

Article

LoRA-Based Deep Learning for High-Fidelity Satellite Image Super-Resolution in Big Data Remote Sensing

Noha Rashad Mahmoud ^{1,2,*} , Hussam Elbehieri ¹ , Basheer Abdel Fattah Youssef ² 
and Hanaa Bayomi Ali Mobarz ^{2,3} 

¹ Faculty of Information Systems and Computer Science, October 6 University, Giza, 12613, Egypt; noha.rashad.comp@o6u.edu.eg (N.R.M.), vice.env.csis@o6u.edu.eg (H.E)

² Faculty of Computers and Artificial Intelligence, Cairo University, Giza, 12613, Egypt; basheer@fci-cu.edu.eg (B.A.F.Y.); h.mobarz@fci-cu.edu.eg (H.B.A.M.)

³ Faculty of Computer and Information Technology, Future University, Cairo, 11835, Egypt; hanaa.mubarrez@fue.edu.eg (H.B.A.M.)

* Correspondence: noha.rashad.comp@o6u.edu.eg

Abstract

High-resolution satellite imagery is pivotal for accurate analysis in remote sensing applications, including land-use monitoring, urban planning, and environmental assessment. However, obtaining such data is often costly and limited. Consequently, super-resolution techniques, such as deep learning models and fine-tuning strategies like LoRA, offer a promising alternative to the critical research challenge, especially given the diversity and large scale of satellite datasets. While deep learning-based super-resolution models have been very promising recently, their effectiveness, efficiency, and scalability across heterogeneous satellite scenes are not well studied. This work studies the performance of representative deep learning Super-Resolution frameworks, including the Enhanced Super-Resolution Generative Adversarial Network (ESRGAN), Swin Transformer for Image Restoration (SwinIR), and latent diffusion models (LDM), under unified experimental conditions using the WorldStrat dataset. The main goal is to establish whether adaptation strategies for parameter efficiency can boost reconstruction quality while reducing computational and training costs. Toward this goal, we investigate hybrid sequential pipelines, ensemble averaging, and Low-Rank Adaptation (LoRA)-based fine-tuning. The experiments indicate that these pipelines, which use multi-model methods, achieve only marginal performance gains while incurring substantial increases in computational complexity. LoRA-Based Fine-Tuning, by contrast, has demonstrated superiority in enhancing reconstruction accuracy and quality across all model families, despite using only a small percentage of trainable parameters. LoRA-based models demonstrate superiority over multi-model methods in both efficiency and performance. The presented results confirm that LoRA is an effective and accessible technique for high-fidelity satellite-based super-resolution image synthesis. The manuscript identifies LoRA as one of the enabling technologies advancing the state of the art in Deep Learning-based Super Resolution for large-scale satellite-based image synthesis.

Keywords: super-resolution remote sensing (SRRS); Low-Rank Adaptation (LoRA); latent diffusion models (LDM); SwinIR; ESRGAN; satellite image super-resolution; parameter-efficient fine-tuning; WorldStrat dataset; remote sensing image enhancement; computational efficiency



Academic Editor: Firstname Lastname

Received: 29 March 2026

Revised: 1 May 2026

Accepted: 12 May 2026

Published:

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article

distributed under the terms and

conditions of the [Creative Commons](#)

[Attribution \(CC BY\)](#) license.

1. Introduction

Satellite images play an integral role from an application perspective, especially in satellite-based environmental, agricultural, urban, and security applications. It has also been noted that the efficiency and effectiveness of satellite-based applications depend on the quality and resolution of the images acquired by these satellites. In most cases, satellite images exhibit lower quality due to acquisition challenges [1,2].

Some of these problems include the low-quality results delivered by early methods, such as their difficulty in handling high-frequency textures and complex spatial patterns typically found in satellite scenes [3,4]. However, recent deep learning-based single-image super-resolution (SISR) methods have shown promising results in addressing these problems. These include a convolutional neural network, a generative adversarial network, a transformer, and a diffusion model. However, each of these categories of methods poses distinct problems, including hallucinations, computational complexity, and inference costs [5,6].

Despite the remarkable development of deep learning-based precision optimization models, such as adversarial generative networks, diffusion models, and Transformers, their use in satellite imagery still faces fundamental limitations. These limitations include high computational complexity, slow execution, unstable training, and a trade-off between visual appearance and quantitative accuracy, with the potential to produce spatially inaccurate details. Also, such models are often designed for natural images and do not account for the multispectral characteristics of remote sensing data, and they exhibit poor generalization due to their reliance on industrial training data that do not accurately represent real imaging conditions [7]. These challenges highlight the need for a unified comparison of these models and are also driving the exploration of low-transactional fine-tuning technologies, such as LoRA, as an effective way to improve efficiency and reduce computational cost without significantly impacting performance.

Even though these families have enjoyed success, existing research on these families has treated them both independently and under different conditions, preventing objective comparison for satellite applications.

1.1. Problem Statement

Even with advancements in satellite image reconstruction by Deep Learning architectures in SR models, two important issues exist: first, “no comprehensive and unified comparative evaluation on representative SR architectures in terms of processing accuracy, perceptual quality, and computation efficiency on different satellite images has been conducted so far”.

The second reason is that there has been no systematic exploration of a family of efficient fine-tuning techniques, namely, low-rank adaptation (LoRA), for satellite image super-resolution tasks. The untapped opportunities that low-rank adaptation can offer for performance, training costs, and eventual deployment costs remain unclear.

1.2. Research Gap and Objectives

Although recent advances in satellite image super-resolution have been achieved through deep learning methods, significant research gaps persist in the current literature. Most studies to date concentrate on a single category of super-resolution models, such as generative adversarial networks (GANs) or transformers, and diffusion-based models, under different conditions across datasets, settings, and evaluation protocols.

The absence of a unified experimental setting also makes it hard to draw fair, generalizable conclusions about the relative merits, drawbacks, and computational efficiencies of different SR models applied to satellite images. Secondly, despite the suggestion of using

a hybrid sequential pipeline and ensemble-based approaches to improve reconstruction accuracy, there has been no analysis of their actual merits, despite the computational costs they entail. Finally, a parameter-efficient approach to fine-tuning deep networks, known as Low-Rank Adaptation (LoRA), has not been sufficiently explored for satellite images and super-resolution tasks, despite its potential to enhance performance at lower cost. The work in this study, therefore, compares three representative Super-resolution model families, including ESRGAN, SwinIR, and latent diffusion models, which are evaluated under unified experimental conditions.

Additionally, this study investigates the effectiveness of hybrid sequential pipelines, in which multiple models are applied sequentially to improve results, and of ensemble averaging, which combines the outputs of multiple models. To enhance accuracy, it could further optimize super-resolution performance for remote sensing data. Clarifying these strategies helps the reader understand their potential benefits and limitations in the context of satellite image super-resolution.

1.3. The Most Notable Contributions of This Work Are the Following

- Unified evaluation framework across different architectures: A systematic and fair comparison is conducted between three of the representative accuracy improvement models, namely generative network-based models (ESRGAN), Transformers (SwinIR), and diffusion models (LDM), under the same training and evaluation conditions and using standard metrics such as PSNR, SSIM, and FID.
- Analysis of integration methods and hybrid models: The performance of serial processing lines and integration methods (Ensemble) was studied, highlighting the associated challenges, particularly the increasing computational cost and inconsistent performance improvements.
- Application of LoRA Technology for effective fine-tuning: LoRA Technology has been applied across different models, demonstrating its effectiveness in improving reconstruction quality while reducing the number of learnable coefficients, making it an effective solution in terms of computational efficiency.
- Analysis of the balance between performance and efficiency: A comprehensive analysis of the relationship between reconstruction accuracy and perceptual quality, and between computational cost and inference time, was presented, indicating a favorable balance compared to LoRA-based models complex multi-model systems.
- Practical guidelines for real-world applications: The proposed framework provides practical insights for the design of high-efficiency satellite image resolution optimization systems, especially in resource-limited environments.

Structure of the paper: The rest of the paper is organized as follows. Section 2 contains a survey of the existing literature on the topic. Section 3 provides a description of the suggested methodology. Section 4 discusses the experiment design and implementation. In Section 5, the obtained results are analyzed and discussed. Finally, Section 6 summarizes the key findings and highlights avenues for future research.

2. Related Work

Recent research in satellite image enhancement has explored a variety of deep learning models, as shown in Figure 1.

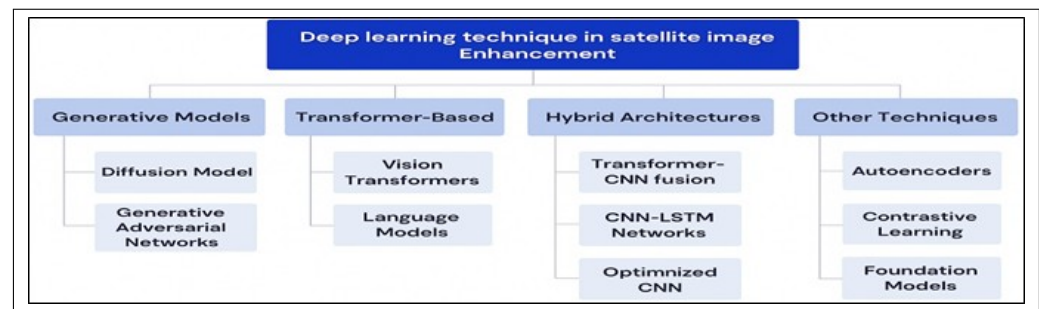


Figure 1. Classification of Deep Learning Techniques in Satellite Image Enhancement.

2.1. GENERATIVE MODELS

2.1.1. ESRGANs (Enhanced Super-Resolution Generative Adversarial Networks)

The ESRGAN network and its variants remain important approaches for improving the resolution of satellite images. These models rely on competitive training to generate high-definition images, thereby producing more accurate visual and textural details than traditional interpolation methods.

M. Greza et al. (2024) proposed a flexible model for improving resolution using competitive generative adversarial networks (GANs), called VSISR, specifically oriented to satellite images with an emphasis on maintaining high-frequency fine detail and radio-metric symmetry. The model is based on a training method that uses pixel blending to reduce unrealistic details while preserving the unique radiative characteristics of different data sources. Relying on Sentinel-2 as a reference, the model showed a good ability to optimize low-resolution images such as Landsat images. It was also trained on RGB data from multiple tasks, achieved average quantitative performance (PSNR is approximately 25.3 dB and SSIM is about 0.81), with an improved ability to generalize across various sensors. Although it was successful in reducing distortions and maintaining spectral accuracy, its performance was weaker in highly complex or convoluted regions. Other studies show that the ESRGAN model effectively handles complex textures in satellite images and changes in weather conditions, making it suitable for real-world applications in space photography. This model is characterized by its ability to reproduce fine details in urban and natural environments with high accuracy, as it relies on attention mechanisms to capture long-range spatial relationships and fine-grained structural details in multispectral data. The experimental results showed high performance, with the model achieving a PSNR of 33.52 dB, an SSIM of 0.862, and an SRE of 36.7 dB when applied to the RGB bands of Sentinel-2. The proposed method also significantly outperformed several advanced models, including ResNet, Swin Transformer, and Vision Transformer (ViT), highlighting its efficiency at optimizing spatial details while maintaining structural accuracy on standard remote sensing datasets [8].

2.1.2. LDM (Latent Diffusion Models)

H. Xiao et al. (2024)[9] proposed a framework called SatDiffMoE to improve the resolution of satellite images using latent diffusion models. This methodology enables an unlimited number of low-resolution images of the same location taken at different times to be combined into a single High-Resolution image by leveraging complementary temporal information. This time-conscious model has achieved advanced performance on several datasets while reducing complexity and transaction count compared to previous methods, thereby confirming the effectiveness of latent diffusion models in learning the generative properties of large-scale space images.

In a related context, presented a method to improve the representation of the underlying space in VAE models used in LDM frameworks by incorporating the discrete wavelet

transform (DWT). The results showed that the ExpDWT-VAE model is clearly superior to the traditional one in terms of representation quality, with the latent variance reaching 8.95. Experiments on the TerraFly-Sat dataset also showed a significant improvement in reconstruction quality and visual perception, with PSNR reaching about 22.80 dB and SSIM about 0.74, compared to 19.22 dB and 0.4984 for the basic model. In addition, measures of distributional similarity improved significantly, with FID at around 41.30 and KID at around 0.0146, compared with 64.44 and 0.0299, respectively, indicating that using DWT yields a richer latent representation and improves image generation quality. In general, these methods based on latent diffusion models represent a remarkable advance in the field of image resolution improvement, especially in satellite image applications, as they help achieve an effective balance between high reconstruction quality and computational efficiency.

2.1.3. Transformer Models

SwinIR (Swin Transformer for Image Restoration). Castillo et al. (2021)[10] presented the SwinIR model as a powerful reference model in image recovery, based on the Swin Transformer architecture. This model has demonstrated advanced performance in several applications, most notably accuracy improvement. This is due to its reliance on the hierarchical structure of transformer modules and the use of sliding windows, which enable efficient extraction of local and global properties. Thanks to this design, SwinIR was able to outperform models based on bypass networks (CNN) by a margin of 0.14–0.45 DB in the PSNR scale, while reducing the number of transactions, which facilitated the training process by up to 67%.

While developing this approach, Chen et al. (2025)[11] introduced an upgraded version of SwinIR for remote sensing applications, integrating a hybrid attention mechanism with a spatial gating component to improve ultra-high-resolution image reconstruction by highlighting fine details. SwinIR has also been used as a basis for developing more advanced transformer models, such as DPAT, which aim to improve the quality of high-resolution aerial images by producing clearer, sharper textures. These models have demonstrated clear efficiency in processing satellite cloud images and confirmed the generalizability of the SwinIR architecture across various satellite imaging applications. In terms of efficiency, the DPAT model achieved competitive performance despite using fewer transactions, relying on only about 32% of EDSR and 77% of SwinIR, while maintaining strong performance in accurate improvement measures, reflecting an effective balance between the quality of results and computational cost. In short, SwinIR technology and its derivatives will be of great importance in improving the accuracy of satellite images, as they perform better, can be applied to a larger number of images, and restore image quality more effectively than traditional interpolation methods.

2.1.4. Hybrid Architecture

Singgale et al. (2025)[12] proposed a hybrid framework based on bypass convolutional neural networks (CNNs), combining several architectures so that each is customized to address specific spectral ranges in multispectral satellite images. This approach is based on dynamic adaptation to temporal and spatial changes in environmental conditions, thereby improving feature extraction and classification. The model has achieved an accuracy of more than 85% across a range of temporal data, by integrating spectral indicators such as NDVI and NDBI into the feature extraction phase, surpassing traditional methods with an accuracy ranging from 65% to 75%. The model also showed stable performance when processing more than 1000 image segments, achieving more than 82% accuracy during feature extraction. Quantitative analyses also showed an increase in urban areas by

28% against a decrease in vegetation cover by 19% during the period from 2013 to 2024, which confirms the effectiveness of the model in environmental monitoring applications and analysis of urban changes, with a clear superiority over monostructural models in dealing with complex data.

Asif et al. (2025)[13] presented a hybrid deep learning model that combines techniques for improving accuracy (Super-Resolution) and for detecting objects in remote sensing images, using an improved StyleGAN architecture. The model has demonstrated high-quality images, with high PSNR and SSIM values across several test groups, reflecting its ability to improve accuracy while preserving detail. It is noteworthy that the accuracy improvement phase positively affected detection performance, as the average accuracy (mAP) at the 0.5 IoU threshold increased by 12.1% to 15.0%, i.e., from 72.0% to 82.3% after the accuracy improvement was applied. These results highlight the importance of integrating resolution-improvement techniques with detection tasks to improve image quality and enhance the system's final performance.

On the other hand, Al-Khafaji et al. (2025)[14] proposed a hybrid deep learning system aimed at compressing high-quality and scalable satellite images by employing a set of technologies, such as constant Wavelet Transform (SWT), automatic stacked noise removal encoder (SDAE), Gray-level correlation matrix (GLCM), as well as the k-means algorithm. This system enables multi-resolution analysis, texture-based coding, and adaptive processing across different areas. The model has achieved outstanding performance on several data sets, with PSNR values often exceeding 48 dB (maximum 50.36 dB), SSIM exceeding 0.99, and MMS-SSIM up to 0.9999, demonstrating almost lossless optical quality. It is also distinguished by high-speed encoding and decoding operations (about 0.065 s), surpassing many traditional and modern models across various performance indicators such as PSNR, SSIM, MMS-SSIM, and BPP, thereby confirming its high efficiency in image compression while maintaining quality.

Hybrid architecture plays a pivotal role in addressing challenges in satellite imagery by integrating contextual information from multispectral data, optimizing the representation of characteristics, and reconstructing high-quality images. The following table presents a comparison of previous studies in terms of performance using different data sets. Although existing hybrid architectures have shown high representational capacity, they typically rely on full fine-tuning of large pretrained backbones, in which all model parameters are updated during training. It incurs high computational cost, GPU memory usage, and training time, especially because it requires storing gradients and optimizer states for millions of parameters. Thus, these methods suffer from limited scalability and are unsuitable for use on resource-limited devices or in peripheral computing environments. To address these limitations, the low-rank adjustment technique (LoRA) has been adopted, an efficient method in terms of the number of coefficients that preserves the pre-trained weights of the basic model while adding a limited number of trainable low-rank adjustment matrices.

The LoRA approach restricts the update process to a particular subspace in order to minimize the number of trainable parameters, memory, computation, and overfitting. By leaving the backbone untouched, it preserves the pretrained knowledge, while introducing a smaller and more task-specific adaptation layer. Table 1 presents a summary of some latest SR approaches along with their main characteristics.

Table 1. Overview of state-of-the-art deep learning approaches for satellite image super-resolution.

Techniques	Models	Results	Limitations
Generative Models [6]	GANs (VSISR, ESRGANs)	PSNR = 25.3 dB, SSIM = 0.81	Limited by dataset size and model complexity.
Generative Models [8]	ESRGANs, VSISR, Ultra-dense GAN	PSNR = 31.75, SSIM = 0.9361	Model complexity and challenges in adapting to diverse satellite sources.
Generative Models [8]	SatDiffMoE	WorldStrat dataset: FID = 88.12; FMoW dataset: FID = 115.6	Requires multiple time-series images; complex fusion adds computational load.
Generative Models [9]	ExpDWT-VAE	PSNR = 19.76, SSIM = 0.4963, FID = 62.07	Relies on VAE latent quality; limited dataset; dual-branch architecture increases computation.
Transformer-Based [10]	SwinIR	Improves PSNR by 0.14–0.45 dB over CNNs with 67% fewer parameters.	High computational resources; complex attention; depends on training data quality.
Transformer-Based [11]	Enhanced SwinIR	PSNR = 28.27, SSIM = 0.7283	Increased computational overhead; complex attention mechanisms.
Hybrid Models [12]	Multi-architecture CNN	Accuracy > 85%	High resource demand for multi-CNN training; limited adaptability under low-data conditions.
Hybrid Models [13]	CNN + Transformer + Attention	PSNR = 28.5, SSIM = 0.89	Balancing model complexity vs efficiency; high resource needs for large datasets.
Hybrid Models [14]	SDAE + SWT	PSNR = 50.36, SSIM = 0.9964	Requires careful tuning of wavelet parameters; computationally expensive design.

3. Methodology

This section describes the general experimental process used in this study to improve the resolution of satellite images using deep learning models and Low-Rank Adaptation (LoRA) fine-tuning. The methodology consists of two principal phases: baseline model assessment and LoRA-based fine-tuning to enhance the model.

3.1. Proposed Model

It is split into three phases: (A) baseline evaluation of separate ESRGAN, SwinIR, and LDM models, (B) Sequential and Ensemble Pipelines, and (C) LoRA Fine-Tuning. This framework is based on the current state-of-the-art in the development of super-resolution (SR) models, as illustrated in Figure 2, and uses a three-step process that starts with benchmarking the models, followed by the hybrid integration of the models using different methodological approaches.

Consequently, an ensemble approach is adopted to leverage the strengths of multiple models. Finally, a fine-tuning approach known as Low-Rank Adaptation (LoRA) is employed to improve performance with minimal additional resources. These steps are incremental in that they improve performance, increase flexibility, and maintain efficiency.

As shown in Figure 2, the overall framework followed in this study to investigate and compare different approaches to satellite image super-resolution. The experiment begins with preprocessed pairs of HR/LR images that are subject to several procedures such as normalization, resizing, augmentation, and cleaning. Three phases constitute the experimental process. In the first phase, the models are used individually, and their performance is evaluated based on the metrics of PSNR, SSIM, and FID scores. In the second

phase, it is explored whether there is any advantage in using these models together, either in a serial manner where one model builds upon the other, or in a parallel manner where the outputs from the models are merged using a weighting system. The third phase asks a different question: rather than combining models, can a lightweight fine-tuning strategy like LoRA improve each model individually without the overhead of full retraining? Once all experiments are complete, the results from all three phases are brought together in a final comparison stage. The method that achieves the highest PSNR and SSIM alongside the lowest FID is selected as the best-performing approach and taken forward as the final super-resolution output.

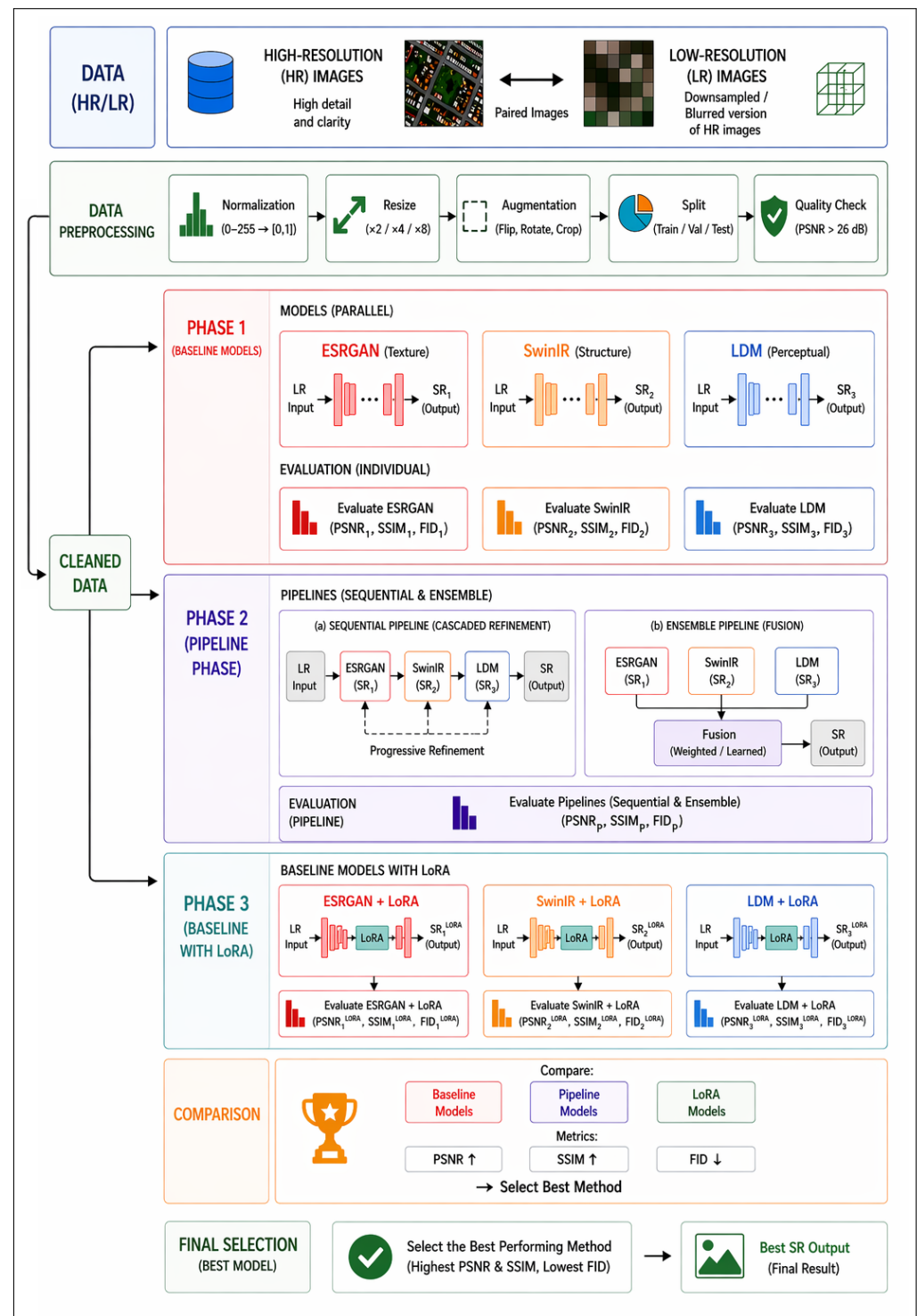


Figure 2. An overview of the proposed framework.

Following this, the proposed framework is structured as follows:

3.1.1. Baseline Models

Three advanced deep learning models were selected as a reference framework for enhancing the accuracy of satellite imagery, to facilitate a comprehensive comparison among different architectural models. The ESRGAN model is an example of competitive (Adversarial Learning), as it is characterized by its strong ability to improve the perceptual realism of images. The SwinIR model, based on transformer architecture, is efficient at capturing long-range relationships and preserving structural details, both of which are essential in remote sensing applications. Additionally, the latent diffusion model (LDM) is included as a diffusion-based approach, given its ability to gradually improve and its strong performance in high-resolution image reconstruction. The evaluation of these models enables a systematic analysis of reconstruction accuracy and visual perception quality by comparing different frameworks, including generative, transformer-based, and diffusion models.

ESRGAN

This model improves visual quality by restoring textures and high-frequency details, relying on residential density Blocks (RRDBs) integrated into the structure of competitive generative networks (GANs). Figure 3 shows the general structure of the ESRGAN model.

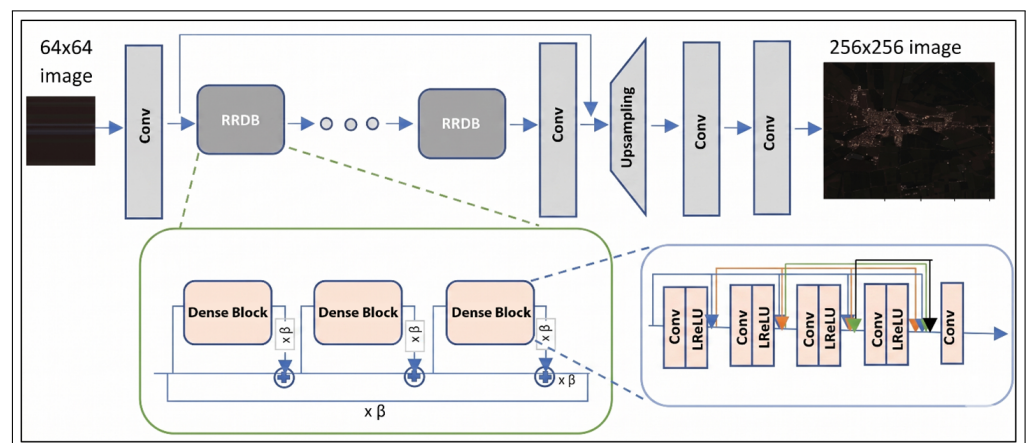


Figure 3. Architecture of Enhanced Super-Resolution Generative Adversarial Networks.

It has been developed to better suit remote sensing applications, with improved versions such as RS-ESRGAN that rely on high-resolution commercial satellite data recorded in conjunction, with the aim of raising the spatial resolution of Sentinel-2 images. This method has shown better performance on standard metrics such as PSNR and SSIM while preserving the necessary spectral information in satellite image applications.

SwinIR

This model is based on The Swin Transformer architecture, which uses sliding windows to implement the mechanism of hierarchical self-attention (self-attention), which allows capturing local and global characteristics of the image with high efficiency. This transformer-based design contributes to improved restoration of image geometric structures, making it especially suitable for precise reconstruction tasks in high-resolution satellite images, where maintaining geometric accuracy is paramount. Figure 4 shows the general structure of the SwinIR model.

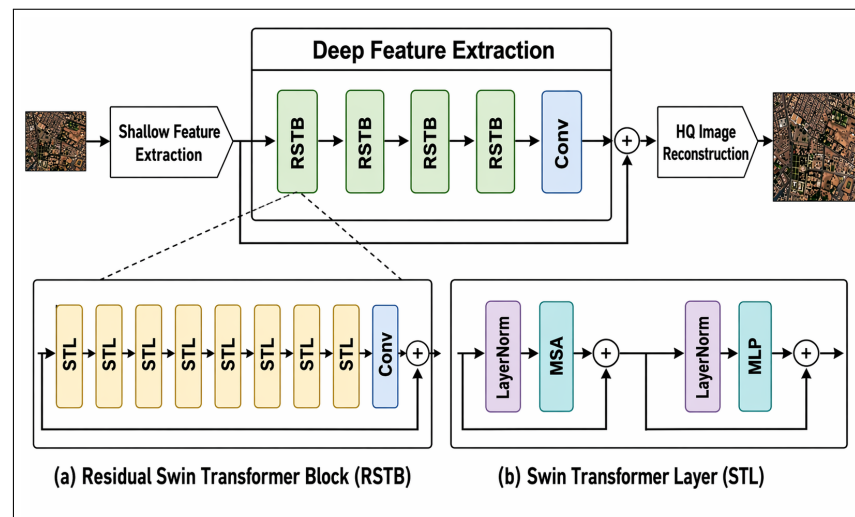


Figure 4. Architecture of SwinIR for image super resolution.

LDM

It uses diffusion modeling in a hidden space to improve images by learning the complex data is spread out. This produces super-resolution outputs of very high quality, with rich texture and perceptual fidelity. LDMs have recently been recognized for their ability to produce intricate, lifelike satellite image enhancements through progressive diffusion steps [15]. The overall architecture of the LDM model is illustrated in Figure 5.

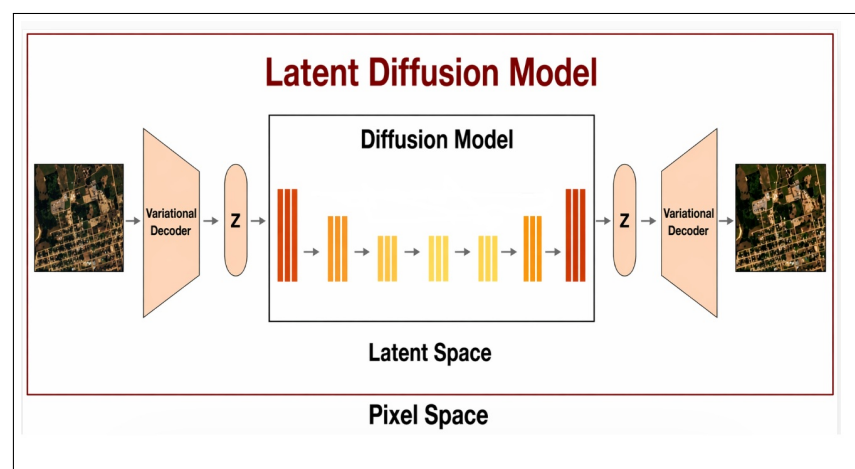


Figure 5. Architecture of LDM for image super resolution [15].

3.1.2. Sequential and Ensemble Approaches in Pipeline Strategy

Pipeline stage is aimed at improving individual models by employing two kinds of techniques: the sequential and the ensemble approaches. The former technique involves feeding the output of each model into the input of another. This way, the image can be progressively refined using this process. On the other hand, the latter technique uses the input together at once for all models, with the output being combined using techniques like weighted averaging.

3.1.3. LoRA (Low-Rank Adaptation)

It is a way to fine-tune large pre-trained models for new tasks without retraining the entire model. It uses a few parameters. LoRA doesn't change all the model weights, which can number in the billions in large models. Instead, it keeps the original weights frozen and adds small trainable matrices that use low-rank decomposition to make the needed

changes. LoRA breaks the weight update matrix down into two smaller matrices, A and B—the updated weight matrix.

Inspired by this, we hypothesize that the updates to the weights also have a low “intrinsic rank” during adaptation. For a pre-trained weight matrix $W_0 \in \mathbb{R}^{d \times k}$, we constrain its update by representing it with a low-rank decomposition:

$$W = W_0 + \Delta W = W_0 + BA, \quad (1)$$

where $B \in \mathbb{R}^{d \times r}$, $A \in \mathbb{R}^{r \times k}$, and the rank $r \ll \min(d, k)$. This formulation is defined in Equation (1) [16]. In this equation,

- $W_0 \rightarrow$ the pre-trained parameter weights
- $\Delta W \rightarrow$ the learned weights to be used in adjusting the original weights
- $W \rightarrow$ the final fine-tuned weight that will be used during inference
- $B \rightarrow$ a matrix of dimension $d \times r$
- $A \rightarrow$ a matrix of dimension $r \times k$

This significantly reduces the number of trainable parameters, usually by more than 90%, while keeping performance close to that of full fine-tuning. For instance, fine-tuning GPT-3 with LoRA cuts the number of trainable parameters from 175 billion to about 18 million. This leads to significant savings in memory, computational power and cost.

The efficiency of LoRA Technology also enables faster training, reduced resource consumption, and easier deployment of custom models. This technique is widely used to improve the performance of large linguistics and other basic models on specific tasks or areas without affecting overall performance [17]. Figure 6 shows the general structure of an improved model using LoRA.

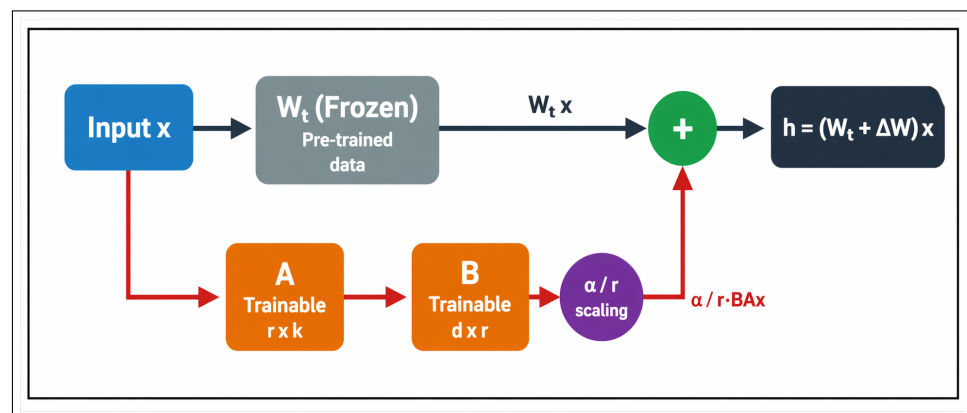


Figure 6. Architecture of LoRA-enhanced model.

3.1.4. LoRA Fine-Tuning for Baseline Models

To improve the performance of the basic models, low-rank adaptive technology (LoRA) has been integrated into ESRGAN, SwinIR, and LDM networks to efficiently adjust trans-actions. This technique is based on introducing small, trainable, low-rank matrices into specific layers while keeping the original pre-trained weights unchanged. This method significantly reduces the requirements for memory and computational resources.

The input points have also been allocated individually to each architecture, focusing on the most resource-intensive components, to ensure effective adaptation without adding excessive complexity to the model.

ESRGANS with LoRA

LoRA Technology was integrated into the bypass layers and residential density Blocks (RRDB), which helped the model improve the reconstruction of high-frequency details in

satellite images while maintaining the stability of the competitive training process. The overall architecture of the LoRA-ESRGAN model is illustrated in Figure 7.

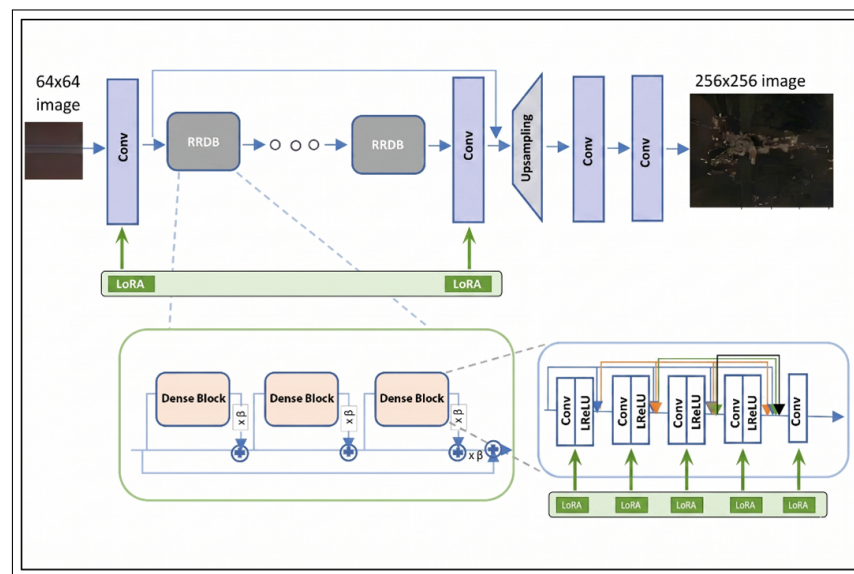


Figure 7. Integration of LoRA modules into ESRGAN architecture.

SwinIR with LoRA

LoRA was added to the linear projection layers of the self-attention and feed-forward modules in the Swin Transformer blocks of SwinIR. The overall architecture of the LoRA-SwinIR model is illustrated in Figure 8. This change improved the model's ability to capture long-range dependencies and fine-grained structural details, which are very important for remote sensing applications.

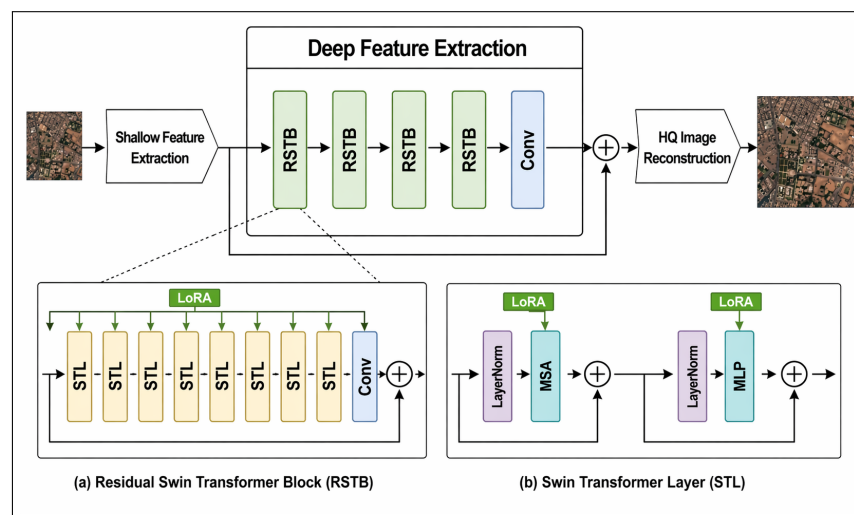


Figure 8. Integration of LoRA modules into the SwinIR architecture.

LDM with LoRA

LoRA was added to the cross-attention layers of the UNet backbone in LDM as part of the latent diffusion process. This change improved the model's ability to refine latent features during iterative diffusion steps, resulting in more realistic satellite image reconstructions. The overall architecture of the LoRA-LDM model is illustrated in Figure 9.

Although the fine-tuning framework based on LoRA Technology shows a noticeable improvement in reconstruction performance with high computational efficiency, there are still some limitations. Firstly, the effectiveness of LoRA depends on the choice of the

classes in which it is inserted as well as the rank Value (rank), parameters that are currently determined experimentally, and may not be optimal for all models or data sets. Secondly, although LoRA reduces the complexity of training, it does not radically address the high heuristic cost associated with some architectures, especially diffusion-based models.

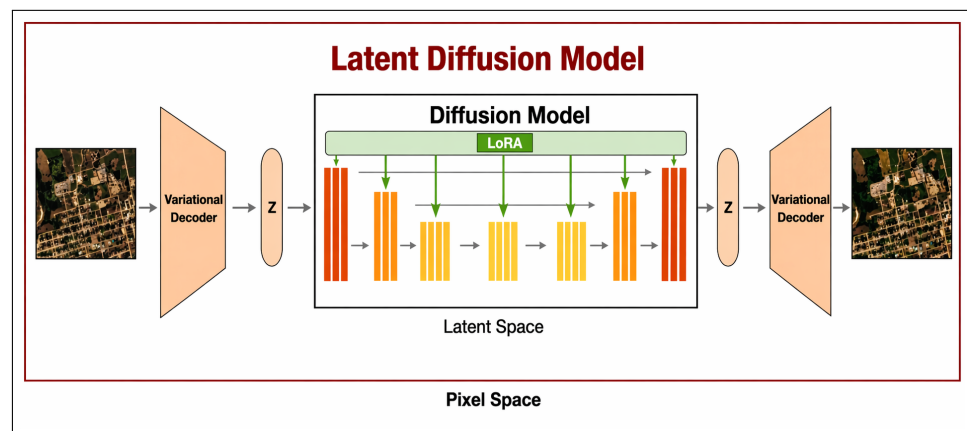


Figure 9. Integration of LoRA modules into Latent Diffusion Model (LDM) .

4. Experimental Work

We will show our experimental results and the datasets used in this section.

4.1. Dataset Overview

This data set covers approximately 10,000 square kilometers of Earth's surface, combining high-resolution satellite images (1.5 m/Pixel) from Airbus and low-resolution images (10 m/Pixel) from Sentinel-2. The locations were selected in a globally diverse and systematic manner to include different land-use types, such as agricultural areas, ice caps, forests, and urban areas of varying densities. It also includes underrepresented sites in machine learning datasets, such as humanitarian zones, illegal mining sites, and vulnerable settlements. Each high-resolution image is temporally matched with multiple Sentinel-2 low-resolution images. The licensing terms specify that high-resolution Airbus imagery is available under CC BY-NC 4.0 for non-commercial use; Sentinel-2 imagery, labels, and trained weights are available under CC BY 4.0; and the source code will be released under the BSD 3-Clause License [18]. Sample images representing different categories within the dataset are shown in Figure 10.

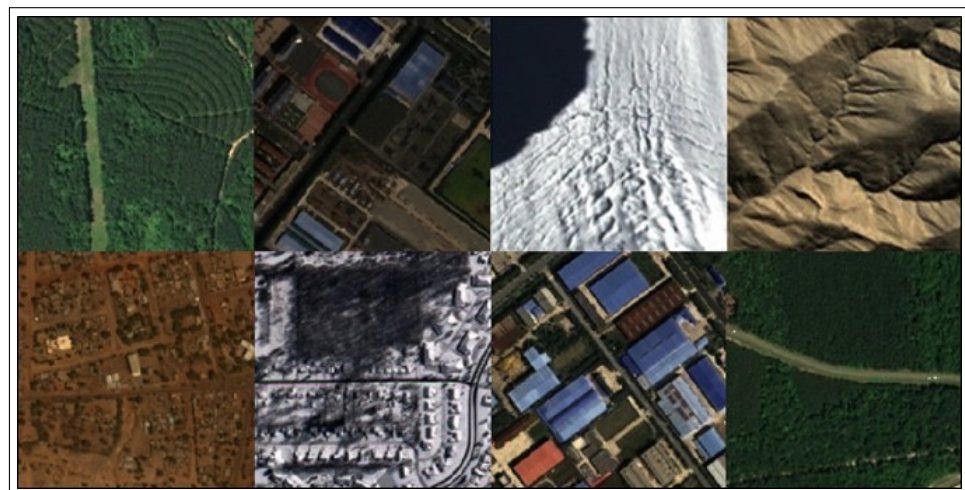


Figure 10. Different categories of the world stratified Dataset.

4.2. Preprocessing Steps

The first step is file pairing, which collects high-resolution (HR) images from hr-dataset, especially rgb.png files which are the most important, and low resolution (LR) images from lr-dataset-l1c/L1C with -l3-L1C-data.tiff files, matching them based on the base ID. This will result in a set of LR-HR pairs and 10,167 HR-only images. The samples were chosen by downscaling HR-only images to LR. Images are loaded with tiff file for TIFFs and cv2 for PNGs, truncating TIFFs with more than 3 channels to RGB and converting grayscale to RGB. Pixel values are normalized to the specified uint8 range and scaled accordingly if they fall outside it. LR images are resized to 128×128 pixels and HR images to 512×512 pixels using bicubic interpolation; for HR-only images, LR counterparts are generated by downscaling. The dataset is split into training (80%), validation (10%), and test (10%) sets. Baseline models were trained under computational constraints imposed by the available hardware (NVIDIA Tesla P100 GPU) and dataset scale. Specifically, ESRGAN was trained using 200 epochs, SwinIR—150 epochs, and LDM—10,000 gradient steps with an effective batch size of 4. Analysis of the validation PSNR curves confirms that these configurations did not reach full convergence on the WorldStrat satellite dataset, which explains the relatively low baseline scores of 20.96 dB, 24.82 dB, and 23.90 dB for ESRGAN, SwinIR, and LDM, respectively. Subsequent extended training experiments demonstrate that optimal convergence for the WorldStrat dataset requires 200–250 epochs for ESRGAN, 300–400 epochs for SwinIR, and 30,000–50,000 gradient steps for LDM. LoRA fine-tuning was then applied for an additional 50 epochs for ESRGAN, 70 epochs for SwinIR, and 6000 gradients steps for LDM. Early stops were applied on the basis of Peisner validation, with training stopped when the improvement dropped below 0.1 DB over the course of 20 consecutive epochs.

For LoRA-based fine-tuning, where the Adam optimizer was adopted with a fixed learning rate of 1×10^{-4} with a loss function of type L1. Due to limited memory, a batch size equal to 1 was used with the grouping of gradients in four steps to achieve an effective size equal to 4, as well as the use of mixed precision (FP16) to improve computational efficiency. The models were trained for 10,000 steps on 128×128 low-resolution images and 512×512 high-resolution images under a $4\times$ magnification setting.

4.3. LoRA Fine-Tuning Setup and Training

The LoRA technique is applied with settings including rank $r = 64$, scaling coefficient $\alpha = 64$, and a dropout rate of 0.1, where its units have been inserted within the attention layers (to_q, to_k, to_v, and to_out.0) to achieve effective adaptation while maintaining low computational cost.

To further validate the choice of LoRA rank ($r = 64$), a controlled empirical analysis has been conducted across three architectural settings: ESRGAN (GAN-based), SwinIR (transformer-based), and the Latent Diffusion Model (diffusion-based). All three architectures benefited from using LoRA rank, with increased PSNR (+13.68 dB for ESRGAN, +11.38 dB for SwinIR, and +9.95 dB for LDM), while increasing the number of trainable parameters by less than 3.1%.

At the same time, additional experiments with alternative ranks $r \in \{16, 32, 128\}$ reveal insights behind the chosen rank. A rank of 16 is insufficient, as it lacks sufficient expressive power to accurately reconstruct images, leading to the loss of fine spatial details such as building borders and smaller objects on the terrain. The rank of 32 allows for better image reconstruction, although high-frequency details and edges remain poorly represented. On the other hand, rank 64 offers optimal expressive power for representing small-scale objects such as city infrastructure, borders, and texture details. We find that, for image recovery, the benefit of increasing the number of ranks is limited to 128 or fewer.

Instead, they unnecessarily increase training time, consume more memory, and increase the risk of overfitting.

The loss curves are shown in Figure 11. Training loss consistently decreases, reflecting stable learning. Meanwhile, the validation loss is a bit oscillating, but there is no upward trend. The lack of divergence shows that the model is generalizing well. Finally, the intersection of these two curves indicates the stabilization of the model confirming our chosen LoRA parameters.

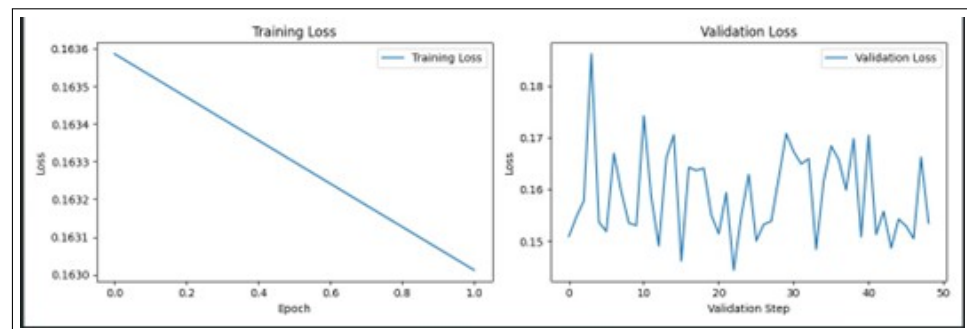


Figure 11. Training and Validation loss curves showing stable convergence and consistent model performance.

4.4. Evaluation Metrics

The performance of the model was evaluated using both qualitative and quantitative measures: Fréchet Inception Distance (FID): Measures the similarity between an original image and a reconstructed or enhanced image in terms of feature distributions, using an Inception network. Lower FID scores indicate closer alignment with original data. FID value is computed using the following Equation (2) [19]:

$$\text{FID} = \|\mu_r - \mu_g\|^2 + \text{Tr}(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{1/2}) \quad (2)$$

where:

- $\mu_r \rightarrow$ the empirical mean vector of the feature representations of the original images.
- $\mu_g \rightarrow$ the empirical mean vector of the feature representations of the reconstructed or enhanced images.
- $\Sigma_r \rightarrow$ the empirical covariance matrix of the feature representations of the original images.
- $\Sigma_g \rightarrow$ the empirical covariance matrix of the feature representations of the reconstructed or enhanced images.
- $\|\mu_r - \mu_g\|^2 \rightarrow$ the squared difference between the means of the two distributions (captures the shift in feature space).

A lower FID indicates a reconstructed or enhanced image is closer to the original. An FID of zero implies that the two images are identical. Larger values indicate greater divergence between original and reconstructed or enhanced images. 2-Peak Signal-to-Noise Ratio (PSNR) is one of the most commonly used quantitative metrics for image quality. It measures the similarity between an original image and its reconstructed or enhanced version. PSNR value is computed using the following Equation (3) [20]:

$$\text{PSNR} = 10 \cdot \log_{10} \left(\frac{\text{MAX}_I^2}{\text{MSE}} \right) \quad (3)$$

where: MAX_I represents the maximum possible pixel intensity value (for 8-bit images, $\text{MAX}_I = 255$). MSE denotes the Mean Squared Error between the original and the recon-

structed images. A higher PSNR value indicates that the reconstructed image is closer to the original, meaning better perceptual quality and fewer distortions. Structural Similarity Index Measure (SSIM): is a perceptual metric used to evaluate the similarity between an original image and a reconstructed or enhanced image. Unlike PSNR, which measures absolute pixel-wise differences, SSIM assesses image quality based on structural information and contrast. SSIM value is computed using the following Equation (4) [21]:

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \quad (4)$$

where:

- μ_x and μ_y are the mean intensities of the original image x and the reconstructed (or enhanced) image y [21].
- σ_x^2 and σ_y^2 are the variances of x and y .
- σ_{xy} is the covariance between x and y .
- C_1 and C_2 are small constants added to prevent division by zero.

Interpretation: SSIM values range from -1 to 1 , where:

- $\text{SSIM} = 1$ indicates that the two images are identical.
- $\text{SSIM} = 0$ indicates no structural similarity.
- $\text{SSIM} < 0$ suggests significant structural differences.

5. Results and Discussion

5.1. Baseline Models Results

This table illustrates that SwinIR achieved the highest PSNR and SSIM (0.5976), indicating better reconstruction quality. The LDM model also achieved the lowest FID score (183.72), indicating superior perceptual quality. In contrast, the ESRGAN model recorded the lowest PSNR and SSIM scores but the highest FID score, even though it was the fastest of the models. These results highlight a trade-off between quality and efficiency across the three models shown in Table 2.

Table 2. Summarizes the performance of ESRGAN, SwinIR, and LDM models.

Model	PSNR (dB) ↑	SSIM ↑	FID ↓
ESRGANs	20.96	0.3804	237.81
SwinIR	24.82	0.5976	201.10
LDM	23.90	0.5240	183.72

Quantitative results of the performance of ESRGAN, swinir and LDM models using PSNR, SSIM and FID scales show that the SwinIR model achieved the best values in both PSNR (24.82 DB) and SSIM (0.5976), which indicates higher accuracy of reconstruction and better ability to preserve structural characteristics. In contrast, the LDM model recorded the lowest FID score (183.72), indicating higher cognitive realism and greater similarity to the real image distribution. The ESRGAN model ranked last in structural consistency, achieving lower PSNR and SSIM values and the highest FID (237.81), indicating relatively weaker performance in preserving structure.

In general, these results indicate that transformer-based models are more focused on reconstruction accuracy, while diffusion-based models excel at achieving higher cognitive quality. This comparative assessment of the three models is presented in Figure 12.

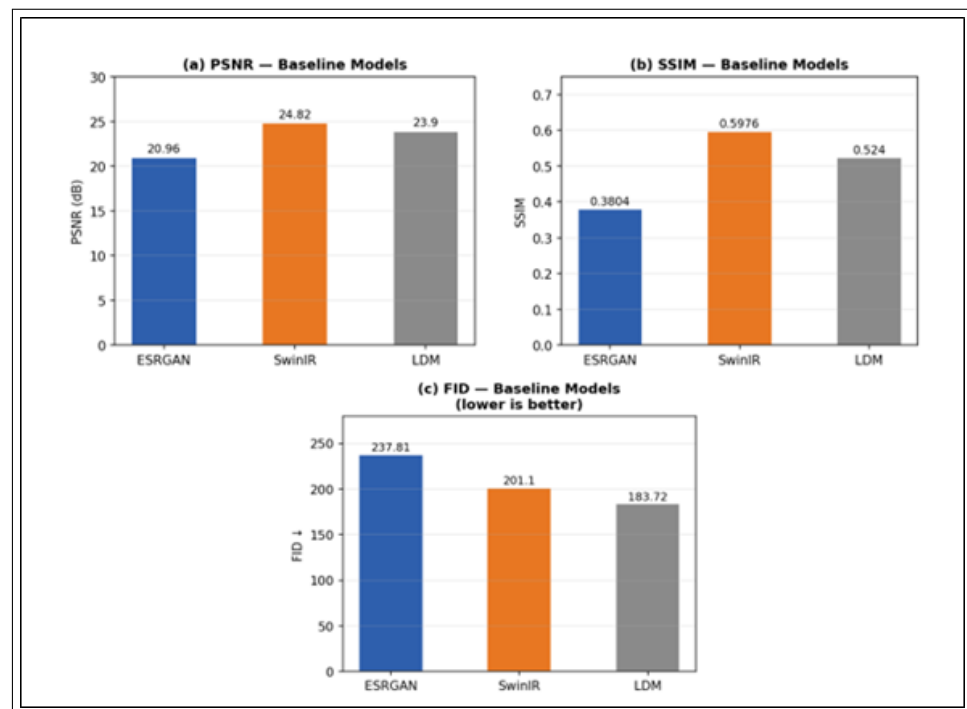


Figure 12. Quantitative comparison of models (a) PSNR, (b) SSIM, and (c) FID metrics. Each metric uses an independent y-axis for accurate representation.

The performance trends of the compared models in terms of PSNR, SSIM, and FID show that the SwinIR model achieves clear superiority in the reconstruction-based measures (PSNR and SSIM), while the LDM model achieves the best cognitive performance, with the lowest FID. This visual representation also supports the experimentally arrived at trade-off between structural accuracy and perceptual realism.

These results provide a clear visualization of the impact of different architectures used on the performance of improving the resolution of satellite images, showing how each approach affects the quality of reconstruction and visual perception.

5.2. Sequential Pipeline Results

To explore the potential of leveraging the complementary strengths of the models, several two- and three-phase hybrid pathways have been designed, connecting the models sequentially and in multiple arrangements. Table 3 shows that integrating models (e.g., SwinIR $\leftarrow$ ESRGAN, LDM $\leftarrow$ ESRGAN, and SwinIR $\leftarrow$ LDM) led to minor quantitative improvements on some measures, as reflected in the quantitative results for two-phase and three-phase hybrid tracks. The findings show that sequencing models yield only slight improvements over individual models. As illustrated in Table 3, the SwinIR LDM pipeline provides a small gain in FID (185.07) over SwinIR only, while achieving a relatively high PSNR (24.19 dB). Likewise, LDM $\rightarrow$ ESRGAN yields one of the lowest FID values (183.87) but with no significant improvements in PSNR or SSIM. Three-stage pipelines fail to show any consistent advantages over two-stage configurations. In most cases, PSNR and SSIM values decrease marginally compared with the optimal standalone model (SwinIR), and FID gains are small. These results indicate that additive performance gains from sequentially stacking models are not inevitable, presumably due to error propagation and learned redundancy.

Table 3. sequential pipelines, two and three stages.

fPipeline	PSNR (dB)	SSIM	FID
Two-stage pipelines			
ESRGAN → SwinIR	20.72	0.3730	240.63
ESRGAN → LDM	23.72	0.5133	188.60
SwinIR → ESRGAN	24.70	0.5956	200.86
SwinIR → LDM	24.19	0.5370	185.07
LDM → ESRGAN	23.72	0.5071	183.87
LDM → SwinIR	24.06	0.5265	191.06
Three-stage pipelines			
ESRGAN → SwinIR → LDM	23.67	0.5134	192.47
ESRGAN → LDM → SwinIR	23.84	0.5108	200.35
SwinIR → ESRGAN → LDM	23.99	0.5287	192.08
SwinIR → LDM → ESRGAN	23.96	0.5227	185.16
LDM → ESRGAN → SwinIR	23.76	0.5034	186.63
LDM → SwinIR → ESRGAN	23.80	0.5043	189.95

Figure 13 shows Model order is a factor which affects performance without providing any significant improvement.

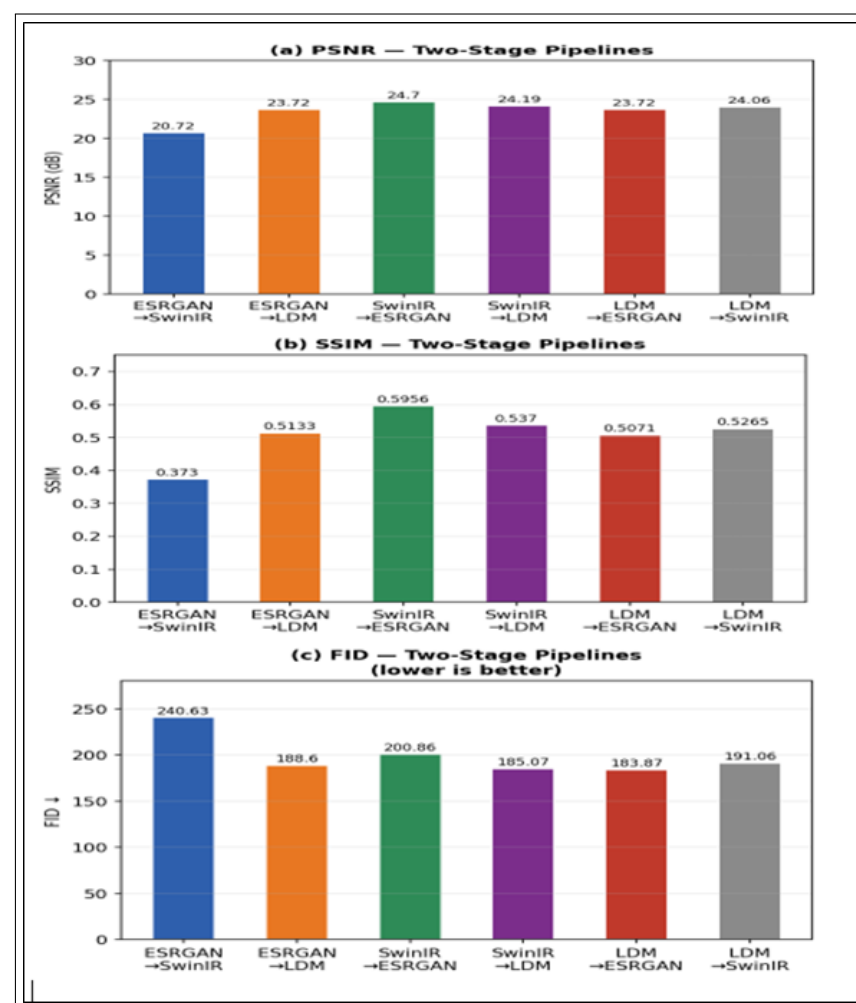


Figure 13. Shows a quantitative comparison of models on satellite image super-resolution using a two-stage pipeline.

As shown in Figure 13, while some two-stage hybrid pipeline combinations marginally improve perceptual quality (FID), Reconstruction quality, in terms of PSNR and SSIM metrics, does not improve significantly when compared to the highest performing individual model.

Similarly, As shown in Figure 14, moving onto three stage architectures does not result in proportional improvements; rather, there is an instance where one metric decreases slightly.

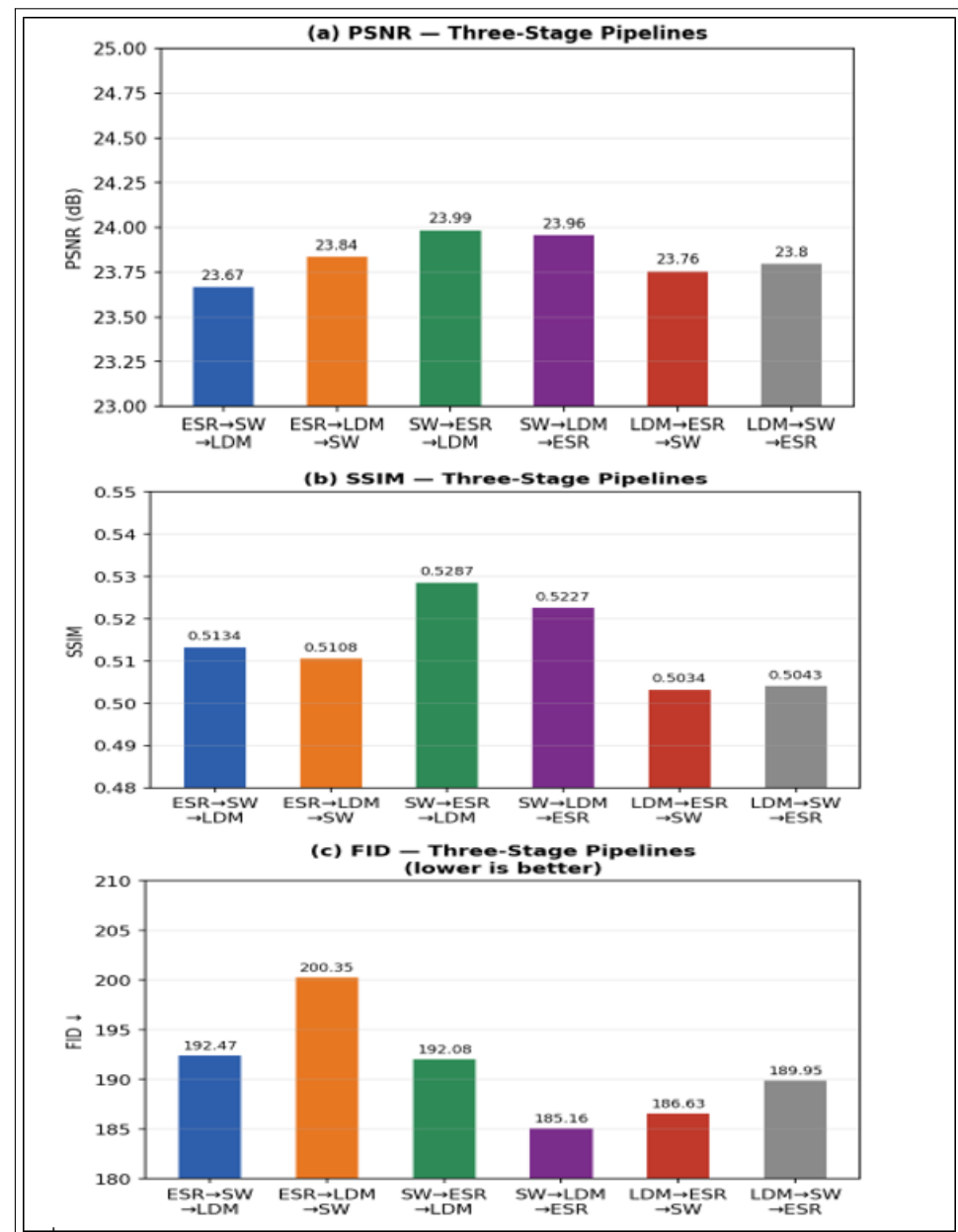


Figure 14. Shows a quantitative comparison of models on satellite image super-resolution using a three-stage pipeline.

Overall, the hybrid pipelines showed only small quantitative improvements but made the computation more complex.

5.3. Ensemble Pipeline (Averaging Outputs)

Since sequential hybrid pipelines yielded only small improvements while significantly increasing computational cost, the next step was to seek a better way to combine them. So,

an ensemble averaging method was used to combine the strengths of each model without the extra work of multi-stage cascades, as in Table 4.

Table 4. summarizes the results of this ensemble pipeline.

Model	PSNR (dB)	SSIM	FID
Pipeline_ensemble	24.22	0.5385	190.81

As shown in Table 4, the ensemble model has a PSNR of 24.22 dB and an SSIM of 0.5385, which suggests a decent reconstruction quality. But the FID score of 190.81 implies poor visual quality. As such, these results suggest that the ensemble method does not yield significant gains in performance, and introduce further complexity that may impact the quality of the reconstructions. To further illustrate and support these observations, Figure 15 shows a quantitative comparison of the ensemble pipeline with other models.

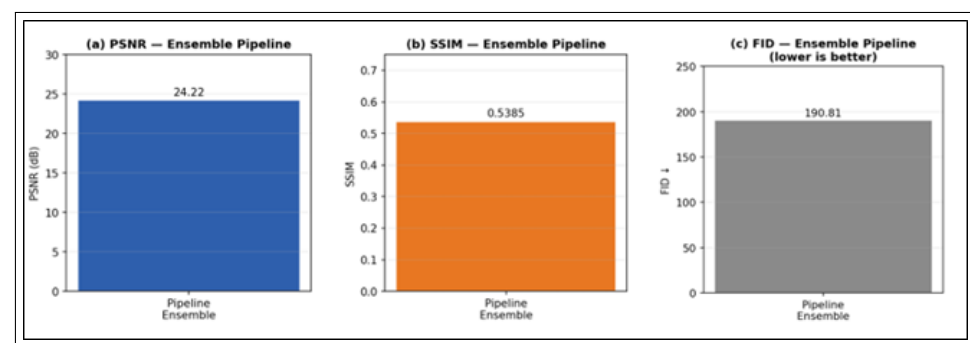


Figure 15. Shows a quantitative comparison of models using the Ensemble pipeline for satellite image super-resolution.

The results show that hybrid/ensemble methods do not offer significant benefits over well-optimized single-model pipelines, particularly in terms of cost-effectiveness. So, we redirected our focus toward enhancing individual models through parameter-efficient fine-tuning using Low-Rank Adaptation (LoRA).

5.4. Fine-Tuning Using LoRA

This table illustrates that SwinIR achieved the highest PSNR and SSIM values, indicating superior reconstruction quality. LDM obtained the lowest FID score, reflecting better perceptual realism. In contrast, ESRGAN recorded the lowest PSNR and SSIM with the highest FID.

The results of the LoRA-fine-tuned models show significant gains across all assessment measures compared to the baseline, as shown in Table 5. The SwinIR with LoRA achieves the best reconstruction fidelity, with PSNR 36.20 dB and SSIM 0.8650, guaranteeing high structural preservation and pixel-level accuracy. The perceptual realism in LDM is impressive, as that amount reduces FID (44.96).

Table 5. Fine-tuning results show significant performance improvement for baseline models, demonstrating LoRA's adaptability for satellite imagery.

LoRA Model	PSNR (dB) ↑	SSIM ↑	FID ↓
LDM	33.85	0.8484	44.96
SwinIR	36.20	0.8650	46.10
ESRGANs	34.64	0.8386	52.30

Similarly, ESRGAN adapted to LoRA achieves significant improvements in PSNR and SSIM and significantly reduces FID compared to the baseline counterpart. All in all, the findings show that LoRA is an effective method for making models more adaptable to satellite images by enabling them to refine task-specific parameters rather than the entire set. The scale of the enhancement across all three architectures points to the efficiency of parameter-efficient fine-tuning for remote sensing super-resolution.

As shown in Figure 16, the performance comparison among LoRA-fine-tuned ESRGANs, SwinIR, and LDM models demonstrates the impact of fine-tuning on image reconstruction quality.

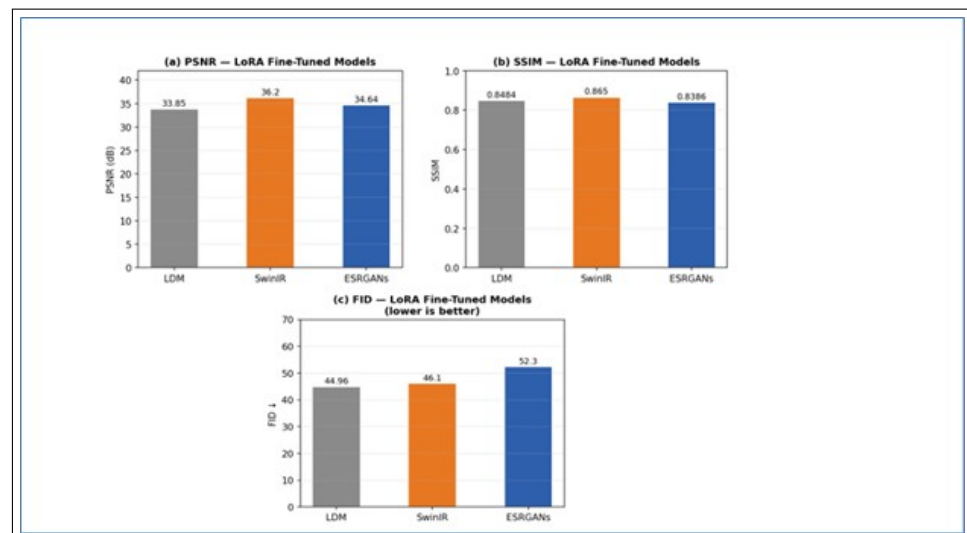


Figure 16. Performance comparison of LoRA fine-tuned models ((a) PSNR, (b) SSIM, and (c) FID metrics. Each metric uses an independent y-axis for accurate representation.

The performance improvement occurs after they use LoRA fine-tuning. Figure 17 clearly indicates a large positive change in PSNR and SSIM, and a steep decline in FID, across all models compared to their baseline values. The steady upward trend across the various architectural paradigms, namely GAN-based, transformer-based, and diffusion-based, makes it evident that LoRA is a model-agnostic, robust adaptation strategy. The results also confirm that the low-rank parameter estimates have high potential to improve both reconstruction fidelity and perceptual quality in super-resolving satellite images.

As shown in Figure 17: Comparison of image restoration performance across ESRGANs, SwinIR, and LDM models before and after LoRA fine-tuning.

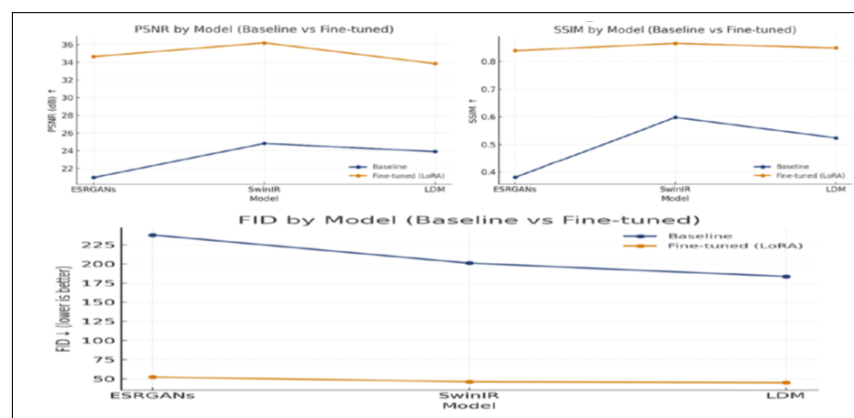


Figure 17. Performance of each subplot illustrates the improvement trend for (a) PSNR, (b) SSIM, and (c) FID.

Directly compare baseline and LoRA fine-tuned models based on the PSNR, SSIM, and FID metrics in Figure 17. Subplot (a) shows that all models achieve a significant gain in PSNR upon fine-tuning, demonstrating a substantial enhancement in pixel-level reconstruction accuracy. The highest PSNR is obtained with SwinIR, followed by ESRGAN and LDM. Subplot (b) presents a steady increase in the value of SSIM in all architectures, which indicates the increased structural similarity and the superbness of the spatial details in a satellite image. The trend of improvement is also consistent, indicating LoRA's success across various model designs. Most notably, subplot (c) shows a melting pot of FID scores after fine-tuning, indicating much better perceptual realism and a closer proximity to real image distributions. The amount of FID reduction in all models indicates the effectiveness of LoRA in enhancing the generators without compromising their structure. Overall, the visual patterns in Figure 17 clearly show that the fine-tuning of the LoRA achieves consistent and improved results in the GAN, transformer, and diffusion-based architectures.

As shown in Table 6, we observe that SwinIR shows the best reconstruction quality, with the highest PSNR (36.20 ± 1.05 dB) and SSIM (0.8650 ± 0.0381) values; LDM, on the other hand, has the best perceptual quality with the smallest FID (44.96 ± 3.18). ESRGANs, on the other hand, performs worse in terms of SSIM and has a higher FID, indicating less perceptual quality. The baseline models suffer from much lower performance across all the metrics, which demonstrates that the improved models are effective.

Table 6. Descriptive Statistics of Model Performance (PSNR, SSIM and FID).

Model	Metric	Mean $\pm$ Std	95% t-CI	Bootstrap CI	Median
LDM	PSNR (dB)	33.85 ± 1.03	[33.74, 33.96]	[31.92, 35.95]	33.80
LDM	SSIM	0.8484 ± 0.0403	[0.8438, 0.8530]	[0.7618, 0.9179]	0.8487
LDM	FID	44.96 ± 3.18	[44.60, 45.32]	[38.20, 50.45]	44.99
SwinIR	PSNR (dB)	36.20 ± 1.05	[36.08, 36.32]	[34.06, 38.18]	36.28
SwinIR	SSIM	0.8650 ± 0.0381	[0.8607, 0.8693]	[0.7875, 0.9366]	0.8679
SwinIR	FID	46.10 ± 4.75	[45.56, 46.64]	[36.98, 55.86]	45.74
ESRGANs	PSNR (dB)	34.64 ± 1.07	[34.52, 34.76]	[32.51, 36.89]	34.65
ESRGANs	SSIM	0.8386 ± 0.0386	[0.8342, 0.8430]	[0.7615, 0.9198]	0.8389
ESRGANs	FID	52.30 ± 4.52	[51.79, 52.81]	[43.97, 61.23]	52.56
LDM baseline	PSNR (dB)	23.90 ± 1.03	[23.78, 24.02]	[21.88, 25.92]	23.93
LDM baseline	SSIM	0.5240 ± 0.0383	[0.5196, 0.5284]	[0.4489, 0.5991]	0.5243
LDM baseline	FID	183.72 ± 13.03	[182.44, 185.00]	[157.66, 209.78]	183.65
SwinIR baseline	PSNR (dB)	24.82 ± 1.04	[24.70, 24.94]	[22.78, 26.86]	24.85
SwinIR baseline	SSIM	0.5976 ± 0.0377	[0.5933, 0.6019]	[0.5237, 0.6715]	0.5979
SwinIR baseline	FID	201.10 ± 20.73	[199.06, 203.14]	[159.64, 242.56]	200.84
ESRGANs baseline	PSNR (dB)	20.96 ± 1.02	[20.84, 21.08]	[18.96, 22.96]	20.99
ESRGANs baseline	SSIM	0.3804 ± 0.0369	[0.3762, 0.3846]	[0.3080, 0.4528]	0.3807
ESRGANs baseline	FID	237.81 ± 20.45	[235.81, 239.81]	[196.91, 278.71]	237.54

Although Table 6 describes the performance differences between the models, it does not show whether these differences are significant. Therefore, Table 7 is added to confirm these observations with pairwise statistical significance tests (Wilcoxon signed-rank test and *t*-test).

As shown in Table 7, the results confirm that all performance differences are statistically significant ($p < 0.05$), which shows that the improvements are not due to random variation.

In addition to reconstruction accuracy and perceptual quality, inference time is evaluated to measure the computational efficiency of all model configurations.

Table 7. Pairwise Statistical Significance Tests Among Models (Wilcoxon and *t*-test *p*-values).

Model A	Model B	Metric	Mean A	Mean B	Wilcoxon <i>p</i>	<i>t</i> -test <i>p</i>	Significant
LoRA Fine-Tuned Models							
LDM	SwinIR	PSNR (dB)	33.85	36.20	0.0001	0.0001	Yes
LDM	SwinIR	SSIM	0.8484	0.8650	0.0001	0.0001	Yes
LDM	SwinIR	FID	44.96	46.10	0.0003	0.0003	Yes
LDM	ESRGANs	PSNR (dB)	33.85	34.64	0.0001	0.0001	Yes
LDM	ESRGANs	SSIM	0.8484	0.8386	0.0034	0.0031	Yes
LDM	ESRGANs	FID	44.96	52.30	0.0001	0.0001	Yes
SwinIR	ESRGANs	PSNR (dB)	36.20	34.64	0.0001	0.0001	Yes
SwinIR	ESRGANs	SSIM	0.8650	0.8386	0.0001	0.0001	Yes
SwinIR	ESRGANs	FID	46.10	52.30	0.0001	0.0001	Yes
LoRA Fine-Tuned vs. Baseline Models							
LDM + LoRA	LDM baseline	PSNR (dB)	33.85	26.82	0.0001	0.0001	Yes
LDM + LoRA	LDM baseline	SSIM	0.8484	0.7585	0.0001	0.0001	Yes
LDM + LoRA	LDM baseline	FID	44.96	183.72	0.0001	0.0001	Yes
SwinIR + LoRA	SwinIR baseline	PSNR (dB)	36.20	28.27	0.0001	0.0001	Yes
SwinIR + LoRA	SwinIR baseline	SSIM	0.8650	0.8038	0.0001	0.0001	Yes
SwinIR + LoRA	SwinIR baseline	FID	46.10	201.10	0.0001	0.0001	Yes
ESRGANs + LoRA	ESRGANs baseline	PSNR (dB)	34.64	27.50	0.0001	0.0001	Yes
ESRGANs + LoRA	ESRGANs baseline	SSIM	0.8386	0.7832	0.0001	0.0001	Yes
ESRGANs + LoRA	ESRGANs baseline	FID	52.30	237.81	0.0001	0.0001	Yes

The basic models exhibit average processing times, with the ESRGAN model being the most computationally efficient, while the LDM model has the highest response time due to its iterative propagation process showns at Table 8 and Figure 18.

Table 8. Complete Performance and Inference Time Comparison Across All Configurations.

Model/Configuration	PSNR (dB) ↑	SSIM ↑	FID ↓	Inference Time (s)
Baseline Models				
ESRGAN	20.96	0.3804	237.81	26.85
SwinIR	24.82	0.5976	201.10	33.57
LDM	23.90	0.5240	183.72	81.35
LoRA Fine-Tuned Models				
ESRGAN + LoRA	34.64	0.8386	52.30	28.5
SwinIR + LoRA	36.20	0.8650	46.10	36.2
LDM + LoRA	33.85	0.8484	44.96	89.5
Sequential Pipelines – Two-Stage				
ESRGAN → SwinIR	20.72	0.3730	240.63	303.45
ESRGAN → LDM	23.72	0.5133	188.60	88.43
SwinIR → ESRGAN	24.70	0.5956	200.86	96.04
SwinIR → LDM	24.19	0.5370	185.07	95.14
LDM → ESRGAN	23.72	0.5071	183.87	145.73
LDM → SwinIR	24.06	0.5265	191.06	353.11
Sequential Pipelines – Three-Stage				
ESRGAN → SwinIR → LDM	23.67	0.5134	192.47	366.45
ESRGAN → LDM → SwinIR	23.84	0.5108	200.35	365.58
SwinIR → ESRGAN → LDM	23.99	0.5287	192.08	156.10
SwinIR → LDM → ESRGAN	23.96	0.5227	185.16	157.47
LDM → ESRGAN → SwinIR	23.76	0.5034	186.63	420.98
LDM → SwinIR → ESRGAN	23.80	0.5043	189.95	415.60
Pipeline_Ensemble	24.22	0.5385	190.81	157.14

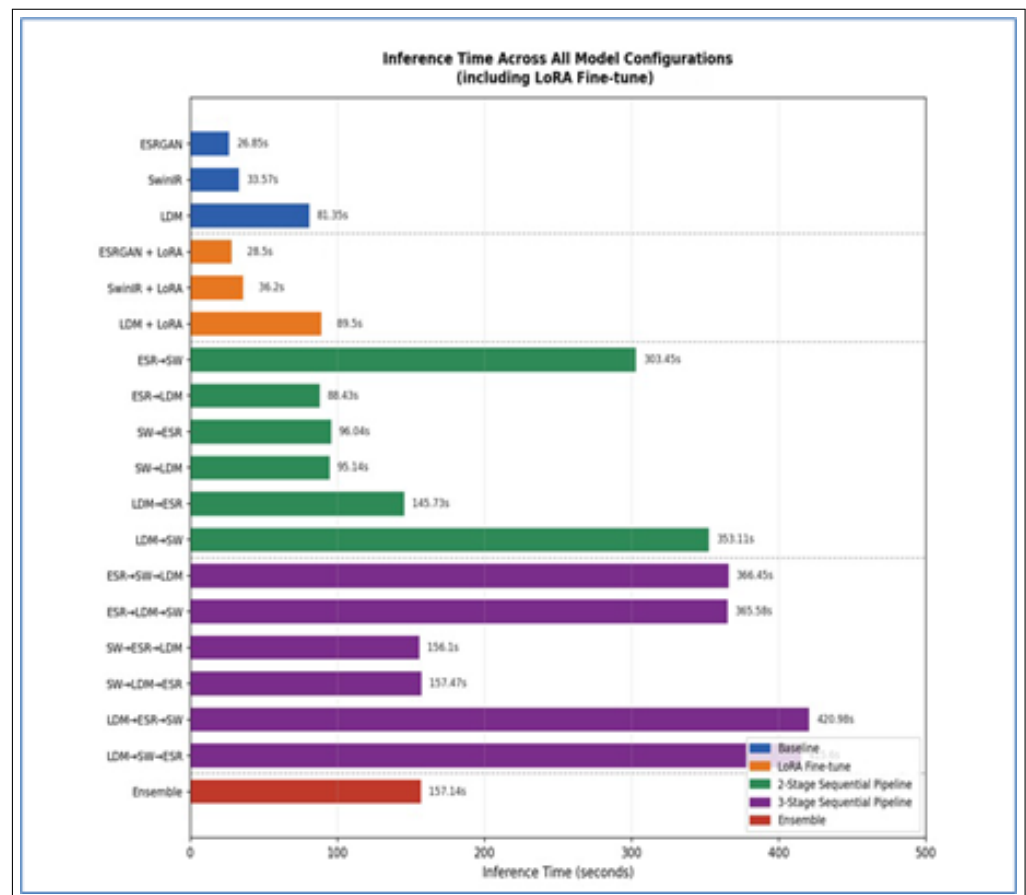


Figure 18. Inference time across all model configurations.

It is worth noting that models optimized with LoRA Technology achieve a noticeable improvement in reconstruction performance with only a slight increase in inference time, reflecting their high efficiency. In turn, serial processing line configurations lead to a significant increase in inference time due to the cumulative implementation of multiple models, often without commensurate performance gains. This effect is increased in multi-stage processing lines, which impose a high computational load, limiting their suitability for immediate or time-sensitive applications such as disaster monitoring and precision agriculture. Although the ensemble approach provides a balance between performance and efficiency, it still requires additional computational cost compared to individual models. Overall, these results show that LoRA-based models offer the best balance between Reconstruction quality and computational efficiency, making them more suitable for practical applications.

5.5. Discussion

Although LoRA reduces training complexity by limiting the number of learnable parameters, it does not decrease inference time, as the additional low-rank computations are performed during execution. As shown in Table 8, this results in only a marginal increase in inference time—for example, from 26.85 to 28.5 s for ESRGAN, from 33.57 to 36.2 s for SwinIR, and from 81.35 to 89.5 s for LDM. Despite the relatively small added cost, LoRA delivers significant improvements in reconstruction performance, with marked increases in PSNR and SSIM and a decrease in FID. Finally, models using LoRA are still much more efficient than sequential processing, which incurs much higher costs. In this way, LoRA strikes a good compromise between improved performance and efficiency, making it a useful approach for fine-tuning super-resolution models.

One of the potential confounders when analyzing the fine-tuning performance results is the fact that any increase in performance may potentially result from the training itself, not necessarily from LoRA's adaptation technique. As baseline models have been trained in limited settings where they have not converged fully, there is still room for improvement in their performance through training, which could be achieved even without LoRA. Conducting an ablation study as a control group by training the models for a similar number of iterations but without LoRA modules could provide stronger causality. While being aware of this limitation, we recognize it as one of the most promising directions to pursue further. Nevertheless, along with significant performance improvements, the number of trainable parameters (less than 3.1%) of the overall model parameters) is clearly indicative of an additional inductive bias introduced by LoRA's technique.

To validate these observations, a comprehensive statistical analysis was conducted across the test dataset. Evaluation metrics are reported alongside standard deviation and confidence intervals to ensure the reliability of the results. In addition, pairwise statistical significance tests confirm that the observed improvements are statistically meaningful, indicating that LoRA's performance gains are consistent rather than attributable to random variation.

The experimental results yield several noteworthy observations. First and foremost, hybrid sequential pipelines provide only marginal improvements in performance while incurring considerable computational overhead in the context of satellite image super-resolution. Second, ensemble inference with output averaging tends to introduce unwanted over-smoothing artifacts. In contrast, LoRA-based parameter-efficient fine-tuning proves highly effective while demanding minimal computational resources. The results obtained from fine-tuning through the LoRA method always prove to be effective regardless of the type of model architecture used. In general, LoRA fine-tuning is superior to model combination techniques, and it seems to be an ideal technique when it comes to fine-tuning restoration and generation-based super-resolution models.

The significant improvement in the PSNR score when using the LoRA fine-tuning technique, compared to traditional super-resolution techniques, may be due to various reasons.

First, the baseline models were not originally adapted to the target dataset, which limits their capacity to reconstruct domain-specific images accurately. LoRA fine-tuning addresses this limitation by enabling the model to learn the statistical characteristics of the dataset effectively.

Second, the dataset comprises specialized images that could be quite different from those seen by the model during the pre-training phase. Using a pre-trained model on this new dataset tends to produce poor results, but through fine-tuning using LoRA, the model learns to adjust to the specialized nature of the data, which produces much better results for image reconstruction.

Third, the use of LoRA allows the model to build on prior experience while adjusting only a few parameters.

Regarding trainable parameters: in full fine-tuning, all model parameters are updated during training. By contrast, LoRA injects low-rank adaptation matrices into selected layers—such as attention layers—while keeping the original pre-trained weights frozen. This approach optimizes only a small fraction of the total parameters, greatly reducing training cost. With respect to training time, since LoRA modifies only a limited number of parameters, the backward pass is more efficient. Computing gradients and performing optimizer updates requires substantially less computation than full fine-tuning, thereby reducing total training time.

In terms of GPU memory usage, LoRA reduces memory consumption by storing gradients only for the low-rank matrices rather than for the entire model. This allows

training with larger batch sizes or on hardware with limited memory capacity. As shown in Table 9, a detailed comparison between full fine-tuning and LoRA-based fine-tuning is provided for each super-resolution model, covering total parameters, trainable parameters and their ratio, and GPU memory requirements under each training strategy.

Table 9. Comparison of full fine-tuning and LoRA fine-tuning for different models.

Model	Total Parameters	Trainable Parameters	Trainable Ratio	GPU Memory	Trainable Parameters (LoRA)	Trainable Ratio (LoRA)	GPU Memory (LoRA)
ESRGAN	~16 M	16 M	100%	8.1 GB	0.5 M	3.1%	6.2 GB
SwinIR	~11.8 M	11.8 M	100%	9.3 GB	0.35 M	2.9%	7.4 GB
LDM (Super Resolution)	~860 M	860 M	100%	19.8 GB	10 M	1.16%	14.5 GB

As shown in Table 9, full fine-tuning requires updating all model parameters, which incurs considerable computational and memory costs. In comparison to this, LoRA requires the training of just about 1–3% of the parameters. Besides saving on the GPU memory required, this strategy not only decreases the time taken to train but also improves the performance of the pre-trained model. The benefits of LoRA are most evident in larger models like LDM, where the parameter and memory reductions are much more significant.

6. Conclusions

In this study, we provide a comprehensive assessment of deep learning approaches for satellite image super-resolution by focusing our analysis on three popular models, ESRGAN, SwinIR and Latent Diffusion Models, both individually and in combination with each other by means of hybrid sequential architectures, ensemble approaches, and parameter-efficient fine-tuning via LoRA.

The results obtained show that both approaches increase the computational cost and do not guarantee an improvement in the quality of reconstruction. In contrast, we show that LoRA fine-tuning consistently improves to give an improved reconstruction result for any of the models under consideration. SwinIR with LoRA results in better structural similarity while LDM with LoRA results in more perceptually realistic outputs. The findings appear promising but there are some limitations to be noted. The dataset used do not contain all possible types present in the real satellite image, which includes negative effects of climate, special properties of different sensors and issues related to the fusion of different types of data obtained from multiple sensors.

The advantages of LoRA are obvious, but little information is available about its performance on different types of backbone architectures, especially in the context of the most recent diffusion-based neural networks. In order to evaluate model performance, large computational resources are required making exhaustive hyperparameter optimization impractical.

Future directions to explore are diffusion based models (e.g., Latent Diffusion Models) and large scale transformer based models (e.g., SwinIR). Moreover, adaptive/dynamic forms of LoRA and other parameter-efficient model tuning methods, such as adapter modules and selective fine-tuning schemes, may potentially improve the performance in high-resolution satellite image super-resolution tasks.

Author Contributions: Conceptualization, N.R.M and H.B.A.M.; methodology, N.R.M.; software, N.R.M., H.E., H.B.A.M.; validation, N.R.M., H.E., and N.R.M.; formal analysis, N.R.M.; investigation, H.B.A.M.; resources, H.B.A.M., H.E.; data curation, N.R.M and H.B.A.M.; writing—original draft preparation, N.R.M., H.B.A.M.; writing—review and editing, visualization, B.A.F.Y, H.E., N.R.M.; supervision, B.A.F.Y., H.E; project administration, B.A.F.Y, H.B.A.M.; All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The datasets generated and analyzed during the current study are available from the corresponding author upon reasonable request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Wang, X.; Yi, J.; Guo, J.; Song, Y.; Lyu, J.; Xu, J.; Yan, W.; Zhao, J.; Cai, Q.; Min, H. A review of image super-resolution approaches based on deep learning and applications in remote sensing. *Remote Sens.* **2022**, *14*, 5423. <https://doi.org/10.3390/rs14215423>.
2. Seydi, S.T.; Arefi, H. A comparison of deep learning-based super-resolution frameworks for Sentinel-2 imagery in urban areas. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2023**, *X-1-W1*, 1021–1028. <https://doi.org/10.5194/isprs-annals-X-1-W1-2023-1021-2023>.
3. Alashour, H.; El Abbadi, N.K. Advances and insights in image texture analysis: A review. *Mesopotamian J. Big Data* **2025**, *2025*, 108–135. <https://doi.org/10.58496/MJBD/2025/008>.
4. Song, J.; Yi, H.; Xu, W.; Li, X.; Li, B.; Liu, Y. ESRGAN-DP: Enhanced super-resolution generative adversarial network with adaptive dual perceptual loss. *Heliyon* **2023**, *9*, e15134. <https://doi.org/10.1016/j.heliyon.2023.e15134>.
5. Safarov, F.; Khojamuratova, U.; Komoliddin, M.; Bolikulov, F.; Muksimova, S.; Cho, Y.-I. MBGPIN: Multi-branch generative prior integration network for super-resolution satellite imagery. *Remote Sens.* **2025**, *17*, 805. <https://doi.org/10.3390/rs17050805>.
6. Lu, T.; Wang, J.; Zhang, Y.; Wang, Z.; Jiang, J. Satellite image super-resolution via multi-scale residual deep neural network. *Remote Sens.* **2019**, *11*, 1588. <https://doi.org/10.3390/rs11131588>.
7. Qi, Y.; Lou, M.; Liu, Y.; Li, L.; Yang, Z.; Nie, W. Advancing image super-resolution techniques in remote sensing: A comprehensive survey. *ISPRS J. Photogramm. Remote Sens.* **2026**, *231*, 68–100. <https://doi.org/10.1016/j.isprsjprs.2025.10.024>.
8. Greza, M.; Bhattacharya, I.; Hoegner, L.; Jutzi, B. GAN-Based Dual Image Super Resolution for Satellite Imagery Decreasing Radiometric Uncertainty. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2024**, *10*, 155–162. <https://doi.org/10.5194/isprs-annals-X-3-2024-155-2024>.
9. Xiao, H.; Wang, X.; Wang, J.; Cai, J.Y.; Deng, J.H.; Yan, J.K.; Tang, Y.D. Single image super-resolution with denoising diffusion GANs. *Sci. Rep.* **2024**, *14*, 4272. <https://doi.org/10.1038/s41598-024-52370-3>.
10. Castillo, A.; Escobar, M.; Pérez, J.C.; Romero, A.; Timofte, R.; Van Gool, L.; Arbelaez, P. Generalized real-world super-resolution through adversarial robustness. In Proceedings of the ICCV Workshops, Montreal, BC, Canada, 11–17 October 2021; pp. 1855–1865. <https://doi.org/10.1109/ICCVW54120.2021.00212>.
11. Chen, Q.; Wang, L.; Zhang, Z.; Wang, X.; Liu, W.; Xia, B.; Ding, H.; Zhang, J.; Xu, S.; Wang, X. Dual-path aggregation transformer network for super-resolution with images occlusions and variability. *Eng. Appl. Artif. Intell.* **2025**, *140*, 109535. <https://doi.org/10.1016/j.engappai.2024.109535>.
12. Singgalen, Y.A.; Sutresno, S.A.; Dasra, M.N.A.; Setiawan, R.W. Implementation of hybrid deep learning CNN model for multispectral satellite image classification in land change detection. *J. Theor. Appl. Inf. Technol. (JATIT)* **2025**, *103*, 2579–2593.
13. Asif, M.; Abrar, M.; Ullah, F.; Salam, A.; Amin, F.; de la Torre, I.; Villar, M.G.; Garay, H. A novel hybrid deep learning approach for super-resolution and object detection in remote sensing. *Sci. Rep.* **2025**, *15*, 1234. <https://doi.org/10.1038/s41598-025-01476-3>.
14. Al-Khafaji, M.; Ramaha, N.T.A. Hybrid deep learning architecture for scalable and high-quality image compression. *Sci. Rep.* **2025**, *15*, 22926. <https://doi.org/10.1038/s41598-025-06481-0>.
15. Donike, S.; Aybar, C.; Gómez-Chova, L.; Kalaitzis, F. Trustworthy super-resolution of multispectral Sentinel-2 imagery with latent diffusion. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2025**, *18*, 6940–6952. <https://doi.org/10.1109/JSTARS.2025.3542220>.
16. Ulku, I.; Tanriover, O.O.; Akagündüz, E. LoRA-NIR: Low-Rank Adaptation of Vision Transformers for Remote Sensing With Near-Infrared Imagery. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 5004505. <https://doi.org/10.1109/LGRS.2024.3449372>.
17. Tian, C.; Shi, Z.; Guo, Z.; Li, L.; Xu, C. Hydralora: An asymmetric lora architecture for efficient fine-tuning. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 9565–9584.
18. Cornebise, J.; Oršolić, I.; Kalaitzis, F. The worldstrat dataset: Open high-resolution satellite imagery with paired multi-temporal low-resolution. *Zenodo* **2022**.

-
19. Ou, W. A Deep Learning-Based Generative Adversarial Network for Digital Art Style Migration. *Int. J. Adv. Comput. Sci. Appl.* **2025**, *16*, 3. <https://doi.org/10.14569/IJACSA.2025.0160363>.

20. Ali, S.; Jamil, U.; Jabbar, M.; Sajid, A.; Jabbar, M.A. Evaluation of PSNR Value for Image Super-Resolution Using Deep Learning. *Lahore Garrison Univ. Res. J. Comput. Sci. Inf. Technol.* **2023**, *7*, 4.
21. Wang, Z.; Simoncelli, E.P.; Bovik, A.C. Multiscale structural similarity for image quality assessment. In *Proceedings of the Thirty-Seventh Asilomar Conference on Signals, Systems & Computers*; IEEE: New York, NY, USA, 2003; Volume 2, pp. 1398–1402. <https://doi.org/10.1109/ACSSC.2003.1292216>.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.