation operation.

Since directly computing D^{-1} is usually impractical , DL models can be used to approximate this inverse mapping:

$$\mathbf{I}_{HR} \approx f(\mathbf{I}_{LR}) \tag{3}$$

where f is a non-linear mapping function learned by the DL models.

2.2. Literature

The review followed a systematic search process that encompasses 124 documents from the Scopus database [17]. The search strategy involved using keywords "super resolution" AND "remote sensing" AND "deep learning" AND ("single image" OR "single frame" OR "single input" OR "single view") along with classic DL architecture papers and reviews, as shown in Figure 2. The subject areas were limited to Computer Science, Earth and Planetary Sciences, Engineering, Mathematics, Environmental Science, ensuring English-language, peer-reviewed articles, conference proceedings, and scientific book sections. The review applied specific inclusion criteria for the selected documents, which included: (a) utilization of RS images, instead of natural images; (b) exploration of DL architecture improvements; (c) publications in reputable academic literature. By employing these criteria and conducting a rigorous search, the review aimed to provide a comprehensive and well-rounded overview of the state of RSSISR and the associated DL architectures.

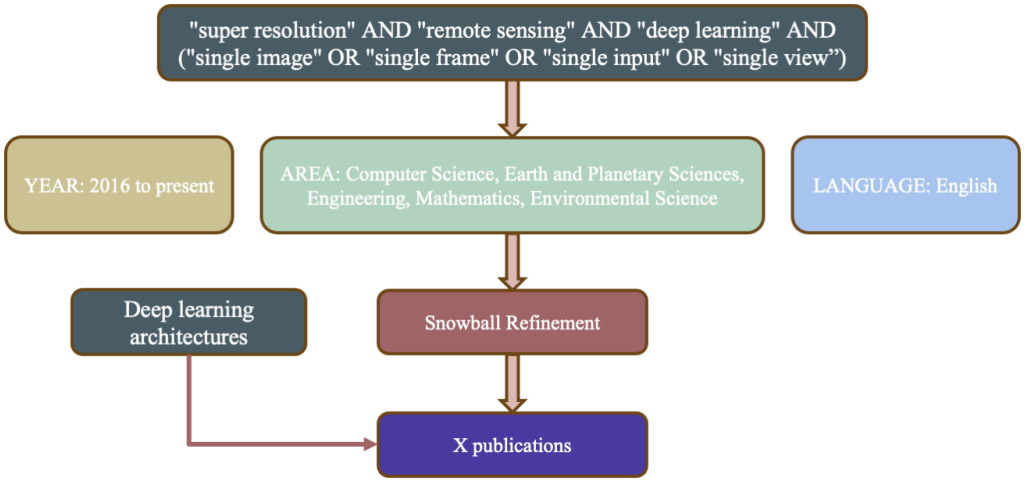

Figure 2: Flowchart of the methodology applied for paper selection.

3. RS Datasets

Table 1 summarized commonly used datasets in RSSISR field. Note that these datasets do not actually contain HR-LR pairs, the LR images are interpolated using the bicubic method [18].

AID: The AID dataset contains 10,000 aerial images across 30 scene categories, including airport, forest, stadium, and residential areas [19]. Each image is 600×600 pixels with spatial resolution ranging from 0.5 to 8 meters.

DOTA: This dataset includes 2806 aerial images from different sensors and platforms [20]. Each image is 4000 × 4000 pixels. It supports object detection tasks with annotations across 15 categories such as ship, vehicle, and airplane. The dataset features diverse imaging conditions, including scale variation, orientation, and shape complexity.

UC Merced: It consists of 21 classes of land-use and possibly object images with 256×256 pixels selected from aerial orthoimagery [21]. The images are extracted from orthorectified aerial photos over 21 U.S. regions. They are then cropped into small regions of 256×256 pixels. It is widely used for land cover classification and is one of the most benchmarked datasets in RSSR field.

WHU-RS19: The WHU-RS19 dataset includes 1005 images from Google Earth, each with 600×600 pixels [22]. The dataset covers 19 classes such as airport, viaduct, and commercial area. Earlier versions included only 12 classes, but new updates introduced 7 more. It is commonly used for scene classification under various spatial resolutions.

RESISC45: The NWPU-RESISC45 dataset consists of 31500 images across 45 scene classes such as mountain, desert, and storage tanks [23]. Each image is 256×256 RGB pixels. It was collected globally from over 100 countries using Google Earth, and is known for high intra-class variability and diversity in geography.

RSCNN7: This dataset contains 2,800 images (400×400 pixels) from Google Earth across 7 scene types such as farmland, industrial region, and residential areas [24]. Each class includes 400 images sampled across four spatial scales, angles, and seasons, making it suitable for evaluating robustness in scene classification.

Pavia Center: The Pavia Center dataset includes 7,456 hyperspectral samples (1096×1096 pixels) with 102 spectral bands, acquired by the ROSIS sensor [25]. It captures urban scenes in northern Italy and includes 9 labeled categories. The dataset is frequently used for spectral-spatial super-

resolution and classification tasks.

Table 1: Commonly Used Datasets in RSSISR Field.

Dataset	Data Source	Amount	Image Size	Usage
AID [19]	Aerial	10,000	600×600	Scene Classification
DOTA [20]	JL-1, GF-2	2,806	4000×4000	Object Detection
UC Merced [21]	Air- and Space-borne	2,100	256×256	Land Cover Classification
WHU-RS19 [22]	Google Earth	1,005	600×600	Scene Classification
RESISC45 [23]	Google Earth	31,500	256×256	Scene Classification
RSCNN7 [24]	Google Earth	2,800	400×400	Scene Classification
Pavia Center [25]	ROSIS sensor	7,456	1096×1096	Scene Classification

Figure 3 shows the usage frequency of datasets listed in Table 1. Among research papers examined, UC Merced, RESISC45, and AID are the most popular datasets in RSSISR. Furthermore, researchers also customize satellite images for method development. This shows that RSSISR needs real-world data validation and applications for practical usage.

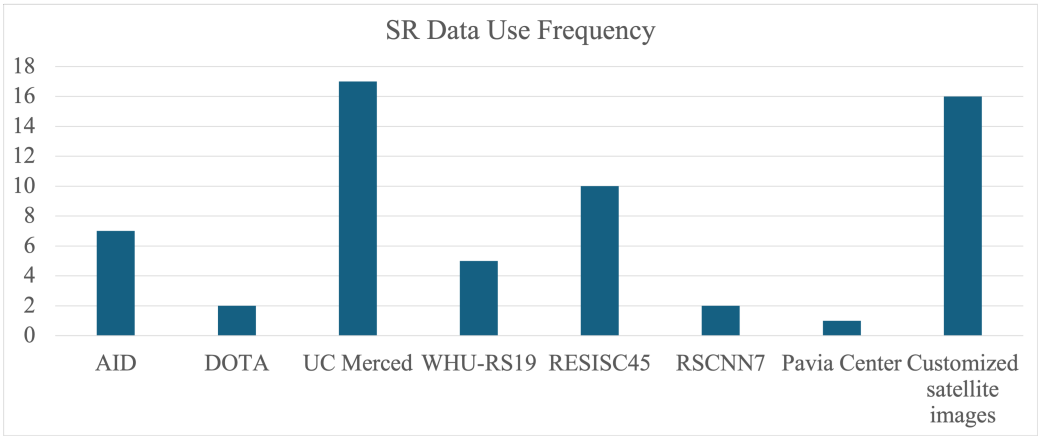

Figure 3: Data Use Frequency. We only summarized papers cited in this study.

4. Deep Learning Architectures

Based on the characteristics and limitations of commonly used RSSISR datasets, it becomes obvious that robust and adaptable deep learning architectures are essential. This section reviews the impact of DL architectures on RSSISR, presenting developing history and key architectures, including

Convolutional Neural Network (CNN) architectures, Attention-based architectures, Generative Adversarial Network (GAN) architectures, Diffusion architectures, and Graph Neural Network (GNN) architectures. As the complexity and volume of RS data increase, so does the need for DL architectures capability.

4.1. CNN Architectures

Originating in the 1980s and gaining popularity in the 1990s for handwritten digit recognition, CNNs is predominant in computer vision until now, with applications such as image classification, semantic segmentation, and object detection [26, 27].

A typical CNN architecture consists of a sequence of convolutional layers, followed by nonlinear activation functions (e.g., ReLU, Tanh), pooling operations for spatial downsampling, and sometimes normalization layers (e.g., BatchNorm, LayerNorm). However, as network depth increases, plain CNNs often suffer from vanishing gradients and performance degradation.

To address these limitations, Figure 4 shows a classic CNN-based block, which is ResNet block [28]. Based on the a plain CNN network, one can insert shortcut connections to turn the vanilla CNN block into its residual version [28]. The identity shortcuts can be directly used when the input and output are of the same dimensions (solid line shortcuts in Figure 4 (A)). In this way, residual networks are easier to optimize, and can boost accuracy from considerably increased depth. CNNs can perform effective feature extraction to identify spatial patterns of images but are limited in receptive fields due to the size of convolution operations.

Several landmark CNN-based models have been proposed to advance SISR. SRCNN [8] is one of the earliest DL-based SR models, introducing a simple three-layer convolutional architecture that significantly improved performance over traditional interpolation-based and reconstruction-based methods. VDSR [9] builds on this by employing deeper residual learning (20 layers), allowing for faster convergence and improved accuracy. EDSR [10] further optimizes the residual structure by removing batch normalization, enabling even deeper networks and achieving better results on many benchmarks. In contrast to supervised methods, ZSSR [29] is a zero-shot framework that trains on internal image statistics at test time, offering flexibility in real-world scenarios with unknown degradation or limited training data. These CNN-based models have laid a strong foundation for transferring SR techniques into RS imagery. For instance, GEOSR [30] integrates

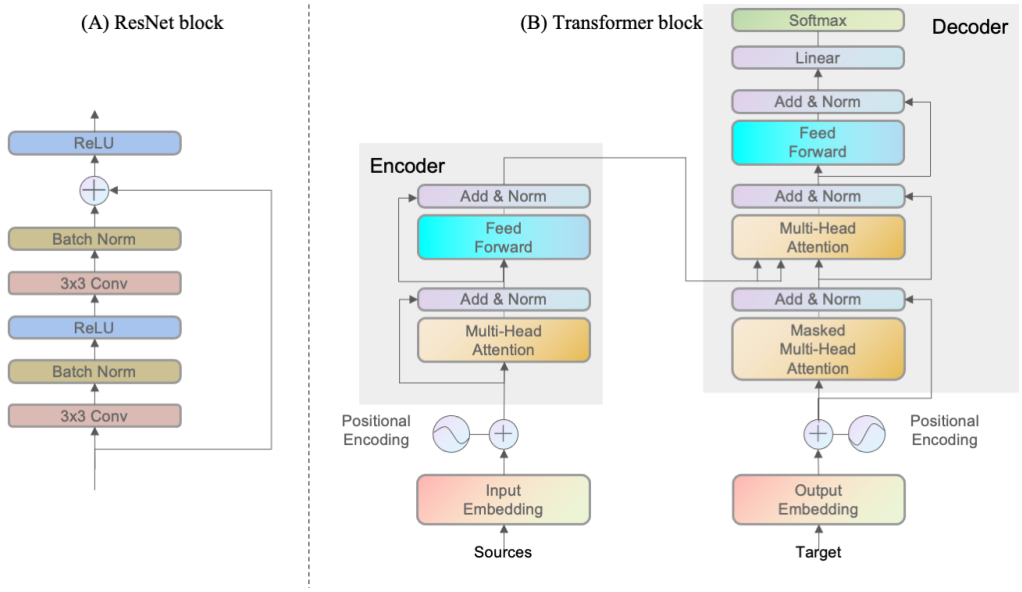

Figure 4: Comparison of ResNet block (A) and Transformer block (B).

and adapts these classical models for RS tasks, demonstrating the utility of these architectures in domain-specific SR.

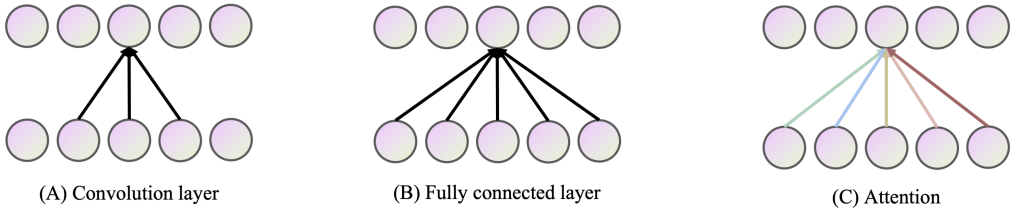

Figure 5: Comparison of convolution layer (A), fully connected layer (B) and attention mechanism (C). Attention weights are dynamic and input-dependent (colorful lines), unlike the fixed learned weights in convolution and fully connected layers (black lines).

4.2. Attention-based Architectures

Vision Transformer (ViTs) typically include several Transformer blocks [31], where each consists of the input embedding, the positional encoding, and the transformer encoders, as shown in Figure 4(B). ViT first deconstructs an image into a series of 16×16 pixel patches. Each patch is mapped into a

vector through input embedding. To help the model keep track of patch sequence, a positional encoding is added.

Then what makes ViTs unique and exceptional is the self-attention mechanism for capturing global and local context [32]. This enables ViTs to process global data representations in parallel, drastically improving computational efficiency. As shown in Figure 5, in self-attention, each element of the input sequence interacts with every other element with dynamic weights. This interaction is facilitated by the computation of Query, Key, and Value vectors for each input token. The query asks: "what am I looking for?" The key answers: "what do I have?" The value carries the information needed if there's a match. The dot product of Query and Key vectors determines the attention scores, which are then normalized using the softmax function to compute the weights. These weights are used to aggregate the Value vectors, producing the self-attention output. This enables the ViTs to capture long-range relationships in the image. On the contrary, in convolution layers and fully connected layers, weights are fixed, as shown in Figure 5(B)(C). Compared to CNNs, ViTs are more flexible in learning global context, especially useful when trained with large datasets. Examples include the Multi-scale Attention Network (MAN) [33], which integrates attention modules across multiple scales to enhance the representation of the network, therefore achieving superior performance on many SR benchmarks.

4.3. GAN Architectures

A GAN consists of two models: a discriminator and a generator [34], as shown in Figure 6(A). A discriminator estimates the probability of a given sample coming from the real dataset. It works as a critic and is optimized to distinguish the fake samples from the real ones. A generator outputs synthetic samples given a noise variable input (shown as latent code z in Figure 6(A)). It is trained to capture the real data distribution so that its generative samples can be as real as possible, or in other words, can enforce the discriminator to distinguish image details. These two models compete against each other during the training process: the generator is trying hard to trick the discriminator, while the discriminator is trying hard not to be cheated. This competitive game between two models motivates both to improve their functionalities.

This approach was first introduced by Super-Resolution GAN (SRGAN) [11], pushing the output towards more realistic textures and sharper details.

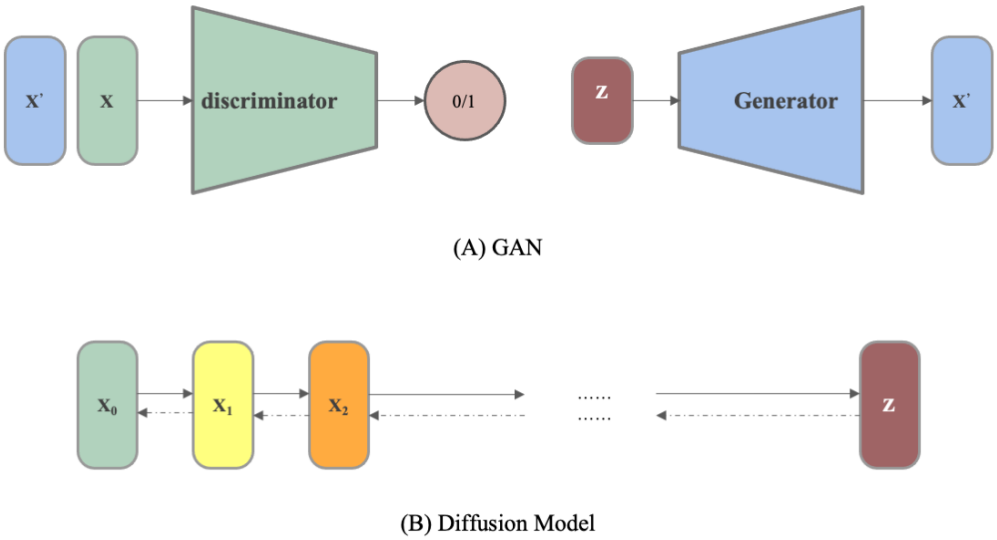

Figure 6: Overview of typical generative models: (A) GAN and (B) Diffusion model.

Building on this, ESRGAN (Enhanced SRGAN) [12] further refines the SRGAN architecture with residual-in-residual dense blocks and a perceptual loss. These improvements lead to both better perceptual quality and higher quantitative metrics. More recent works have extended ESRGAN to domain specific tasks—for instance, applying ESRGAN to infrared (IR) image super-resolution [35], demonstrating its adaptability to diverse data modalities beyond the natural color (RGB) domain.

4.4. Diffusion Architectures

Diffusion models define a Markov chain of diffusion steps to slowly add random noise to degrade original data and then learn to reverse the diffusion process to construct desired data samples from the noise [36], as shown in Figure 6 (B). the process starts with a clean image x_0, and adds small amounts of noise over multiple steps to create increasingly noisy versions $x_1, x_2, \ldots$, until reaching pure noise z. During training, the model learns to reverse this process to transform random noise z back into a clean image by predicting and removing noise at each step $\ldots, x_2, x_1, x_0$. Since diffusion models rely on diffusion steps to generate samples, it can be quite expensive in terms of time and compute. Methods have been proposed to make the

process much faster, such as DDIM, but the sampling process is still slower than GAN [37].

To address this, new methods aim to accelerate inference. Recently, studies demonstrate that the diffusion prior, embedded in Stable Diffusion [38], can be applied to various downstream content creation tasks, offering adaptability and competitive performance [39]. For example, StableSR [40] adds trainable spatial feature transform layers to exploit Stable Diffusion priors. Even if escaping from training diffusion process from scratch, StableSR still needs 200 steps during inference. One-Step Effective Diffusion network (OSEDiff) [41] introduces a pretrained text-to-image model as generator and regularizer, and simplifies the inference step from 200 to 1. Moreover, the Pixel-level and Semantic-level Adjustable Super-resolution (PiSA-SR) [42] is a dual approach, characterizing pixel-level and semantic-level information, achieving results in both quality and efficiency in 1 diffusion step. In RS area, DiffusionSat [43] is a notable example. It leverages RS image metadata (longitude, latitude, ground-sampling distance, cloud cover, timestamp) as additional embeddings, enabling effective RSSR and inpainting.

4.5. Graph Neural Network Architectures

Graph Neural Networks (GNNs) have significant potential for processing graph-structured data, such as transportation networks, due to their ability to model irregular, non-Euclidean structures [44]. GNNs are designed to account for the relationships between nodes and the transmission of information through edges. These models aggregate information from neighboring nodes to predict values for nodes or edges. Due to the ability of modeling long-range dependencies, GNNs are introduced to image restoration tasks. As an example, Internal Graph Neural Network (IGNN) [45] constructs a graph between similar patches across different image scales, and then aggregates information from a LR version to guide HR reconstruction.

5. Deep Learning in RSSISR

In this section, we categorize existing deep learning-based RSSISR methods into four paradigms: supervised learning, unsupervised learning, self-supervised learning, and training-free learning, based on their reliance on labeled data and learning strategies. We then discuss commonly used evaluation metrics to assess the effectiveness and efficiency of these methods, followed by real-world application scenarios.

5.1. Supervised Learning

5.1.1. CNN

The msiSRCNN [46] pioneers its application of RSSR in multispectral RS data; RS-SRCNN [47] is also a fine-tuned SRCNN version for super-resolution, image denoising, and haze removal. RS-SR [49] adapted VDSR in Pleiades multispectral images, demonstrating the better overall performance that both Bicubic and SRCNN. After that, more CNN architectures have been proposed for RSSR. In this review, we provide detailed information on supervised CNN methods in Table 2, and summarize five types of commonly used modules.

Residual Learning: As mentioned previously, ResNet blocks are essential and widely used in CNN architectures. For example, Remote Sensing Deep Residual Learning (RS-DRL) [48] and Wide Remote Sensing Residual network (WRSR) [53, 54] both incorporate the residual blocks. Moreover, Multi-Losses Function Network (MLFN) [50] proposes two branch networks: one is residual network, and the other is loss network, including pixel-wise spatial loss and spectral loss to drive the learning of the entire reconstruction model. With the advanced computing power, residual learning with more depth has also been proposed, such as Improved Deep Recursive Residual Network(IDRRN) [59]. Also, Multi-residual U-Net [60] enhances the original U-Net by incorporating residual connections at multiple levels. Furthermore, Enhanced Deep Pyramidal residual networks for Super-Resolution (EDPSR) [66] is designed to leverage a hierarchical pyramidal architecture with residual learning, making it well-suited for high-quality image reconstruction.

Wavelet and Shearlet Transform: Wavelet Transform-based methods decompose an image into multiscale frequency components, learning high-frequency details. On the other hand, Shearlet Transform-based methods utilize the shearlet transform to represent images in multiscale and multi-directional frequency components, allowing for a more accurate approximation of image structures, such as edges.

The Wavelet Transform Combined with the Recursive Resnet (WTCRR) [51] and the Discrete Wavelet Transform SR (DWTSR) [52] apply the wavelet transform to separate LR inputs into multiple subbands (LL, LH, HL, HH). These components are then processed using deep residual networks to estimate their high-resolution (HR) counterparts. The final super-resolved image is reconstructed through inverse transforms. On the other hand, Deep Shearlet Residual Learning Network (DSRLN) [55] adopts a dual-branch design to

separately process high- and low-frequency shearlet coefficients. In summary, wavelet-based methods provide efficient frequency-aware decomposition, and shearlet-based methods offer directional sensitivity and edge representation, making them particularly suitable for preserving fine structural details.

Markov Random Field: Markov random field (MRF) is a method that models how each pixel in an image is influenced by the pixels around it. The MRF and two-dimensional phase congruency-based single-image super resolution reconstruction method (MRF-SRR) [58] uses phase congruency to compute edge and texture maps, allowing for more accurate structural feature representations. The MRF then models the spatial dependencies for HR reconstruction. By considering both the spatial relationships between pixels and the structural patterns in the image, MRF-SRR helps preserve fine edge details and reduce unwanted artifacts. In simple terms, MRF-based methods make the output image look cleaner and more accurate by referring textures and edges.

Channel Fusion: Channel fusion refers to the process of integrating information from different spectral bands. Channel fusion techniques learn the importance of individual channels or combine multi-spectral features to improve reconstruction accuracy, particularly for texture, edge, and detail recovery. For example, the Dual-branch Multiscale channel fusion Unfolding Network (DMUNet) [61] reconstructs texture and edge information separately, supported by a multiscale channel fusion module that enables cross-scale and cross-channel information exchange. The Selective Channel Processing Network (SCPN) [62] designed a dynamically learnable feature map using a channel selection matrix in the training phase. During inference, only these selected channels are processed, reducing computation. However, if the channel selection strategy is not robust and well-trained, the module might consistently prioritize some channels while neglecting others, leading to sub-optimal feature extraction. The Spatial and Channel Aggregation Network (SCAN) [63] unifies spatial and channel attention in a structured pipeline. It starts with standard and dilated convolutions, then employs Spatial and Channel Aggregation modules composed of Temporal-Spatial Self-Attention (TSSA) and Channel Attention (CA). TSSA captures spatial dependencies; CA captures inter-channel relationships.

Multi-task Learning: More recently, multi-task learning is often applied to jointly handle related tasks or applications, enabling shared representation learning and mutual performance enhancement [87]. In RSSR, MTL is often applied to jointly handle SR and other low-level vision tasks

such as deblurring or denoising. For example, the Joint Super Resolution and Deblurring Network (JSRDNet) [65] can perform joint SISR and image deblurring on low-resolution blurry inputs. In the network, one branch is responsible for SR feature extraction, which uses residual blocks to extract detailed structures for resolution enhancement. Another branch is deblur feature extraction, which uses hierarchical convolutional layers to minimize blur artifacts. The extracted features from both branches are then merged (e.g., concatenation, element-wise multiplication, and gating) to fuse spatial and channel information effectively. JSRDNet demonstrates the advantage of multi-task learning by providing an end-to-end solution that generates sharp, high-quality images from blurry low-resolution inputs.

5.1.2. Attention

Attention Module Integration: Plugging attention modules into other architectures can enhance features selectively. For example, Self-Attention Fusion (SAF) module can be placed after the CNN backbone and before the up-sampling module [68]. SAF could combine spatial attention and channel attention in parallel and dynamically using learnable parameters. Attention modules (spatial/channel/self-attention) are integrated as lightweight add-ons to previous models with minimal architectural disruption.

Multi-branch Architecture: Multi-branch architectures aim to decouple different paths to specialize in related tasks or extract different feature types. For instance, the Dual-resolution Local Attention unfolding Network (DLANet) [70] flattens extracted features, processes them independently along row-wise and column-wise directions. These attention maps are then folded back into the original spatial domain to aggregate locally enhanced features. Two-Branch Multiscale Residual Attention Network (TBMRA) [72] begins with a convolutional layer followed by multiple TBMRA blocks. Each TBMRA consists of two parallel branches with 3×3 and 5×5 kernels, designed to capture low and high level spatial patterns. These branches are further enhanced by efficient channel attention and spatial attention mechanisms. After attention refinement, the feature maps are concatenated to generate the high-resolution output.

Progressive Feature Refinement: This design pattern leverages multistage modules to iteratively refine feature representations. It is often paired with attention or transformer structures to enhance contextual understanding and reconstruction fidelity. For instance, Transformer-based Enhancement Network (TransENet) [69] introduces a multistage enhancement framework

using transformer encoders and decoders. The encoder embeds multilevel features, while the decoder fuses them for super-resolution reconstruction. TransENet can be integrated into conventional SR networks to boost performance. It achieves consistent gains over SRCNN, with up to +1.26 dB in PSNR and +0.0370 in SSIM on the UC Merced dataset at scale ×3, and +0.98 dB in PSNR and +0.0348 in SSIM on AID at scale ×4. Another example is Pyramid Vision Transformer-Residual Feature Aggregation Network (PVT-RFANet) [71], which has a three-part structure: a head with a 1×1 convolution and Residual Feature Attention, a trunk composed of Enhanced Spatial Attention, Channel Attention, and Pyramid Vision Transformer (PVT) blocks, and a final reconstruction module. PVT blocks incorporate patch embeddings, position embeddings, and spatial reduction to capture multiscale global context. This progressive attention-transformer fusion yields the highest reported performance (PSNR: 22.01 dB, SSIM: 0.50) but with slower inference (3.25–3.32 images/sec) compared to traditional models like SRCNN (6.17 images/sec), illustrating a trade-off between reconstruction quality and computational speed.

5.1.3. GAN

Recent studies in RSSISR extend GAN-based SR frameworks with architectural simplifications, progressive refinement strategies, attention mechanisms, and saliency-guided feedback designs. Below, we group recent GAN-based SR models by four design patterns:

Optimized GANs: Transferred Generative Adversarial Network (TGAN) [73] improves upon SRGAN by simplifying the generator architecture—removing batch normalization—to reduce computational overhead. The network is pretrained on DIV2K and fine-tuned on remote sensing data. This streamlined architecture enhances visual performance and speeds up training, while mitigating normalization-induced artifacts. RS-ESRGAN [75] builds upon ESRGAN by introducing Residual-in-Residual Dense Blocks (RRDBs) without batch normalization, along with residual scaling and support for four-band imagery. These modifications enhance computational stability and enable deeper feature propagation. It significantly improves PSNR, SSIM over SRGAN, demonstrating better spatial and spectral reconstruction.

Progressive Feature Refinement: This design leverages multi-scale feature representations, starting from a low-level feature and adding new blocks that model increasingly fine details as training progresses [88]. For instance, Multiple Scale Super Resolution GAN (MSSRGAN) [74] integrates

GAN learning with multi-scale progressive training. The generator has three upsampling stages, each increasing resolution by a scale factor of 2x. On the WorldView-3 dataset, MSSRGAN surpasses both SRGAN and MSSRNet in PSNR and SSIM, showcasing superior structural preservation and fidelity.

Saliency-guided GANs: A saliency map is a visual representation that highlights the most important or attention-worthy regions in an image — areas that are most likely to be relevant for the SR task. SD-GAN [77] adopts dense feature extraction and uses saliency as a static preprocessing module to guide learning. Although less dynamic than feedback-based methods, it maintains high expressivity through iterative refinement and paired discriminators. Saliency-Driven Feedback GAN SDFBGAN [76, 79] introduces a saliency-driven feedback mechanism using paired-feedback blocks (PFBBs) and recurrent structures. Saliency maps reflect texture complexity and guide the generator in restoring regions with varying levels of detail.

Attention-enhanced GANs: Super Resolution Attention GAN (SRAGAN) [80] fuses attention mechanisms with residual learning to significantly improve performance over SRGAN and ESRGAN. It achieves SOTA PSNR and SSIM at 2× and 4× scales while maintaining low ERGAS scores. The model balances effectiveness and moderate computational complexity. Residual Balanced Attention generator with UNet discriminator (RBAN-UNet) [1] models realistic degradation with blur kernels and integrates both spatial and channel attention in a residual generator. Using a UNet discriminator for pixel-wise realism, RBAN-UNet achieves superior results over SRGAN on the AID dataset. GAN with a joint-attention module JOA-GAN [82] proposes a joint attention module combining Efficient Channel Attention (ECA) and Integrated Spatial Attention (ISA) to guide both the generator and discriminator. The model further incorporates multi-scale ERRDB blocks in the generator and relative discrimination in the discriminator. Across UC Merced, NWPU-RESISC4, and AID datasets, JOA-GAN consistently outperforms SRGAN and ESRGAN in PSNR, SSIM, and perceptual quality, demonstrating robustness across scale factors (4×, 8×).

5.1.4. Diffusion Models

Diffusion models have recently emerged as the most powerful generative paradigm by modeling the data distribution through iterative denoising processes [37, 39]. For example, Detail Complement mechanism (DMDC) [83] introduces a two-stage approach, where the model is first trained on randomly masked LR images to learn to recover missing patches. In the second

stage, SR is applied to the intermediate output to refine spatial resolution. Conditional guidance based on unmasked reference images allows DMDC to iteratively restore fine details. Quantitative results across Potsdam and Vaihingen datasets demonstrate consistent superiority over DDPM, especially in SSIM and BRISQUE metrics, affirming its ability to enhance perceptual quality. TESR (Two-stage approach for Enhancement and super-resolution) [84] combines SwinIR-based structural enhancement with a U-Net-based diffusion model for fine detail restoration. The first stage extracts shallow and deep features, while the second stage applies iterative denoising via diffusion to refine textures. TESR outperforms SRCNN and TransENet across all scales, with significant margins in PSNR and SSIM, highlighting the effectiveness of combining transformer-based learning and generative diffusion modeling. Dual-Diffusion [85] introduces two parallel DDPM-based modules: one for kernel estimation and another for image reconstruction. The kernel predictor learns the mapping between latent and degradation kernels, while the reconstructor uses this information to generate high-quality HR images from LR inputs. This dual-path design enhances robustness and adaptability to diverse degradation conditions.

5.1.5. GNN

GNNs are an alternative model design to CNNs by explicitly modeling relationships between non-local image patches, making them particularly useful for handling scale-variant objects and self-similar patterns in RS imagery. Dual Learning-based Graph Neural Network (DLGNN) [86] addresses the limitations of single-level feature representation by aggregating cross-scale patch similarities via GNN-based message passing. A dual learning strategy is employed, which jointly learns both the forward mapping from LR to HR and a reverse mapping from HR to LR, improving generalization and reconstruction fidelity. Evaluations on the Massachusetts Roads and 3K VEHICLE SR datasets show that DLGNN achieves the highest PSNR and SSIM compared to methods like TransENet and RDBPN, confirming its effectiveness in leveraging self-similarity and scale-awareness.

5.2. Unsupervised Learning

Unsupervised CNN-based methods for SR aim to enhance spatial details and structural integrity without requiring paired low- and high-resolution image datasets.

5.2.1. CNN

UDGN (Unsupervised Deep Generative Network) [89] enhances SISR performance by incorporating edge information from the gradient map of the LR image. The architecture fuses the original image and its Sobel-filtered gradient map, followed by multi-stage processing that includes feature extraction, residual learning, and pixel-shuffle upsampling. While UDGN demonstrates promising performance in preserving textures and edges, its effectiveness diminishes in complex regions, particularly around object boundaries. However, UDGN is unsuitable for complex image features inside edges.

Fusion-based framework for Unsupervised Single-Image Super-Resolution (FUSISR) [90] is a Self-Fusion architecture. It first decomposes LR image into low-frequency (I_1), mid-frequency (I_2), and high-frequency (I_3) components, which capture smoothing features, textures, and edge details, respectively. These components are fused through a learnable weighting mechanism and processed by an autoencoder to reconstruct the final SR image. Training is guided by a no-reference image quality metric (NR-IQM), enabling fully unsupervised learning. Comparative evaluation using BRISQUE, NIQE, and PIQUE metrics shows FUSISR excels at preserving perceptual quality and naturalness, particularly at high scaling factors, while ZSSR remains stronger at edge detail preservation.

5.2.2. GAN

Unsupervised GAN [91] bypasses the need for HR labels by generating SR images from interpolated LR inputs and comparing them to downsampled SR images in the discriminator. This setup enables adversarial learning without HR supervision. The model demonstrates excellent performance at $2\times$ scale and competitive performance at $4\times$, effectively balancing spatial and spectral fidelity.

Closed-Loop Network for Single Infrared Remote Sensing Image Super-Resolution (CLN4SR) [92] proposes a dual-generator closed-loop structure: one network performs downsampling, and the other performs super-resolution. Their interconnection allows for self-supervised learning by enforcing consistency between the input and reconstructed images. In infrared SR tasks, CLN4SR outperforms ZSSR [29], achieving superior PSNR (40.1415 vs. 38.1762) and SSIM (0.9645 vs. 0.9597), validating its ability to preserve both fine structural details and overall image quality under real-world degradation.

5.3. Self-supervised Learning

Self-supervised learning (SSL) for SR leverages intrinsic structures or degradation properties within unpaired data to guide model learning, eliminating the need for ground truth HR images.

5.3.1. CNN

Degradation Guided Adaptive Network (D2U) [93] learns degradation-aware representations via a contrastive learning framework. Assuming spatially uniform degradation across an image, embeddings from the same image are treated as positive pairs, while those from different images are negatives. The resulting degradation embeddings are integrated with CNN-extracted features and passed through a dual-wise modulation network comprising Dual-wise Modulation Groups (DMGs). These DMGs include Dual-wise Modulation Blocks that refine features via both additive and multiplicative adjustments. Upsampling is performed with pixel shuffle to reconstruct the HR output. On the AID dataset, D2U significantly outperforms the unsupervised baseline ZSSR, achieving a PSNR of 40.1415 and SSIM of 0.9645, compared to ZSSR's 38.1762 and 0.9597, respectively. This demonstrates the effectiveness of combining self-supervised degradation modeling with adaptive modulation for SR under unknown degradations.

5.3.2. Attention

Cross-dimension Attention guided Self-supervised remote sensing single-image super-resolution method (CASSISR) [94] introduces a Cross-Dimension Attention Module (CDAM) that jointly models channel and spatial dependencies to enhance internal feature learning. The self-supervised training pipeline first downsamples the LR input to 1/s×LR (where s is the upscaling factor) and trains the network to reconstruct the original LR image. After convergence, the full LR input is used to infer the HR output. This strategy encourages learning from internal patterns and supports structure-aware reconstruction. CASSISR achieves best performance on multiple datasets over all compared methods. For example, on the RSSCN7-Blur dataset, it reaches a PSNR of 30.01 and SSIM of 0.8142, outperforming SRCNN by +2.15 dB and +0.0831 in SSIM. On WHU-RS19-Blur, CASSISR achieves 32.66/0.8831 versus SRCNN's 29.08/0.7936, indicating significant improvements in blur reduction and structural preservation. However, its reliance on internal image redundancy may limit performance in scenes lacking repetitive texture.

5.4. Training-free Learning

Training-free methods bypass the need for large annotated datasets and explicit training phases by leveraging the inherent structure of neural networks as image priors. These approaches are particularly valuable for data-scarce domains like hyperspectral imaging.

Deep Hyperspectral Prior (Deep HS Prior) [95] extends the concept of deep image priors to hyperspectral image restoration tasks. The method evaluates both 2D and 3D convolutional architectures: the 2D model independently processes each spectral band, capturing spatial features but ignoring inter-band correlations, while the 3D model jointly models spatial and spectral dependencies through volumetric convolutions.

Despite the 3D model's theoretical advantage in modeling spectral-spatial relationships, the 2D architecture consistently outperforms it in practice. It achieves higher mean PSNR (MPSNR: 33.67 vs. 32.31) and lower Spectral Angle Mapper error (SAM: 4.211 vs. 4.692), while the 3D variant exhibits slightly better MSSIM (0.967 vs. 0.945). Additionally, the 2D model is more computationally efficient, requiring significantly less memory and time. These results suggest that 2D architectures, although simpler, are often sufficient for effective hyperspectral restoration, while 3D convolutions remain theoretically important for capturing spectral priors.

5.5. Evaluation Metrics

5.5.1. Model Performance

Peak-Signal-Noise-Ratio (PSNR): an image quality evaluation index based on the mean square error (MSE) between the ground-truth HR and the reconstructed SR, which can be formulated as:

$$\mathrm{MSE}(HR, SR) = \frac{1}{mn} \sum_{p=1}^{m} \sum_{q=1}^{n} \left(HR(p,q) - SR(p,q)\right)^2 \tag{4}$$

$$\mathrm{PSNR}(HR, SR) = 10 \log_{10} \left(\frac{255^2}{\mathrm{MSE}(HR, SR)} \right) \tag{5}$$

Where m and n are the width and the height of the image, respectively; p and q index the pixel location at row and column, respectively. PSNR is an essential evaluation method in SR tasks, and an image with a larger PSNR has a higher reconstruction quality in the pixel level. However, PSNR

has a weakness in measuring fidelity signal as it fails to capture structural or perceptual differences between images [96].

Structural Similarity Index Measure (SSIM): used to evaluate the structural similarity between the ground-truth HR and the reconstructed SR image, and it can be calculated as:

$$\text{SSIM}(HR, SR) = l(HR, SR) \cdot c(HR, SR) \cdot s(HR, SR) \tag{6}$$

$$l(HR, SR) = \frac{2\mu_{HR}\mu_{SR} + C_1}{\mu_{HR}^2 + \mu_{SR}^2 + C_1} \tag{7}$$

$$c(HR, SR) = \frac{2\sigma_{HR}\sigma_{SR} + C_2}{\sigma_{HR}^2 + \sigma_{SR}^2 + C_2} \tag{8}$$

$$s(HR, SR) = \frac{\sigma_{HRSR} + C_3}{\sigma_{HR}\sigma_{SR} + C_3} \tag{9}$$

where μ_{HR} and μ_{SR} represent the mean value of HR and SR, respectively, σ_{HR} and σ_{SR} denote the standard deviation of HR and SR, respectively, $\sigma HRSR$ is the covariance between HR and SR, and C1, C2 and C3 are constants. A higher value of SSIM indicates a higher quality. SSIM is designed to account for structural similarity and perceptual differences, providing better alignment with human visual judgment than PSNR [96].

Spectral Angle Mapper (SAM): it calculates the angle between two images by computing the dot product divided by the 2-norm of each image. This index indicates higher similarity between images as it approaches zero:

$$SAM(HR, SR) = \arccos\left(\frac{HR \cdot SR}{\|HR\|_2 \|SR\|_2}\right) \tag{10}$$

Correlation Coefficient (CC): calculates the average Pearson correlation coefficient across all bands between SR and HR:

$$\text{CC}(SR, HR) = \frac{1}{n_{\text{bands}}} \sum_{j=1}^{n_{\text{bands}}} \frac{\sum_{k=1}^{n}(SR_{kj} - \bar{SR}_j)(HR_{kj} - \bar{HR}_j)}{\sqrt{\sum_{k=1}^{n}(SR_{kj} - \bar{SR}_j)^2}\sqrt{\sum_{k=1}^{n}(HR_{kj} - \bar{HR}_j)^2}} \tag{11}$$

Where n is the number of pixels in each band, and $\bar{SR}_j$, $\bar{HR}_j$ are the means of band j. Both SAM and CC are designed for spectral similarity evaluation, and are commonly used in multispectral and hyperspectral image SR, where spectral integrity across multiple channels is essential.

Normalized Root-Mean-Square Error (NRMSE): can be computed as follows, and the smaller the value of NRMSE is the better quality the reconstructed HR image has:

$$\mathrm{NRMSE}(SR, HR) = \frac{\sqrt{\mathrm{MSE}(SR, HR)}}{255} \tag{12}$$

Erreur Relative Globale Adimensionnelle de Synthese (ERGAS): is put forward to measure the quality of reconstructed HR images by taking the scaling factor into consideration, and it can be formulated as:

$$\mathrm{ERGAS}(SR, HR) = 100 \cdot \frac{1}{s} \cdot \sqrt{\frac{1}{c} \sum_{i=1}^{c} \left(\frac{\mathrm{RMSE}_i}{\mu_{HR_i}} \right)^2} \tag{13}$$

Where s represents the scale factor, c denotes the channel number of the image. The smaller the value of ERGAS, the better the quality of the reconstructed HR image.

5.5.2. Model Efficiency

Model efficiency is an essential consideration in RSSISR, especially for real-time or resource-constrained applications such as onboard satellite processing. Two widely used metrics to evaluate the computational efficiency of super-resolution models are Floating Point Operations (FLOPs) and the number of trainable Parameters (Params).

FLOPs: FLOPs refer to the total number of operations (multiplications and additions) required during a single forward pass of the network. For a standard 2D convolutional layer, the FLOPs can be estimated as:

$$\mathrm{FLOPs} = 2 \cdot C_{\mathrm{in}} \cdot C_{\mathrm{out}} \cdot K^2 \cdot H_{\mathrm{out}} \cdot W_{\mathrm{out}} \tag{14}$$

where C_{in} is the number of input channels, C_{out} is the number of output channels, K is the kernel size , and $H_{\mathrm{out}}, W_{\mathrm{out}}$ is the height and width of the output feature map. The factor 2 accounts for both multiplication and addition operations per output pixel.

Trainable Params: The number of learnable parameters determines the model size and storage cost. For a single convolutional layer, it is computed as:

$$\mathrm{Params} = C_{\mathrm{in}} \cdot C_{\mathrm{out}} \cdot K^2 + C_{\mathrm{out}} \tag{15}$$

where the last term accounts for the bias in each output channel (if applicable).

These metrics help quantify the trade-off between model complexity and computational efficiency, especially when comparing lightweight designs with more elaborate architectures involving self-attention, diffusion, or graph-based modules.

5.6. Downstream Applications of SR Images

Currently, most RSSISR methods are developed in an application-independent manner, often without case-specific or real-world evaluations. However, practical deployments demand robust models designed to diverse geospatial scenarios. Critical applications such as agricultural forecasting, flood mapping, and urban monitoring require SR methods that adapt to domain-specific features and geographical contexts. For instance, land cover and land use applications often rely on urban-focused training data, emphasizing features such as physical infrastructure and built environments. Wildfire damage assessments, by contrast, prioritize forested and rural regions, while flood mapping necessitates accurate water-body delineation and topographic sensitivity during training phases.

An example of application-aware method is JSRDNet [65], which demonstrates the utility of RSSR for land cover classification. Post-SR classification using supervised models reveals that JSRDNet outperforms seven other SR methods, achieving the highest classification accuracy (86.94%) and Kappa coefficient (0.8006), surpassing the next best DASR (85.6%, 0.7799). These improvements highlight its robustness for land-use mapping. Another example is the Dual Super-Resolution (DSR) framework [97], which jointly integrates SISR and Semantic Segmentation Super-Resolution. DSR significantly outperforms interpolation baselines with a PSNR of 35.422 and SSIM of 0.776. Experiments on VDSR [98, 99] also show that SISR can enhance Global Navigation Satellite System Reflectometry (GNSS-R) data and sea surface temperature reconstruction across two resolution transitions (15km→5km and 5km→1km). Multi-Residual U-Net [60] is a method designed to super-resolve MODIS land surface temperature data from 1,000m to 250m resolution, enabling finer-scale climate and environmental analyses. BLiSR [67] also validates SR's relevance by enhancing object detection (mAP50: 92.8%) and semantic segmentation (mIoU: 66.3%) across NWPU and Vaihingen datasets. These examples confirm the value of domain-adapted

SR methods in improving the performance of RS-related tasks across diverse applications.

6. Challanges

6.1. Training Data Quality

The scarcity and domain specificity of high-quality training data remains a critical challenge in RSSR. Despite the availability of open-access Earth observation datasets such as Landsat8 and Sentinel-2 [100], difficulties persist in curating real-world datasets that meet the spatial, spectral, and temporal requirements of SR tasks. Many current SR models still rely on synthetically generated HR-LR pairs using bicubic downsampling, which fail to reflect the complexity of real-world degradation processes, particularly for modalities like hyperspectral and infrared imagery [92]. This mismatch reduces the applicability of such models in practical settings [7].

To address this issue, recent methods have begun to explore blind SR under unknown and variable degradations. However, the lack of real-world benchmark datasets and the difficulty of estimating reasonable degradation kernels still hinder progress [93]. Moreover, accurate HR-LR pair generation from satellite platforms (e.g., Sentinel-2, PlanetScope) is non-trivial. As [101] points out, SR dataset construction in RS must consider multi-temporal alignment, atmospheric correction, co-registration, histogram matching, and patch selection based on perceptual quality metrics (e.g., PSNR, SSIM). These steps are essential to ensure spatial, temporal, and radiometric consistency for supervised learning. In summary, training data scarcity and the mismatch between synthetic degradation assumptions and real-world conditions continue to challenge the development of robust RSSR models.

6.2. Generalization and Transferability

Another key limitation in RSSISR models is the limited generalization and transferability across geographic regions and sensor domains. First, models trained on data from one geospatial region often fail to perform well in other regions due to variations in land cover patterns, atmospheric conditions, and data acquisition time [102]. This geographic specificity poses a significant barrier to cross-geospatial regions, and global-scale deployment. Moreover, the limited generalizability of models across geographic regions

and sensor types also introduces challenges. In particular, differences in spatial resolution and sensor characteristics introduce distribution shifts. Although models may be trained at one target scale (e.g., Landsat-8 30 m $\rightarrow$ Sentinel-2 10 m), they often generalize poorly to data from other target scales (e.g., Sentinel-2 10 m $\rightarrow$ PlanetScope 3 m, MODIS 500 m $\rightarrow$ Landsat-8 30 m). RS data at different resolutions and from different sensors have different statistical distributions due to radiometric and geometric discrepancies [103]. While some studies have made initial attempts to address generalization, such as ISRGAN's cross-region and cross-sensor evaluation [102], the experiments were constrained to only two locations within one country and in moderate-resolution (Landsat 8) sensors, which may not fully capture the complexity of global-scale variation in land cover, sensor characteristics, and degradation processes. As such, these discrepancies hinder model scalability.

6.3. High Scale Factor Super Resolution

While most current RSSISR models demonstrate satisfactory results at moderate scale factors (e.g., 2×, 3×, 4×), their performances can be inferior at higher scale factors (e.g., 6×, 8×, 10×, even 32×) [15]. This limitation arises from the increasing difficulty in recovering high-frequency details when the low-resolution input lacks sufficient contextual information. As the scale factor increases, the degradation gap widens, leading to artifacts such as over-smoothing or checkerboard effect [104]. Moreover, many public RS datasets cannot provide reliable ground truth at ultra-high resolution, making supervised learning at higher scales particularly challenging. Recent efforts, such as spectra-guided generative adversarial networks (SpecGAN), have reported success with extremely large scale factors (e.g., 32×) in RSSR, yet such results are based on synthetic degradations using interpolation-based downsampling [104], which cannot reflect the complexity of real-world degradation. As such, the applicability to operational scenarios remains to be further validated.

6.4. Lack of Application-Oriented Design

Among all research papers we examined in this review, most methods are validated in application-independent scenarios, often using simulated LR images, resulting in overly optimistic conclusions [105]. In addition, results on simulated benchmarks have limited relevance to real downstream applications. As such, it is unclear whether existing SR methods that can generate

realistic and fidelity-preserved datasets for downstream applications (e.g., crop yield prediction, or active fire detection).

A comparative analysis [106] highlights the sensitivity of SRGAN models to the training and application data domains. In particular, models trained on agricultural versus urban imagery exhibit similar PSNR/SSIM performance on non-satellite benchmarks, yet diverge notably when applied to real-world satellite datasets. This underscores the importance of domain-aligned and diverse training data in achieving generalizable performance. In downstream tasks, such as image classification and object detection, training data characteristics significantly influence performance. Classification tasks marginally benefit from SR pre-processing when the training and test images are domain-aligned. For instance, SRGANs trained on ship imagery achieved slightly higher validation accuracy (98.72%) compared to raw imagery (98.59%). These results highlight that domain specific and semantic alignment with target tasks are crucial for exploring the benefits of SR in downstream applications.

6.5. Computational Cost

RSSR methods—especially those based on training transformers, GANs, and diffusion models from scratch—are computationally expensive, limiting real-time or onboard use. More recent and powerful models such as diffusion-based architectures are also expensive to do inference, often requiring several minutes on high-end GPUs, which restricts the accessibility. For example, models like RFA-PVTNet [71] or TransENet [69] achieve strong PSNR/SSIM but are slow (3.25–3.32 images/sec vs. SRCNN's 6.17 images/sec). A comparative analysis [107] shows that although progressive reconstruction predicts HR images more accurately, the multiple upsampling steps have considerably increased computational costs. These constraints are especially critical in time-sensitive RS applications, such as disaster responses, where inference time must be within seconds.

6.6. Limited Physical Integration

Most SISR models treat degradation as black-box (e.g., bicubic downsampling to generate LR images) [107]. However, actual LR images that are encountered in real-world scenarios have a totally different distribution compared to the ones generated synthetically using bicubic interpolation [107]. As a result, SR networks trained on artificially created degradations do not generalize well to actual LR images in practical scenarios. Furthermore,

current deep networks for SR are data-driven models that are learned in an end-to-end fashion. While this approach has shown excellent results in general, it proves to be sub-optimal when a particular class of degradation occurs for which a large amount of training data is non-existent [107]. In such cases, if the information about the sensor, imaged object/scene, and acquisition conditions is known, useful priors can be designed to obtain high-resolution images. Research also [108] shows that many of current models underperform on real-world SISR tasks because they are optimized more for image-to-image translation than for realistic SR reconstruction. This highlights a critical bottleneck in the generalization ability of existing SR models because of limited physical degradation modeling.

7. Opportunities and Directions

7.1. Data Augmentation and Enhancement

Data augmentation and enhancement techniques aim to increase the diversity and richness of training samples, which is particularly important in RSSISR tasks where HR images are scarce. Traditional augmentation strategies—such as rotation, flipping, cropping, and illumination adjustment—have been widely adopted to improve model robustness to geometric and radiometric variability [42]. In the context of RS, task-specific augmentations such as seasonal variation simulation, cloud masking can be applied to better reflect real-world acquisition conditions.

Beyond basic augmentation, data enhancement techniques that incorporate domain-specific auxiliary data have shown promise. For instance, spectral indices such as NDVI, land use/land cover labels, or even gradient edge maps [109, 38, 110] can be injected as additional channels to enrich spatial-spectral representations. These additional inputs can help guide the learning process, especially in weakly supervised or transfer settings. Overall, data augmentation and enhancement serve as foundational strategies to alleviate training data scarcity and improve model generalization in RSSISR pipelines.

7.2. Generative Priors and Large Foundation Models

Generative priors refer to implicit knowledge and statistics embedded within pretrained generative models [111], which can be generalized or fine-tuned to guide downstream or related tasks such as SISR. Recent advances in pretrained diffusion models and large foundation models—such as large

language models (LLMs) and vision language models (VLMs)—have demonstrated remarkable capabilities in image generation, translation, and structure-aware reasoning. These models offer new potential for improving remote sensing SISR through implicit priors, multimodal reasoning, and zero-shot adaptability [38, 40, 41]. For example, StableSR [40] incorporates a time-aware encoder into Stable Diffusion [38], preserving the Diffusion priors and generating realistic images with 200 sampling steps. DiffBIR [112] also freezes a pretrained diffusion model for its generative priors, and introduces a degradation-aware encoder to align the corrupted input with the diffusion model's latent space. Harnessing generative priors makes the method both flexible and efficient, and the model could generate images with 50 steps. However, these models are based on natural images, and their generalization to geospatial and RS datasets are unknown.

Furthermore, VLMs like CLIP [113], which encode rich semantic and spatial knowledge through cross-modal training, may help bridge the gap between pixel-level reconstruction and high-level scene understanding. Recent trends in VLM-based SR suggest promising directions for geospatial adaptation. For example, To enhance RSSR with high-level understanding, SuperCLIP [114] integrates semantic priors from a pretrained CLIP model. It extracts scene-level attributes (e.g., coast, desert, river) via vision-language embeddings and injects them into a Semantic Attribute-Guided Transformer, which fuses these priors with visual features. A visual semantic decoder reconstructs enriched representations, while a semantic projection network enforces alignment between the super-resolved output and CLIP-derived semantics using multi-level losses. This approach demonstrates that pretrained VLMs can guide pixel-level restoration through global semantic consistency.

Therefore, generative priors and foundation models offer a promising avenue to: 1) mitigate the dependency on large-scale, task-specific HR-LR datasets; 2) improve generalization under complex, real-world sensor degradations; 3) provide semantic-aware or task-specific super-resolution using multimodal input (e.g., text, maps, or class labels); and 4) enable zero-shot or few-shot adaptation to new domains.

7.3. Transfer-Learning, Self-Supervised, Unsupervised, and Training-Free Paradigms

Models discussed in this review generally learn from downsampled versions of HR images to simulate LR inputs. However, in many real-world remote sensing scenarios, such as historical archives, cloud-covered acquisitions, or low-budget satellite missions, HR references may not be available.

In such cases, conventional supervised training fails. To overcome this, several alternative learning paradigms have been proposed, including transfer learning, self-supervised learning, unsupervised learning, and training-free approaches.

Transfer learning. Transfer learning has emerged as a powerful solution to address the scarce training data and leverage pretrained model priors. Some approaches freeze encoder layers from pretrained SISR models while adapting decoder components to target domains. For instance, CNMF [115] and TransRes [116] both leverage pretrained architectures and fine-tune them on satellite images from specific regions or sensors. This strategy significantly reduces training time and improves generalization. However, transfer learning can suffer from domain shift when the source and target distributions differ substantially (e.g., urban RGB $\rightarrow$ rural multispectral). Domain adaptation or spectral-aware fine-tuning may be needed to mitigate this gap.

Self-supervised learning. SSL frameworks utilize internal information of the image itself to generate pseudo supervision signals. One notable method is D2U [93], which learns degradation-aware representations through contrastive learning and adapts them to guide SR reconstruction. Another is CASSISR [94], which introduces a cross-dimension attention mechanism and downscales the input further (e.g., 1/s$\times$LR) to generate a pseudo-SR task without HR ground truth. These methods enable learning from unlabelled data and are particularly suited for blind SR settings.

Unsupervised learning. Unsupervised super-resolution typically learns from unpaired LR and HR image sets. A representative work is Unsupervised GAN [91], which replaces the standard HR supervision with an adversarial loss between generated SR images and interpolated LR references. Another example is CLN4SR [92], which constructs a closed-loop between a downsampling network and a super-resolution network, enforcing cycle consistency to improve performance on infrared satellite imagery. These methods remove the need for HR annotation while ensuring fidelity and structure preservation through generative or cyclic losses.

Training-free methods. Training-free approaches, such as Deep Hyperspectral Prior (Deep HS Prior) [95], do not rely on pretraining or external datasets. Instead, they directly optimize a randomly initialized CNN (or 3D-CNN for hyperspectral images) on the input image, exploiting the implicit image prior

encoded in the network's structure. These methods are flexible, lightweight, and suitable for fast deployment on unseen data without domain adaptation. While they often suffer from lower performance compared to trained models, their independence from supervision makes them highly attractive for onboard or resource-constrained scenarios.

These alternative paradigms hold strong potential for remote sensing applications due to their low data requirements and ability to generalize across tasks. Specifically: 1) Transfer learning allows models trained on generic datasets to be adapted to RS domains; 2) SSL leverages unlabeled data efficiently and is robust to real-world degradations; 3) Unsupervised learning bridges the gap between SR and generative modeling without paired data; and 4) Training-free models eliminate the need for offline training or pretraining, making them ideal for dynamic or operational environments. Combining these paradigms with generative priors or task-specific adaptation strategies represents a promising future direction.

7.4. Application-Aware Methods

Despite the rapid progress in DL-based RSSISR, limited attention has been paid to evaluating the practical utility of super-resolved products in real-world downstream tasks [105]. Traditional evaluation relies heavily on pixel-level metrics such as PSNR and SSIM, which, while useful for measuring perceptual quality, may not accurately reflect performance in applied contexts such as agriculture, hydrology, or environmental monitoring. To address this gap, recent studies have introduced task-based evaluation frameworks for RSSR, where the output of SR models is directly assessed based on performance in application-specific tasks [105]. In this study, a benchmarking framework was applied to 76 images from PRISMA and TROPOMI satellites (2020–2022), spanning three scenarios: precision agriculture, inland and coastal water quality monitoring, and air pollution assessment.

For precision farming, the HyperTransformer achieves the highest accuracy at both 15 m and 5 m resolutions, indicating its effectiveness in retaining crop-relevant features. In contrast, PCA-based pansharpening demonstrates superior performance for inland and coastal water quality tasks, outperforming HyperTransformer and other HSI-SR methods, particularly in preserving spectral information critical to water reflectance analysis. Simpler methods such as bicubic interpolation, Bi-3DQRNN, and HAT consistently underperform across tasks, revealing their limited adaptability to complex environmental signals.

These results suggest that SR model effectiveness is strongly tied to task requirements, where spatial fidelity alone is insufficient—spectral integrity and structural consistency are equally critical. In domains like agriculture or aquatic monitoring, errors in spectral reconstruction can lead to misleading predictions, such as incorrect crop classification or water quality estimation. Thus, application-aware models must be optimized not only for perceptual realism but also for downstream robustness.

Future efforts in application-aware SR could benefit from integrating auxiliary modalities (e.g., RGB, NIR, or SAR), which provide complementary information under challenging conditions, such as heterogeneous land surfaces or cloud occlusion [117, 118]. Moreover, dynamic task supervision—where SR networks are trained jointly with downstream objectives like classification, segmentation, or retrieval—may offer an effective strategy to align SR reconstruction with the demands of real-world geospatial analysis.

7.5. Physically and Statistically Guided Methods

Physically and statistically guided super-resolution (SR) methods integrate domain-specific knowledge—such as sensor characteristics, radiative transfer models, and statistical priors—into the SR process. Unlike purely data-driven approaches, these methods aim to enhance reconstruction fidelity and generalizability, particularly in remote sensing applications where data scarcity, sensor noise, and complex degradations are prevalent.

Physically Guided Methods. These approaches incorporate physical models and sensor-specific information into the SR framework. For instance, [1] models realistic degradations of RS images as a convolution with a blur kernel, and also simulate image noises by different statistical distributions, such as Gaussian distribution and Poisson distribution. Another example is the PGRSID framework [119], which introduces a physics-guided SR approach by simulating realistic point spread functions (PSFs) using Zernike polynomials to model optical aberrations, sensor blur, and atmospheric effects. Experimental results on both simulated and real satellite data demonstrate that PGRSID could preserve fine structural details and handling real-world blur and noise conditions. [120] reconstructs high-resolution SWIR images by combining spectral inversion, radiative transfer modeling, and full-link sensor imaging simulation. The method estimates infrared reflectance from high-resolution visible-band data via spectral library matching, then simulates radiance using atmospheric and sensor models. A refinement stage

minimizes discrepancies between the observed LR infrared image and a degraded version of the reconstructed HR image using PSF, downsampling, and noise constraints. This physics-guided pipeline improves structural fidelity (SSIM up to 0.9157) and visual quality (VIF up to 0.6652), while preserving real infrared characteristics.

Statistically Guided Methods. These techniques introduce explicit regularization through Bayesian modeling, Markov Random Fields (MRFs), or uncertainty quantification to reflect prior distributions or spatial dependencies. For example, [121] propose a Bayesian hyperspectral SR model to better capture pixel-level uncertainty and avoid spectral artifacts. Meanwhile, MRF-based models have been used for edge and texture preserving image reconstruction in ill-posed SR settings [122]. These approaches are especially useful in hyperspectral and multimodal fusion tasks, where maintaining spatial coherence and spectral integrity is critical.

Despite their potential, physically and statistically guided SR methods remain underexplored. Future directions include incorporating radiative transfer models, refining uncertainty estimation in pixel reconstruction, and building modular physical priors that are compatible with end-to-end learning. These approaches promise to enhance the interpretability, reliability, and scientific utility of SR models across diverse applications.

7.6. Multi-task learning

Multi-task learning (MTL) enables models to simultaneously optimize for super-resolution and other downstream tasks, such as semantic segmentation, object detection, or change detection. For example, SEG-ESRGAN [123] super-resolves Sentinel-2 imagery to 2m resolution (a scaling factor of 5), with a parallel encoder–decoder branch for semantic segmentation, generating the enhanced land cover map. By sharing representations across tasks, MTL promotes mutual benefits—SR can enhance the quality of downstream inputs, while auxiliary tasks provide semantic guidance to regularize SR outputs. For instance, SEG-ESRGAN designs several skip connections from the SR branch and concatenates with features from the segmentation branch, promoting meaningful information to segmentation task [123]. Moreover, MTL can reduce training and inference costs compared to training multiple single-task networks, and improves generalization by enforcing cross-task consistency [124]. Future directions may focus on adaptive loss balancing, or task-specific decoders to further exploit multi-task synergy in RSSISR.

8. Conclusion

This review provides a comprehensive analysis of single-image super-resolution techniques in the context of remote sensing, covering different training paradigms (supervised, unsupervised, self-supervised, training-free learning), and various model architectures (CNN-based, GAN-based, Diffusion-based, and GNN-based methods).

By systematically categorizing the diverse approaches and evaluating their performance, this study reveals the strengths and limitations of each proposed method under various practical constraints. Despite substantial advancements, we identify key challenges in this field:

(1) Limited training data quality. High-quality LR-HR pairs or real-world degradation simulations are scarce in RS dataset, making it difficult to train robust models.

(2) Limited model generalization and transferability. Models trained on specific regions, resolutions, sensors often fail to generalize across different geographic locations, resolutions, and platforms due to different spatial variations and sensor characteristics.

(3) Lack of high scale factor super resolution methods. Most existing methods are optimized for scale factors up to $2\times$, $3\times$, $4\times$, with limited success on larger magnification factors such as $6\times$ or $8\times$, which are more and more critical and urgent in RS applications.

(4) Lack of application-oriented model designs. Most existing methods focus on improving perceptual quality rather than evaluating HR outputs to real-world downstream applications such as land cover classification or object detection.

(5) High computational cost. Although training deep learning models is particularly computationally demanding, many SOTA architectures also involve complex structures and large memory footprints during inference, limiting their deployment in real-time or resource-constrained environments.

(6) Limited physical integration. Current LR image generation methods (bicubic interpolation and Gaussian blur) rarely incorporate physics-informed and domain-specific knowledge, which can challenge the model's reliability in practical applications.

To address these, promising future research directions include:

(1) Data augmentation and enhancement. Generating diverse training samples through geometric transformations, spectral index incorporation (e.g., NDVI for SR in vegetated areas), and the data pairs enhancement of aux-

iliary data (e.g., land cover, topography for texture in croplands or shaded areas) may improve model robustness and generalization.

(2) Exploiting generative priors and large foundation models. Leveraging pre-trained generative models or foundation models (e.g., diffusion, vision-language models) can provide informative priors, reduce training costs, and improve performance in low-data regimes.

(3) Exploring self-supervised, unsupervised, and training-free learning methods. These approaches reduce dependence on expensive and human-labeled data and may provide scalability across different sensors and spatial regions, particularly valuable for under-annotated RS domains.

(4) Developing application-aware methods. Designing SR models with awareness of downstream tasks (e.g., agriculture, urbanization, forestry, human mobility) ensures that improvements in spatial resolution can translate to real-world application performance.

(5) Physically and statistically guided methods. Incorporating prior knowledge of sensor physics, radiometric distributions, or domain-specific constraints can enhance model interpretability and accuracy.

(6) Multi-task learning. Simultaneously training SR models with related tasks (e.g., classification, segmentation, object detection, denoising) encourages feature representations and improves overall model generalization and efficiency.

Moving forward, we advocate for a close coupling between algorithm design and real-world applications, providing practical solutions in Earth observation and geospatial analytics.

References

[1] J. Zhang, T. Xu, J. Li, S. Jiang, Y. Zhang, Single-image super resolution of remote sensing images with real-world degradation modeling, Remote Sensing 14 (12) (2022) 2895.

[2] C. E. Woodcock, A. H. Strahler, The factor of scale in remote sensing, Remote sensing of Environment 21 (3) (1987) 311–332.

[3] J. Graesser, N. Ramankutty, Detection of cropland field parcels from landsat imagery, Remote sensing of environment 201 (2017) 165–180.

[4] D. Poursanidis, N. Chrysoulakis, Z. Mitraka, Landsat 8 vs. landsat 5: A comparison based on urban and peri-urban land cover mapping, In-

ternational Journal of Applied Earth Observation and Geoinformation 35 (2015) 259–269.

[5] B. Watkins, A. Van Niekerk, Automating field boundary delineation with multi-temporal sentinel-2 imagery, Computers and electronics in agriculture 167 (2019) 105078.

[6] L. Liebel, M. Körner, Single-image super resolution for multispectral remote sensing data using convolutional neural networks, The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences 41 (2016) 883–890.

[7] H. Al-Mekhlafi, S. Liu, Single image super-resolution: a comprehensive review and recent insight, Frontiers of Computer Science 18 (1) (2024) 181702.

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

[9] J. Kim, J. K. Lee, K. M. Lee, Accurate image super-resolution using very deep convolutional networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 1646–1654.

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

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

[12] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, C. Change Loy, Esrgan: Enhanced super-resolution generative adversarial networks, in: Proceedings of the European conference on computer vision (ECCV) workshops, 2018, pp. 0–0.

[13] R. Fernandez-Beltran, P. Latorre-Carmona, F. Pla, Single-frame super-resolution in remote sensing: A practical overview, International journal of remote sensing 38 (1) (2017) 314–354.

[14] X. Wang, J. Yi, J. Guo, Y. Song, J. Lyu, J. Xu, W. Yan, J. Zhao, Q. Cai, H. Min, A review of image super-resolution approaches based on deep learning and applications in remote sensing, Remote Sensing 14 (21) (2022) 5423.

[15] P. Wang, B. Bayram, E. Sertel, A comprehensive review on deep learning based remote sensing image super-resolution methods, Earth-Science Reviews 232 (2022) 104110.

[16] K. Karwowska, D. Wierzbicki, Using super-resolution algorithms for small satellite imagery: A systematic review, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 15 (2022) 3292–3312.

[17] J. F. Burnham, Scopus database: a review, Biomedical digital libraries 3 (2006) 1–8.

[18] D. Han, Comparison of commonly used image interpolation methods, in: Conference of the 2nd International Conference on Computer Science and Electronics Engineering (ICCSEE 2013), Atlantis Press, 2013, pp. 1556–1559.

[19] G.-S. Xia, J. Hu, F. Hu, B. Shi, X. Bai, Y. Zhong, L. Zhang, X. Lu, Aid: A benchmark data set for performance evaluation of aerial scene classification, IEEE Transactions on Geoscience and Remote Sensing 55 (7) (2017) 3965–3981.

[20] G.-S. Xia, X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, M. Datcu, M. Pelillo, L. Zhang, Dota: A large-scale dataset for object detection in aerial images, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 3974–3983.

[21] Y. Yang, S. Newsam, Bag-of-visual-words and spatial extensions for land-use classification, in: Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems, 2010, pp. 270–279.

[22] D. Dai, W. Yang, Satellite image classification via two-layer sparse coding with biased image representation, IEEE Transactions on Geoscience and Remote Sensing 8 (1) (2011) 173–176.

[23] G. Cheng, J. Han, X. Lu, Remote sensing image scene classification: Benchmark and state of the art, Proceedings of the IEEE 105 (10) (2017) 1865–1883. doi:10.1109/jproc.2017.2675998.
URL http://dx.doi.org/10.1109/JPROC.2017.2675998

[24] Q. Zou, L. Ni, T. Zhang, Q. Wang, Deep learning based feature selection for remote sensing scene classification, IEEE Geoscience and Remote Sensing Letters 12 (11) (2015) 2321–2325. doi:10.1109/LGRS.2015.2475299.

[25] M. Ahmad, S. Protasov, A. M. Khan, R. Hussain, A. M. Khattak, W. A. Khan, Fuzziness-based active learning framework to enhance hyperspectral image classification performance for discriminative and generative classifiers, PloS one 13 (1) (2018) e0188996.

[26] Y. LeCun, Y. Bengio, et al., Convolutional networks for images, speech, and time series, The handbook of brain theory and neural networks 3361 (10) (1995) 1995.

[27] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai, et al., Recent advances in convolutional neural networks, Pattern recognition 77 (2018) 354–377.

[28] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[29] A. Shocher, N. Cohen, M. Irani, "zero-shot" super-resolution using deep internal learning, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 3118–3126.

[30] B. Hecht, M. Raubal, Geosr: Geographically explore semantic relations in world knowledge, The European Information Society: Taking Geoinformation Science One Step Further (2008) 95–113.

[31] A. Dosovitskiy, An image is worth 16x16 words: Transformers for image recognition at scale, arXiv preprint arXiv:2010.11929 (2020).

[32] A. Vaswani, Attention is all you need, Advances in Neural Information Processing Systems (2017).

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

[34] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, Advances in neural information processing systems 27 (2014).

[35] K. Vassilo, T. Taha, A. Mehmood, Infrared image super resolution with deep neural networks, in: 2021 11th Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing (WHISPERS), IEEE, 2021, pp. 1–5.

[36] J. Ho, A. Jain, P. Abbeel, Denoising diffusion probabilistic models, Advances in neural information processing systems 33 (2020) 6840–6851.

[37] J. Song, C. Meng, S. Ermon, Denoising diffusion implicit models, arXiv preprint arXiv:2010.02502 (2020).

[38] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-resolution image synthesis with latent diffusion models, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 10684–10695.

[39] C. He, Y. Shen, C. Fang, F. Xiao, L. Tang, Y. Zhang, W. Zuo, Z. Guo, X. Li, Diffusion models in low-level vision: A survey, IEEE Transactions on Pattern Analysis and Machine Intelligence (2025).

[40] J. Wang, Z. Yue, S. Zhou, K. C. Chan, C. C. Loy, Exploiting diffusion prior for real-world image super-resolution, International Journal of Computer Vision 132 (12) (2024) 5929–5949.

[41] R. Wu, L. Sun, Z. Ma, L. Zhang, One-step effective diffusion network for real-world image super-resolution, Advances in Neural Information Processing Systems 37 (2024) 92529–92553.

[42] L. Sun, R. Wu, Z. Ma, S. Liu, Q. Yi, L. Zhang, Pixel-level and semantic-level adjustable super-resolution: A dual-lora approach, arXiv preprint arXiv:2412.03017 (2024).

[43] S. Khanna, P. Liu, L. Zhou, C. Meng, R. Rombach, M. Burke, D. Lobell, S. Ermon, Diffusionsat: A generative foundation model for satellite imagery, arXiv preprint arXiv:2312.03606 (2023).

[44] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, S. Y. Philip, A comprehensive survey on graph neural networks, IEEE transactions on neural networks and learning systems 32 (1) (2020) 4–24.

[45] S. Zhou, J. Zhang, W. Zuo, C. C. Loy, Cross-scale internal graph neural network for image super-resolution, Advances in neural information processing systems 33 (2020) 3499–3509.

[46] L. Liebel, M. Körner, Single-image super resolution for multispectral remote sensing data using convolutional neural networks, The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences 41 (2016) 883–890.

[47] Y. Wei, Q. Yuan, H. Shen, L. Zhang, A universal remote sensing image quality improvement method with deep learning, in: 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), IEEE, 2016, pp. 6950–6953.

[48] N. Huang, Y. Yang, J. Liu, X. Gu, H. Cai, Single-image super-resolution for remote sensing data using deep residual-learning neural network, in: Neural Information Processing: 24th International Conference, ICONIP 2017, Guangzhou, China, November 14-18, 2017, Proceedings, Part II 24, Springer, 2017, pp. 622–630.

[49] C. Tuna, G. Unal, E. Sertel, Single-frame super resolution of remote-sensing images by convolutional neural networks, International journal of remote sensing 39 (8) (2018) 2463–2479.

[50] K. Zheng, L. Gao, B. Zhang, X. Cui, Multi-losses function based convolution neural network for single hyperspectral image super-resolution, in: 2018 Fifth International Workshop on Earth Observation and Remote Sensing Applications (EORSA), IEEE, 2018, pp. 1–4.

[51] W. Ma, Z. Pan, J. Guo, B. Lei, Achieving super-resolution remote sensing images via the wavelet transform combined with the recursive res-net, IEEE Transactions on Geoscience and Remote Sensing 57 (6) (2019) 3512–3527.

[52] Q. Qin, J. Dou, Z. Tu, Deep resnet based remote sensing image super-resolution reconstruction in discrete wavelet domain, Pattern Recognition and Image Analysis 30 (2020) 541–550.

[53] F. Deeba, F. A. Dharejo, Y. Zhou, A. Ghaffar, M. H. Memon, S. Kun, Single image super-resolution with application to remote-sensing image, in: 2020 Global Conference on Wireless and Optical Technologies (GCWOT), IEEE, 2020, pp. 1–6.

[54] F. Deeba, Y. Zhou, F. A. Dharejo, Y. Du, X. Wang, S. Kun, Multi-scale single image super-resolution with remote-sensing application using transferred wide residual network, Wireless Personal Communications 120 (1) (2021) 323–342.

[55] T. Geng, X.-Y. Liu, X. Wang, G. Sun, Deep shearlet residual learning network for single image super-resolution, IEEE Transactions on Image Processing 30 (2021) 4129–4142.

[56] H. Wang, Q. Hu, J. Chi, C. Wu, X. Yu, Multi-receptive-fields convolutional network for remote sensing images super-resolution, in: 2021 33rd Chinese Control and Decision Conference (CCDC), IEEE, 2021, pp. 1525–1530.

[57] R. Dong, L. Zhang, H. Fu, Blind super-resolution on remote sensing images with blur kernel prediction, in: 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS, IEEE, 2021, pp. 2879–2882.

[58] S. Deepak, D. Patra, S. M. Moorthi, Single image super-resolution reconstruction of remotely sensed images using mrf-2d phase congruency model, Journal of Applied Remote Sensing 15 (4) (2021) 046507–046507.

[59] J. Tang, J. Zhang, D. Chen, N. Al-Nabhan, C. Huang, Single-frame super-resolution for remote sensing images based on improved deep

recursive residual network, EURASIP Journal on Image and Video Processing 2021 (2021) 1–19.

[60] B. M. Nguyen, G. Tian, M.-T. Vo, A. Michel, T. Corpetti, C. Granero-Belinchon, Convolutional neural network modelling for modis land surface temperature super-resolution, in: 2022 30th European Signal Processing Conference (EUSIPCO), IEEE, 2022, pp. 1806–1810.

[61] M. Shi, Y. Gao, L. Chen, X. Liu, Dual-branch multiscale channel fusion unfolding network for optical remote sensing image super-resolution, IEEE Geoscience and Remote Sensing Letters 19 (2022) 1–5.

[62] H. Zhu, H. Tang, Y. Hu, H. Tao, C. Xie, Lightweight single image super-resolution with selective channel processing network, Sensors 22 (15) (2022) 5586.

[63] X. Wu, L. Zuo, F. Huang, Spatial and channel aggregation network for lightweight image super-resolution, Sensors 23 (19) (2023) 8213.

[64] A. Carbone, R. Restaino, G. Vivone, Efficient hyperspectral super-resolution of sentinel-5p data via dynamic multi-directional cascade fine-tuning, IEEE Geoscience and Remote Sensing Letters (2024).

[65] T. Barman, B. Deka, A deep learning-based joint image super-resolution and deblurring framework, IEEE Transactions on Artificial Intelligence (2023).

[66] İ. Babaoğlu, S. Kahveci, A. Kılıç, Enhanced pyramidal residual networks for single image super-resolution, Neural Computing and Applications (2024) 1–15.

[67] Y. Wang, H. Zhang, X. Zeng, B. Wang, W. Li, W. Ding, Binary lightweight neural networks for arbitrary scale super-resolution of remote sensing images, IEEE Transactions on Geoscience and Remote Sensing (2025).

[68] H. Mei, H. Zhang, Z. Jiang, Self-attention fusion module for single remote sensing image super-resolution, in: 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS, IEEE, 2021, pp. 2883–2886.

[69] S. Lei, Z. Shi, W. Mo, Transformer-based multistage enhancement for remote sensing image super-resolution, IEEE Transactions on Geoscience and Remote Sensing 60 (2021) 1–11.

[70] M. Shi, Y. Gao, L. Chen, X. Liu, Dual-resolution local attention unfolding network for optical remote sensing image super-resolution, IEEE Geoscience and Remote Sensing Letters 19 (2022) 1–5.

[71] Y. Cai, H. He, Z. He, M. A. Chapman, J. Li, L. Ma, J. Li, Enhancing spatial resolution of building datasets using transformer-based single-image super-resolution, in: IGARSS 2023-2023 IEEE International Geoscience and Remote Sensing Symposium, IEEE, 2023, pp. 6338–6341.

[72] A. Patnaik, M. K. Bhuyan, K. F. MacDorman, A two-branch multiscale residual attention network for single image super-resolution in remote sensing imagery, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (2024).

[73] W. Ma, Z. Pan, J. Guo, B. Lei, Super-resolution of remote sensing images based on transferred generative adversarial network, in: IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium, IEEE, 2018, pp. 1148–1151.

[74] K. Tran, A. Panahi, A. Adiga, W. Sakla, H. Krim, Nonlinear multiscale super-resolution using deep learning, in: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2019, pp. 3182–3186.

[75] L. Salgueiro Romero, J. Marcello, V. Vilaplana, Super-resolution of sentinel-2 imagery using generative adversarial networks, Remote Sensing 12 (15) (2020) 2424.

[76] J. Ma, H. Wu, J. Zhang, L. Zhang, Sd-fb-gan: Saliency-driven feedback gan for remote sensing image super-resolution reconstruction, in: 2020 IEEE International Conference on Image Processing (ICIP), IEEE, 2020, pp. 528–532.

[77] J. Ma, L. Zhang, J. Zhang, Sd-gan: Saliency-discriminated gan for remote sensing image superresolution, IEEE Geoscience and Remote Sensing Letters 17 (11) (2019) 1973–1977.

[78] F. A. Dharejo, F. Deeba, Y. Zhou, B. Das, M. A. Jatoi, M. Zawish, Y. Du, X. Wang, Twist-gan: Towards wavelet transform and transferred gan for spatio-temporal single image super resolution, ACM Transactions on Intelligent Systems and Technology (TIST) 12 (6) (2021) 1–20.

[79] H. Wu, L. Zhang, J. Ma, Remote sensing image super-resolution via saliency-guided feedback gans, IEEE Transactions on Geoscience and Remote Sensing 60 (2020) 1–16.

[80] Y. Li, S. Mavromatis, F. Zhang, Z. Du, J. Sequeira, Z. Wang, X. Zhao, R. Liu, Single-image super-resolution for remote sensing images using a deep generative adversarial network with local and global attention mechanisms, IEEE Transactions on Geoscience and Remote Sensing 60 (2021) 1–24.

[81] R. Islam, M. Khatun, S. H. Popy, Tl-gan: Transfer learning with generative adversarial network model for satellite image resolution enhancement, in: 2023 26th International Conference on Computer and Information Technology (ICCIT), IEEE, 2023, pp. 1–5.

[82] Z. Gao, L. Shen, Z. Song, H. Yan, Joa-gan: An improved single-image super-resolution network for remote sensing based on gan, IET Image Processing 18 (12) (2024) 3530–3544.

[83] J. Liu, Z. Yuan, Z. Pan, Y. Fu, L. Liu, B. Lu, Diffusion model with detail complement for super-resolution of remote sensing, Remote Sensing 14 (19) (2022) 4834.

[84] A. M. Ali, B. Benjdira, A. Koubaa, W. Boulila, W. El-Shafai, Tesr: two-stage approach for enhancement and super-resolution of remote sensing images, Remote Sensing 15 (9) (2023) 2346.

[85] M. Xu, J. Ma, Y. Zhu, Dual-diffusion: Dual conditional denoising diffusion probabilistic models for blind super-resolution reconstruction in rsis, IEEE Geoscience and Remote Sensing Letters (2023).

[86] Z. Liu, R. Feng, L. Wang, W. Han, T. Zeng, Dual learning-based graph neural network for remote sensing image super-resolution, IEEE Transactions on Geoscience and Remote Sensing 60 (2022) 1–14.

[87] M. Crawshaw, Multi-task learning with deep neural networks: A survey, arXiv preprint arXiv:2009.09796 (2020).

[88] T. Karras, T. Aila, S. Laine, J. Lehtinen, Progressive growing of gans for improved quality, stability, and variation, arXiv preprint arXiv:1710.10196 (2017).

[89] M. Qin, L. Hu, Z. Du, Y. Gao, L. Qin, F. Zhang, R. Liu, Achieving higher resolution lake area from remote sensing images through an unsupervised deep learning super-resolution method, Remote Sensing 12 (12) (2020) 1937.

[90] D. Mishra, I. Dror, O. Hadar, D. Choukroun, S. Maman, D. G. Blumberg, A fusion-based framework for unsupervised single image super-resolution, in: International Symposium on Cyber Security, Cryptology, and Machine Learning, Springer, 2023, pp. 85–95.

[91] N. Zhang, Y. Wang, X. Zhang, D. Xu, X. Wang, An unsupervised remote sensing single-image super-resolution method based on generative adversarial network, IEEE Access 8 (2020) 29027–29039.

[92] H. Zhang, C. Zhang, F. Xie, Z. Jiang, A closed-loop network for single infrared remote sensing image super-resolution in real world, Remote Sensing 15 (4) (2023) 882.

[93] Y. Xiao, Q. Yuan, K. Jiang, J. He, Y. Wang, L. Zhang, From degrade to upgrade: Learning a self-supervised degradation guided adaptive network for blind remote sensing image super-resolution, Information Fusion 96 (2023) 297–311.

[94] W. Jiang, L. Zhao, Y. Wang, W. Liu, B. Liu, Cross-dimension attention guided self-supervised remote sensing single-image super-resolution, Remote Sensing 13 (19) (2021) 3835.

[95] O. Sidorov, J. Yngve Hardeberg, Deep hyperspectral prior: Single-image denoising, inpainting, super-resolution, in: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, 2019, pp. 0–0.

[96] D. R. I. M. Setiadi, Psnr vs ssim: imperceptibility quality assessment for image steganography, Multimedia Tools and Applications 80 (6) (2021) 8423–8444.

[97] S. Abadal, L. Salgueiro, J. Marcello, V. Vilaplana, A dual network for super-resolution and semantic segmentation of sentinel-2 imagery, Remote Sensing 13 (22) (2021) 4547.

[98] H.-Y. Wang, J.-C. Juang, Retrieval of ocean wind speed using super-resolution delay-doppler maps, Remote Sensing 12 (6) (2020) 916.

[99] N. Saxena, Efficient downscaling of satellite oceanographic data with convolutional neural networks, SIGSPATIAL Special 12 (3) (2021) 46–47.

[100] J. Michel, J. Vinasco-Salinas, J. Inglada, O. Hagolle, Sen2venμs, a dataset for the training of sentinel-2 super-resolution algorithms, Data 7 (7) (2022) 96.

[101] A. Malczewska, J. Malczewski, B. Hejmanowska, Challenges in preparing datasets for super-resolution on the example of sentinel-2 and planet scope images, The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences 48 (2023) 91–98.

[102] Y. Xiong, S. Guo, J. Chen, X. Deng, L. Sun, X. Zheng, W. Xu, Improved srgan for remote sensing image super-resolution across locations and sensors, Remote Sensing 12 (8) (2020) 1263.

[103] J. Michel, E. Kalinicheva, J. Inglada, Revisiting remote sensing cross-sensor single image super-resolution: the overlooked impact of geometric and radiometric distortion, HAL open science (2024).

[104] Y. Meng, W. Li, S. Lei, Z. Zou, Z. Shi, Large-factor super-resolution of remote sensing images with spectra-guided generative adversarial networks, IEEE Transactions on Geoscience and Remote Sensing 60 (2022) 1–11.

[105] M. Kawulok, P. Kowaleczko, M. Ziaja, J. Nalepa, D. Kostrzewa, D. Latini, D. De Santis, G. Salvucci, I. Petracca, V. La Pegna, et al., Hyperspectral image super-resolution: task-based evaluation, IEEE Journal

of Selected Topics in Applied Earth Observations and Remote Sensing (2024).

[106] M. Ciolino, D. Noever, J. Kalin, Training set effect on super resolution for automated target recognition, in: Automatic Target Recognition XXX, Vol. 11394, SPIE, 2020, pp. 105–117.

[107] S. Anwar, S. Khan, N. Barnes, A deep journey into super-resolution: A survey, ACM computing surveys (CSUR) 53 (3) (2020) 1–34.

[108] U. Shami, B. Khan, Z. Zafar, M. Fraz, Bridging the resolution gap in remote sensing: A comparative analysis of deep learning models for real-world single image super-resolution, in: 2024 4th International Conference on Digital Futures and Transformative Technologies (ICoDT2), IEEE, 2024, pp. 1–8.

[109] K. Simonyan, A. Vedaldi, A. Zisserman, Deep inside convolutional networks: Visualising image classification models and saliency maps, arXiv preprint arXiv:1312.6034 (2013).

[110] M. Ghaffar, A. McKinstry, T. Maul, T. Vu, Data augmentation approaches for satellite image super-resolution, ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences 4 (2019) 47–54.

[111] H. Zhang, Y. Zhang, H. Li, T. S. Huang, Generative bayesian image super resolution with natural image prior, IEEE Transactions on Image processing 21 (9) (2012) 4054–4067.

[112] X. Lin, J. He, Z. Chen, Z. Lyu, B. Dai, F. Yu, Y. Qiao, W. Ouyang, C. Dong, Diffbir: Toward blind image restoration with generative diffusion prior, in: European Conference on Computer Vision, Springer, 2024, pp. 430–448.

[113] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International conference on machine learning, PmLR, 2021, pp. 8748–8763.

[114] D. Rambabu, C. Gayathri, R. Datla, S. Babu, Superclip: Semantic attribute-guided transformer with super resolution and clip for zero-shot remote sensing scene classification, IEEE Geoscience and Remote Sensing Letters (2025).

[115] Y. Yuan, X. Zheng, X. Lu, Hyperspectral image superresolution by transfer learning, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 10 (5) (2017) 1963–1974.

[116] Y. Zhang, R. Zong, J. Han, D. Zhang, T. Rashid, D. Wang, Transres: a deep transfer learning approach to migratable image super-resolution in remote urban sensing, in: 2020 17th annual IEEE international conference on sensing, communication, and networking (SECON), IEEE, 2020, pp. 1–9.

[117] K. Li, D. Dai, L. Van Gool, Hyperspectral image super-resolution with rgb image super-resolution as an auxiliary task, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2022, pp. 3193–3202.

[118] T. Y. Han, D. H. Kim, S. H. Lee, B. C. Song, Infrared image super-resolution using auxiliary convolutional neural network and visible image under low-light conditions, Journal of Visual Communication and Image Representation 51 (2018) 191–200.

[119] F. Ji, J. Wang, S. Cui, J. Li, X. Tang, F. Xu, Physics-guided optical simulation and psf analysis for remote sensing images deblurring, IEEE Transactions on Geoscience and Remote Sensing (2024).

[120] W. Chen, S. Jiang, F. Wang, X. Zhi, J. Hu, Y. Zhang, W. Zhang, Infrared remote-sensing image super-resolution based on physical characteristic deduction, Results in Physics 64 (2024) 107897.

[121] N. Akhtar, F. Shafait, A. Mian, Bayesian sparse representation for hyperspectral image super resolution, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3631–3640.

[122] S. Deepak, D. Patra, S. M. Moorthi, Single image super-resolution reconstruction of remotely sensed images using mrf-2d phase congruency model, Journal of Applied Remote Sensing 15 (4) (2021) 046507–046507.

[123] L. Salgueiro, J. Marcello, V. Vilaplana, Seg-esrgan: A multi-task network for super-resolution and semantic segmentation of remote sensing images, Remote Sensing 14 (22) (2022) 5862.

[124] H. Wang, H. Zhao, B. Li, Bridging multi-task learning and meta-learning: Towards efficient training and effective adaptation, in: International conference on machine learning, PMLR, 2021, pp. 10991–11002.

Table 2: Prior work based on supervised learning.

Data Source	Data Size	Model	SR Scale	Results (%)	Innovations	Limitations
1. CNN						
Sentinel-2	4224 images (244x244)	msiSRCNN [46]	x2	PNSR: 60.6527 SSIM: 0.9979	Capability of scaling multichannel images	Poor generalization due to limited training set
UC Merced	500 images (256x256)	RS-SRCNN [47]	x2, x4	PNSR: 29.27 SSIM: 0.9229	Generalization to image denoising and haze removal	Lack of model improvement
Sentinel-2	39675 images	RS-SRL [48]	x2	PNSR: 65.0622 SSIM: 0.9996	Tailored to RS images	Not specific to RS image characterictics
SPOT, Pleiades	- (250x250)	RS-SR [49]	x2, x3, x4	PNSR: 38.68 SSIM: 0.97 SAM: 0.34 ERGAS: 0.49	Comparison of SRCNN and VDSR	Lack of model improvement
CAVE, Pavia Centre	32, 7456 images (512x512, 1096x1096)	MLFN [50]	x2	PNSR: 49.674, 36.059 SSIM: 0.9985, 0.9762 SAM: 1.4596, 3.4593	Consider spectral loss	Lack of model improvement for spectra
RESISC45	700 images (256x256)	WTCRR [51]	x2, x3, x4	PNSR: 31.80 SSIM: 0.9051 NRMSE: 0.0257 ERGAS: 2.1818	Exploration of different frequency bands	Loss of high-frequency information
RESISC45	291 images (250x250)	DWTSR [52]	x2, x3, x4	PNSR: 32.1900 SSIM: 0.9087	Exploration of different frequency bands	Limited training data size
DIV2K	800 images	WRSR [53, 54]	x2, x3, x4	PNSR: 27.75 SSIM: 0.8189	Reduced computational cost	Limited model innovation
RESISC45	4680 images (256x256)	DSRLN [55]	x2, x3, x4	PNSR: 31.05 SSIM: 0.9221	High freqneucy details using Shearlet transform	Implementation relies on multiple tools
RESISC45	700 images (256x256)	MRFSR [56]	z2, x3, z4	PNSR: 33.44 SSIM: 0.908	Increased reception fields	Limited data category (only airplane images)
Self-collect	17000 images (480x480)	BKPSR [57]	x4	PNSR: 27.8848 SSIM: 0.7620	Captured degradation-specific features	Dependency on accurate kernel estimation
UC Merced	2100 images (256x256)	MRF-SRR [58]	x4	PNSR: 27.4028 SSIM: 0.8669	High freqneucy details using phase congruency	Computational cost
RESISC45	4500 images (256x256)	IDRRN [59]	x2, x3, x4	PNSR: 25.81 SSIM: 0.8077 ERGAS: 6.3351	Model efficiency	Vanishing gradient risk
MODIS LST	2573 images (64x64)	Multi-residual U-Net [60]	x4	PNSR: 28.40 SSIM: 0.85 MSE: 0.39	Real-world RSSISR	Limited to cloud-free images
UC Merced, WHU-RS19	900 images (96x96)	DMUNet [61]	x2, x3, x4	PNSR: 30.81 SSIM: 0.860	Fuse texture and edge information	No real-world scenario experiments
UC Merced, RESISC45	500 images (256x256)	SCPN [62]	x2, x3, x4	PNSR: 29.32 SSIM: 0.7961	Model efficiency	Bias in channel selection
Flickr2K, DIV2K	3450 images (192x192)	SCAN [63]	x2, x3, x4	PNSR: 30.34 SSIM: 0.9068	Lightweight	Not specific to RS image characterictics
Seninel 5P	Multiple	S5Net-dyn [64]	-	PNSR: 30.944 SAM: 2.082 EGRAS: 3.473	Consideration of spectral correlation	Computational cost
PatternNet, AID2 LISS-IV, LISS-III	(96x96)	JSRDNet [65]	x2, x3, x4	PNSR: 28.95 SSIM: 0.791	Generalization to image deblurring	Single level feature in SR branch
UC Merced, RESISC45	33,600 images (256x256)	EDPSR [66]	x2, x4	PNSR: 31.95 SSIM: 0.7455	Pyramid feature fusion with skip connections	Model complexity
GeoEye-1, Google Earth	Multiple	BLiSR [67]	x1-x4(arbitrary)	PNSR: 30.05 SSIM: 0.8831	Lightweight	Potential quantization errors from binary operations
2. Attention						
UC Merced	2100 images (256x256)	SAF [68]	x4	PNSR: 28.31 SSIM: 0.7784	Generalization	Increased computational
AID, UC Merced	12100 images (256x256, 600x600)	TransENet [69]	x2, x3, x4	PNSR: 29.38 SSIM: 0.7909	Multilevel feature embedding	Model complexity
UC Merced, WHU-RS19	900 images (96x96)	DLANet [70]	x2, x3, x4	PNSR: 30.87 SSIM: 0.849	Fuse information using flexible weights	Only considered local dependency
WHU Building		PVT-RFANet [71]	x4	PNSR: 22.01 SSIM: 0.50	Pyramid feature fusion	Model efficiency
WHU-RS19	1005 images (600x600)	TBMRA [72]	x2, x3, x4	PNSR: 30.837 SSIM: 0.816	Model stability	Computational cost Specialized use case
3. GAN						
UC Merced	100 images (256x256)	TGAN [73]	x4	PNSR: 27.62 SSIM: 0.78	Shared knowledge by transfer learning	Limited data size
WorldView-3	10000 images (320x320)	MSSRGAN [74]	x2, x4, x8	PNSR:22.931 SSIM: 0.4564	Capability of small scale object recovery	Computationally expensive
WorldView-2/3, Sentinel-2	Multiple	RS-ESRGAN [75]	x5	PNSR: 28.099 SSIM: 0.622 ERGAS: 25.389 SAM: 0.0954 CC: 0.958	Enhanced feature representation	Computationally expensive
GeoEye-1	298 images (512x512)	SD-FB-GAN [76]	x3, x4	PNSR: 22.36 SSIM: 0.6306	Inject varying levels of texture by saliency maps	High computational complexity
GeoEye-1	137 images (512x512)	SD-GAN [77]	x3, x4	PNSR: 20.9714 SSIM: 0.6964	Inject varying levels of texture by saliency maps	High computational complexity
DIV2K, UC Merced	Multiple	TWIST-GAN [78]	x2, x3, x4	PNSR: 35.21 SSIM: 0.9617	Exploration of different frequency bands	Other WT techniques
GeoEye-1	137 images (512x512)	SG-FBGAN [79]	x2, x3, x4, x8	PNSR: 26.81 SSIM: 0.86	Inject varying levels of texture by saliency maps	High computational complexity
RSCNN7, RESISC45, AID, DOTA, and UC Merced	Multiple	SRAGAN [80]	x2, x4	PNSR: 27.205 SSIM: 0.7710 MSE: 205.36 ERGAS: 1.6028	Consider local and global dependency	Computationally expensive
AID, UC Merced, and RSIs-CB256	10,000 images (256x256)	RBAN-UNet [1]	x4	PNSR: 25.676 SSIM: 0.7336	Simulating real-world sensor degradation	Model complexity
UC Merced	200 images (256x256)	TL-GAN [81]	x4	PNSR: 36.28 SSIM: 0.8340	Feature enhancement using transfer learning	Model complexity
AID, RESISC45 UC Merced	300 images (256x256)	JOA-GAN [82]	x4, x8	PNSR: 30.28 SSIM: 0.8016	High frequency region enhancement	No degradation diversity
4. Diffusion						
Potsdam, Vaihingen	(64, 128, 256, 512)	DMDC [83]	x2, x4, x8	PNSR: 23.46 SSIM: 0.6696 BRISQUE: 19.6959	Avoid over-smoothing problem	Sampling time cost
UC Merced	200 images (256x256)	TESR [84]	x2, x3, x4	PNSR: 31.951 SSIM: 0.90456	Restoration of context details	Sampling time cost
GeoEye-1, Google Earth		Dual-Diffusion [85]	x2, x3, x4	PNSR: 24.26	Degradation simulation	Model complexity
5. GNN						
3K VEHICLE$_S R$, Massachusetts Roads	8370 images (512x512)	DLGNN [86]	x2, x3, x4	PNSR: 24.050 SSIM: 0.636	Combine cross-scale features	Model complexity

Table 3: Prior work based on unsupervised learning.

Data Source	Data Size	Model	SR Scale	Results (%)	Innovations	Limitations
1. CNN						
Landsat 8, MODIS	Multiple	UDGN [89]	x2, x4, x8	PNSR: 35.1123 SSIM: 0.9726 RMSE: 0.0176 SAM: 0.0543	Injected geographic details by NDWI	Dependence on egde information
UC-Merced	2100 images (256x256)	FUSISR [90]	x2, x3, x4	BRISQUE: 52.2345 NIQE: 14.2345 PIQUE: 64.6754	Model efficiency	Failure in fine details
2. GAN						
UC Merced, RESIS45, WHU-RS19	All 3 datasets	Unsupervised GAN [91]	x2, x4	PNSR: 24.20 SSIM: 0.7136 SAM: 1.356 ERGAS: 4.626	super-resolve SR without HR labels	Limited in high-frequent info reconstruction
PROBA-V NIR Landsat 8	566 images (384x384)	CLN4SR [92]	x2, x3, x4	PNSR: 39.1266 SSIM: 0.9635	Can be trained in multiple modes (supervised, weakly supervised and unsupervised learning)	Not SOTA in supervised learning mode

Table 4: Prior work based on self-supervised Learning.

Data Source	Data Size	Model	SR Scale	Results (%)	Innovations	Limitations
1. CNN						
AID, DOTA	Multiple	D2U [93]	x4	PNSR: 30.18 SSIM: 0.7906	Simulation of real-world scenarios	Model complexity
2. Attention						
RSSCN7, RSC11, WHU-RS19, UC-Merced, AID, NWPU45	Multiple	CASSISR [94]	x2, x4	PNSR: 36.69 SSIM: 0.9514	No prior training	Dependence on repetitive features

Table 5: Prior work based on training-free learning.

Data Source	Data Size	Model	SR Scale	Results (%)	Innovations	Limitations
1. CNN						
Pavia Center	-, (150x150)	Deep HS Prior [95]	x2	PNSR: 33.67 SSIM: 0.967 SAM: 4.211	Image prior within a CNN itself	High computation cost

Table 6: Mapping of Research Directions to Identified Challenges

Challenge / Direction	**Data Augmentation and Enhancement**	**Generative Priors and Foundation Models**	**Self-/Unsupervised and Training-Free**	**Application-Aware Methods**	**Physically / Statistically Guided**	**Multi-task / Learning**
Training Data Quality	✓	✓	✓			
Generalization and Transferability	✓	✓	✓	✓		✓
High Scale Factor Super Resolution		✓				
Lack of Application-Oriented Design				✓	✓	✓
Computational Cost		✓	✓	✓	✓	
Limited Physical Integration					✓	✓