DEPARTMENT OF INFORMATICS

TECHNISCHE UNIVERSITÄT MÜNCHEN

Bachelor's Thesis in Informatics

Investigating Blind Image Super-Resolution of Sentinel-2 Satellite Data and Its Applications

Ron Mühlhaus

DEPARTMENT OF INFORMATICS

TECHNISCHE UNIVERSITÄT MÜNCHEN

Bachelor's Thesis in Informatics

Investigating Blind Image Super-Resolution of Sentinel-2 Satellite Data and Its Applications

Untersuchung der Blind Image Super-Resolution für Sentinel-2-Satellitendaten und deren Anwendungen

Author:	Ron Mühlhaus
Supervisor:	Prof. Dr. Daniel Cremers
Internal Advisor:	Cecilia Curreli
External DLR Advisor:	Sandeep Kumar Jangir
Submission Date:	18.09.2025

I confirm that this bachelor's thesis in informatics is my own work and I have documented all sources and material used.

Munich, 18.09.2025 Ron Mühlhaus

Acknowledgments

I would like to express my sincere gratitude to my advisors, Cecilia Currelli (TUM) and Sandeep Kumar Jangir (DLR), for their dedication and support throughout this thesis. This was an incredible learning opportunity for me that would not have been possible without their guidance.

I would also like to thank Paul Karlshöfer (DLR, Department of Imaging Spectroscopy) for his support in the evaluation of the super-resolution images in the field detection downstream task.

Furthermore, I am grateful to the German Aerospace Center (DLR) for providing me with this opportunity and for granting access to the necessary resources. I gratefully acknowledge the computational and data resources provided through the joint high-performance data analytics (HPDA) project Terrabyte of the German Aerospace Center (DLR) and the Leibniz Supercomputing Center (LRZ).

Finally, I would like to thank my family and friends for their encouragement and support during this journey.

Abstract

High-resolution Earth observation data are crucial for applications such as agriculture, urban planning, and environmental monitoring. Although commercial and expensive satellites can capture sub-meter imagery, open-access alternatives like Sentinel-2 are limited to resolutions around 10m, which is insufficient for many applications. In this thesis, we investigate image super-resolution (SR) as a method to bridge this resolution gap, improving the performance of downstream tasks on freely available satellite data.

We developed two 16-bit single-band datasets with different spatial resolutions, using Sentinel-2 (20m $\rightarrow$ 10m) and VENµS (10m $\rightarrow$ 5m), with the goal of training and benchmarking four different super-resolution methods. To this end, we adapted three transformer models (SwinIR, Mat, PFT) and one diffusion model (EDiffSR) to our unique satellite data. After training them with three different dataset mixes, we evaluated their performance quantitatively utilizing standard reference-based metrics (PSNR, SSIM). With FID and custom-trained NIQE models, we assessed the native upscaling capabilities of all twelve model configurations. In addition, we evaluated their impact on a practical downstream application, a Sentinel-2 field boundary detection.

Our experiments demonstrate that the Transformer models performed well in terms of PSNR and SSIM, as well as in our downstream application, proving the value of using super-resolution as a preprocessing step. EDiffSR achieved sharper and perceptually more realistic imagery, outperforming our Transformers on FID and NIQE, but failed to beat bicubic upsampling on our downstream task. These findings highlight that super-resolution can be used to make low-resolution satellites more competitive against commercial imagery.

Kurzfassung

Hochauflösende Satellitenbilder sind auf dem Gebiet der Landwirtschaft, Stadtplanung und Umweltüberwachung nicht mehr wegzudenken. Während teure kommerzielle Satelliten hochauflösende Bilder erfassen können, sind Open-Access-Alternativen wie Sentinel-2 auf Auflösungen von etwa 10 m GSD beschränkt, was für viele Anwendungen unzureichend ist. In dieser Bachelorarbeit versuchen wir mit Super-Resolution (SR) diese Auflösungsdifferenz zu überbrücken und die Leistung von Folgeaufgaben auf frei verfügbaren Satellitendaten zu verbessern.

Wir haben zwei 16-Bit-Einband-Datensätze mit unterschiedlichen räumlichen Auflösungen erstellt, basierend auf Sentinel-2 (20m $\rightarrow$ 10m) und VENµS (10m $\rightarrow$ 5m), mit dem Ziel, vier verschiedene Super-Resolution-Methoden zu trainieren und zu vergleichen. Zu diesem Zweck passten wir drei Transformer-Modelle (SwinIR, Mat, PFT) und ein Diffusionsmodell (EDiffSR) an unsere einzigartigen Satellitendaten an. Nach dem Training mit drei verschiedenen Datensatz-Kombinationen bewerteten wir quantitativ ihre Leistung anhand standardisierter Referenzmetriken (PSNR, SSIM). Mit FID und maßgeschneiderten NIQE-Modellen evaluierten wir die nativen Hochskalierungsfähigkeiten aller zwölf Modellkonfigurationen. Darüber hinaus bewerteten wir ihren Einfluss auf eine praktische Folgeanwendung: die Erkennung von Feldgrenzen in Sentinel-2-Daten.

Unsere Experimente zeigen, dass die Transformer-Modelle sowohl bei PSNR und SSIM als auch in der Downstream-Anwendung gute Ergebnisse erzielten und damit den Nutzen von Super-Resolution als Vorverarbeitungsschritt belegten. EDiffSR erzielte schärfere visuell realistischere Bilder und übertraf unsere Transformer-Modelle bei FID und NIQE. Jedoch konnte EDiffSR das Bicubic-Upsampling bei unserer weiterführenden Feldgrenzenerkennung nicht überbieten. Diese Ergebnisse zeigen, dass Super-Resolution genutzt werden kann, um niedrig aufgelöste Satellitenbilder wettbewerbsfähiger gegenüber kommerziellen Bilddaten zu machen.

Contents

1 Introduction

1.1 Motivation: The Need for Trustworthy High-Resolution Earth Observation Data

From global systems preventing famine to the navigation apps in our pockets, from monitoring wildfires and glaciers to planning sustainable cities, from forecasting tomorrow's weather to protecting us from catastrophes, satellite imagery powers some of the most vital infrastructure of our modern world. It helps us monitor global conflicts, track the evolution of our climate, guide the placement of critical infrastructure such as cell towers, and optimize agricultural practices that feed billions.

And that is only what is already possible today. The number of Earth Observation (EO) satellites is expected to almost triple over the next decade [1], equipped with better sensors, finer spatial resolutions, and higher revisit frequencies, getting us closer to global real-time coverage. These advances in spatial resolution, sensor technology, and post-processing pipelines are vital to improve the trustworthiness of available data, making it an even more reliable foundation for important decisions. Unfortunately, the main factor holding back applications and research is not the technology but access to the already available resources.

Most of the high-quality imagery is only commercially available and extremely expensive. This creates a divide between governments, large corporations, and specialized research projects working with submeter satellite data, while many scientific projects, smaller companies, and individual researchers have to rely on the limited resources that are openly available. At the heart of this is the European Union's Sentinel-2 program [2], providing global multispectral data open to everyone for free. However, Sentinel-2 only provides a spatial resolution of 10 meters for its visible bands and 20 meters for many others, which is not high enough for many real-world applications. This raises a critical question: How can we overcome these limitations and unlock the full potential of open-access satellite data, such as Sentinel-2?

1.2 Super-Resolution: The Solution for the Resolution Gap?

To bridge the gap between freely available, but lower-resolution imagery and commercial high-resolution products, researchers have started to explore blind image super-resolution. **Super-Resolution (SR)** is the task of reconstructing a high resolution (HR) image from a low resolution (LR) base using simple algorithms or deep learning methods. Traditional interpolation techniques, such as bicubic interpolation, estimate pixel values based on their neighbors, leading to blurry results. Newer methods leverage machine learning techniques

from simple convolutional neural networks (CNN) to state-of-the-art Transformer [3] and Diffusion [4] architectures to produce high-quality results. To answer the natural question, which super-resolution technique will perform best when applied to satellite imagery?, we evaluated three Transformer models, SwinIR [5], MAT [6], and PFT [7], as well as one Diffusion model, EDiffSR [8], highlighting their respective strengths and weaknesses.

'Blind' in the context of super-resolution means that the models do not assume a fixed degradation kernel but can generalize over inputs with varied degradation distributions. This is especially important for working with real-world data, where so many factors influence image quality, including sensor optics and characteristics, atmospheric scattering, noise, compression artifacts, and differences in preprocessing pipelines.

In our context of Earth Observation (EO), super-resolution could enhance satellite imagery from open projects like Sentinel-2, increasing its spatial resolution and lowering its ground sampling distance. Ground Sampling Distance (GSD) is a standard measurement in meters used in the remote sensing community, describing the area of ground that a pixel covers in the real world. For Sentinel-2, the visible bands have a ground sampling distance of 10m, meaning every image pixel stores the data from a 10m x 10m surface area. If we now utilize super-resolution to upsample these images, with a factor of two, the resulting output would reach a GSD of 5m.

The promise of super-resolution has great potential: if reliable, it could make free missions like Sentinel-2 more competitive with much more expensive commercial products. It could build a foundation for applications that require finer spatial detail, such as field-level agriculture or urban analysis. But all these speculations are contingent on one question: Can we trust super-resolution to produce results that are viable for scientific use and real-world downstream applications?

1.3 Is Super-Resolution a Trustworthy Pre-Processing Step for Downstream Applications?

For a scientific domain like remote sensing, it is not enough that a super-resolved image looks realistic, it needs to reflect reality. This is inherently different from many techniques used for ground imagery, which focus on creating the perceptually best-looking image. Trustworthiness in our context means that upsampling techniques must not introduce artifacts or hallucinate details that were never present, such as a misplaced field boundary. This would not help a downstream task, but mislead it. Especially in high-stakes applications such as disaster response, there is no room for hallucinated false positives. But even in domains that might appear less critical, such as agriculture, our misplaced field boundary can have repercussions. It could distort yield predictions, causing resources to be allocated incorrectly. The same applies to every field where decisions are made on the basis of the data. A mistake in the super-resolution preprocessing step will propagate into further tasks.

But how can we prevent these downfalls from happening? We must thoroughly validate the performance of our super-resolution techniques, making sure that the gain outweighs the risks. SR methods have shown strong improvements in perceptual quality and reference-based

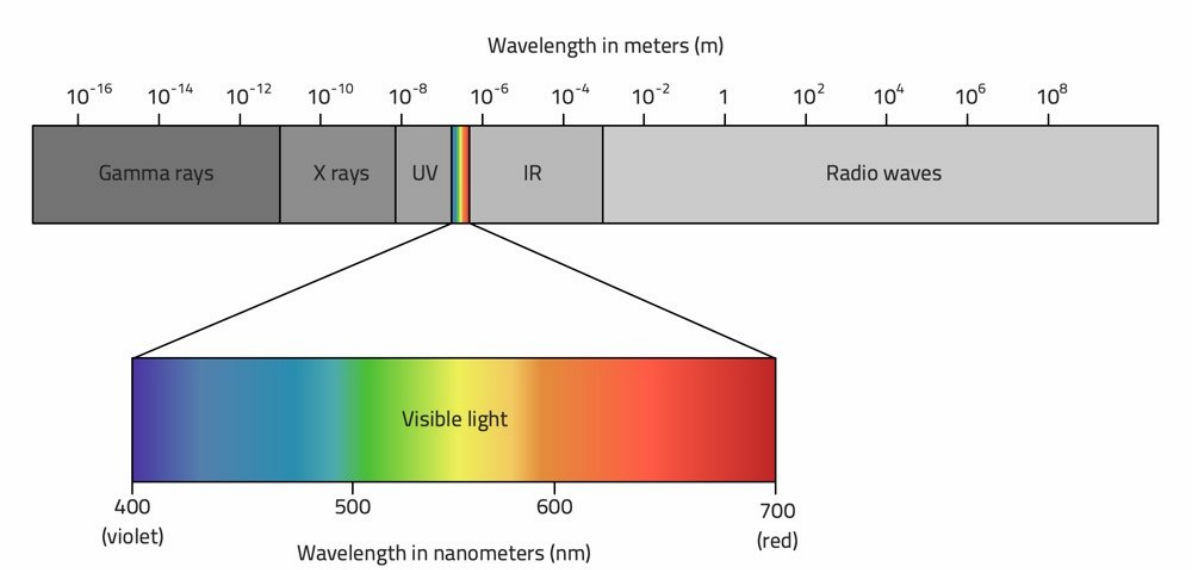

Figure 1.1: Overview over the electromagnetic spectrum. Reproduced from [13].

metrics such as PSNR [9] and SSIM [10]. However, in real-world applications, we lack an image ground truth, which is why we need to find better ways to benchmark the performance of our techniques. This can be addressed by measuring native upscaling performance using no-reference metrics such as FID [11] or NIQE [12]. While these metrics can be used as partial indicators, the only real way to validate the performance of models on downstream tasks is to actually test them on a downstream task and evaluate the results. Even a modest improvement in downstream task performance can justify SR as a viable preprocessing step.

This is exactly the approach taken in this thesis: benchmarking our models not only on PSNR, SSIM, FID, and NIQE, but most importantly on a real-world field boundary detection downstream task. Before turning to these evaluations, we must first understand the unique characteristics and challenges of satellite data itself.

1.4 Challenges in Applying Super-Resolution to Satellite Data

Before we can start getting into the details of satellite data, we first have to understand what the human eye can see. When sunlight reaches the Earth's surface, different materials reflect and absorb it in unique ways. Vegetation, water, soil, or man-made structures reflect a unique spectral signature. The human eye perceives this reflected light as different colors: plants are green, while water is blue. But the reality is that our vision is extremely limited, only capturing a tiny range (roughly 400 - 700 nanometers) of the extensive electromagnetic spectrum (Figure 1.1). Normal cameras are built to mimic our perception and only record images with the three color channels we can see (RGB).

Satellites do not suffer these limitations and are equipped with powerful sensors that can capture a broader range of the electromagnetic spectrum using spectral bands. These bands capture the light intensity in a discrete portion of the spectrum. Sentinel-2 utilizes a multispectral sensor to capture 13 different bands across the spectrum, capturing data from normal visible channels like RGB, but also from near-infrared (NIR) and short-wave infrared (SWIR) sections. These additional bands enable applications that surpass human capabilities, including monitoring vegetation health, estimating soil moisture, or mapping burned areas. To make our super-resolution models work across multiple bands, we do not simply upscale RGB images. Instead, we upscale each band individually, making the model band-agnostic. This allows the model to generalize across different bands, ensuring it works consistently whether it is dealing with the visible spectrum (RGB) or infrared bands (NIR, SWIR).

Unlike normal cameras, satellites do not just capture colors, they measure surface reflectance. This means that each pixel captures how much light is reflected by the surface at a specific wavelength (band). It is vital for further analysis that super-resolution as a pre-processing step does not corrupt this information. Additionally, these surface reflectance measurements are typically stored in a 16-bit data range with 65,536 intensity steps. In comparison, normal RGB ground images utilize 8-bit pixel values with a range of 0-255.

The combination of single-band and 16-bit pixel depth is a significant challenge of this thesis, as it required adapting all models, metrics, and creating our own custom data processing pipelines. Furthermore, most major computer vision libraries and tools do not support 16-bit imagery. A 16-bit data range also requires more complex normalization approaches, which we will go through in depth in Section 3.1.3.

Satellite imagery is typically provided in large tiles that cover extensive geographic areas, such as 10,980 × 10,980 pixels for Sentinel-2 data. These tiles are often too large to process in one go due to memory and computational limitations. To overcome this, satellite images are divided into smaller patches, typically 256 × 256 pixels, which are upscaled individually and recombined again into one tile.

These challenges ranging from spectral consistency and data format to patch-based processing highlight how satellite data and remote sensing super-resolution tasks differ fundamentally from typical natural RGB image super-resolution.

1.5 Contributions

- Developed two satellite datasets for super-resolution tasks, focusing on **16-bit** single-band imagery from **Sentinel-2** and **VENµS**, with two spatial gaps of **20m to 10m** and **10m to 5m**.
- Adapted and trained state-of-the-art super-resolution models **SwinIR**, **MAT**, **PFT**, and **EDiffSR** on three different dataset mixes utilizing different normalization strategies.
- Benchmarked model performances on quantitive metrics **PSNR**, **SSIM** and evaluated their native upscaling capabilities with **FID** and custom trained **NIQE** models.

- Evaluated the super-resolution models on a Sentinel-2 **field boundary detection** downstream task, ensuring their practical applicability for real-world remote sensing applications.

1.6 Thesis Outline

- **Chapter 1: Introduction**
  This chapter sets the stage by discussing the motivation behind remote sensing super-resolution, its use cases, its potential, and its unique challenges.

- **Chapter 2: Related Work**
  In this chapter, we are tracing the evolution of super-resolution from simple algorithms, such as bicubic interpolation, to convolutional neural networks, recent transformer models, and state-of-the-art diffusion architectures. Additionally, we highlight the most important super-resolution research from the remote sensing community, such as relevant datasets/benchmarks and specialized models.

- **Chapter 3: Methodology**
  In the Methodology chapter, we will introduce our two custom datasets, based on Sentinel-2 (20 m $\rightarrow$ 10 m) and VENµS (10 m $\rightarrow$ 5 m). We will look into the architectures of our selected super-resolution models, SwinIR, MAT, PFT, and EDiffSR, explaining how we adapted them to our 16-bit single-band data. We will also highlight all the essential metrics for this thesis: PSNR, SSIM, FID, and NIQE. We conclude the chapter by reviewing the details of our twelve model trainings.

- **Chapter 4: Experiments and Results**
  This chapter presents our experiments, starting with the measurement of inference speeds of our contenders, followed by a quantitative evaluation of our 12 trained models on our Sentinel-2 and VENµS validation datasets, using PSNR and SSIM. In the second major experiment, we test the native upscaling capabilities of our networks with our adapted FID and custom-trained NIQE models on three different setups. With a field-boundary detection downstream task, we will determine the real-world practical applicability of our models to prove the worth of super-resolution as a preprocessing step for remote sensing.

- **Chapter 5: Analysis**
  The discussion chapter interprets the results presented in chapter 4. We will combine the findings of all our experiments, reflecting on the limitations of our different models, and discussing the future of super-resolution in remote sensing.

- **Chapter 6: Conclusion and Future Work**
  The final chapter summarizes the key findings of the thesis, highlights its contributions, and discusses directions for future research in the field of super-resolution for satellite imagery.

2 Related Works

2.1 Evolution of CNN-based Super-Resolution Architectures

The story of Single Image Super-Resolution (SISR) begins with simple mathematical algorithms. One of the most influential first techniques was the **Bicubic Interpolation**. It calculates new pixel values by considering the 16 nearest pixel neighbors in an image. Although it is known to produce blurry, indistinct results, it remains an important baseline for comparing new models to this day.

The Introduction of the **Super-Resolution Convolutional Neural Network (SRCNN)** [14] marked the first major milestone on the journey to using learning-based methods. The model learned, from a paired dataset, a direct mapping between LR and HR images, using a simple Convolutional Neural Network (CNN) architecture. This breakthrough sparked a rapid succession of innovations and optimizations, such as the **FSRCNN** [15], that improved efficiency by moving the upsampling operation to the end of the network. **VDSR** [16], **DRCN** [17], and **SRResNet** [18] started to develop more complex and deeper architectures by incorporating residual connections, resulting in increasingly better PSNR metrics 3.3.1.

However, this focus on the PSNR metric led to new problems. PSNR is based on the mean average pixel error and is known to favor blurry images. Fine details and natural looking textures matter to the human eye, but were often lacking in PSNR-optimized approaches. The **Super-Resolution Generative Adversarial Network (SRGAN)** [18] tried to solve this issue by implementing a novel method to evaluate images, using a Generative Adversarial Network (GAN) [19]. The SRResNet based generator was assessed by a second Evaluator Module, which rates the visual quality based on a novel perceptual loss.

Research on the CNN architecture also led to major improvements. **EDSR** [20] increased performance by simply removing batch normalization layers. Other models introduced novel ways to handle features, from the hierarchical feature fusion in **RDN** [21] to the adaptive feature recalibration in **RCAN's** [22] channel attention mechanism.

The pursuit of perceptual quality reached a new peak with **ESRGAN** [23], a significant improvement to SRGAN. It utilizes a more powerful Residual-in-Residual Dense Block (RRDB) for its generator, which helps with data flow. Its relativistic discriminator follows the new approach of judging whether a real image is more realistic than a fake one, resulting in sharper edges and more convincing details.

As most of these models were trained on simple bicubic degradation, they struggled to handle the complex artifacts of real-world images. **Real-ESRGAN** [24] addressed this issue by training the ESRGAN architecture on a more realistic degradation model. It uses blur, noise, and compression artifacts to mimic the complex real world. With these improved data augmentations, Real-ESRGAN achieved unprecedented generalization to real-world images,

making it, to this day, a decent tool for image enhancement. A more comprehensive overview of CNN-based approaches can be found in Jangir's master's thesis [25].

2.2 Evolution of Transformer-based Models

The introduction of the Transformer architecture marked a paradigm shift, first in Natural Language Processing (NLP), where models like GPT-3 [26] amazed the world and redefined what was possible. It also had a profound impact on computer vision and super-resolution techniques. In this section, we trace the evolution of Transformer-based models, from their foundational concepts to the current state-of-the-art methods of Super Resolution (SR). The survey by Dutta et al. [27] provides a useful performance overview of many foundational models.

2.2.1 "Attention Is All You Need": The Foundation of the Transformer

"Attention Is All You Need" [3] was the groundbreaking paper called, that introduced the Transformer architecture and its novel self-attention mechanism. Self-attention is a powerful mechanism that lets the model see the bigger picture. While CNNs are limited to analyzing a local neighborhood, self-attention allows to take the entire input into consideration. It works by calculating, for each element of the input, which other elements are most closely related. To do this, the model uses three vectors: a Query (Q), a Key (K), and a Value (V). The Query can be thought of as a question that the current element asks. The Keys represent possible answers provided by all other elements. By comparing the Query to the Keys, the model produces similarity scores, which are then normalized. These scores indicate how much attention the model should pay to each value. Finally, the current element is updated by combining the values, weighted by their relevance, leading to a richer representation that captures the context of related elements [28].

The paper also introduced **Multi-Head Attention**, a technique that runs the self-attention mechanism multiple times in parallel. All the results of these parallel attention heads are combined together in order to allow the model to consider multiple diverse perspectives and capture more complex information.

This powerful architecture was first successfully adapted for vision tasks by the **Vision Transformer (ViT)** [29]. It demonstrated how a pure Transformer architecture could achieve state-of-the-art performance in image classification, a field dominated by CNN-based methods. To work with the Transformer architecture, images were split into fixed-sized patches that were treated as sequence elements, similar to words in a sentence. The advantage of this method over the CNN-based method was that it could utilize the context of the entire picture, rather than being limited to local pixel patterns. The main limitation, holding the transformer architectures back, was the massive dataset and computing power required for training.

2.2.2 Adapting Transformers for Image Super-Resolution

ViT was developed for image classification, a task fundamentally different from super-resolution. Super-Resolution (SR) requires working with fine details on high-resolution feature maps. The massive computational cost that is needed for global self-attention did not scale well with this problem [29]. A more efficient solution was required.

This breakthrough came with **SwinIR** [5], a model that made Transformers truly effective for super-resolution. SwinIR solved the computational bottleneck of ViT with a clever trick. Instead of calculating attention across the entire image, it worked on small, non-overlapping local windows. Computing self-attention in these smaller patches massively improves performance. To maintain the benefits of global attention, the famous shifted-window mechanism was introduced. It allows the information to flow between windows across layers. This hierarchical design combines the best of both worlds: global context with fast compute and local focus. SwinIR is to this day the most well-known image transformer for super-resolution, and a common baseline for experiments. As it is one of the core models investigated in this thesis, its architecture is detailed further in Section 3.2.2.

2.2.3 Evolution After SwinIR: Refining Attention

After SwinIR established a strong baseline, the following years saw a rapid evolution of Transformer-based SR models, primarily focused on refining the attention mechanism.

In 2022, research on efficiency continued with the **Efficient SR Transformer (ESRT)** [30]. This lightweight hybrid model combined a CNN backbone for local features with a light Transformer for global dependencies. The **Deformable Attention Transformer (DAT for RefSR)** [31] introduced a more flexible attention. Reference-based SR enhances an LR image by transferring textures and patterns from an additional HR reference image. As DAT is specifically designed for this task, it can adaptively focus on relevant features, proving highly effective for transferring texture.

Rapid developments and many unique approaches were introduced in the following year (2023), starting with an efficient contender, the **Slide-Transformer** [32]: It used highly optimized convolutions to implement local attention, which runs efficiently on any platform. At the same time, new feature aggregation strategies emerged. The **Dual Aggregation Transformer (DAT)** [33] proposed two different types of alternating attention blocks: a standard spatial window attention block for capturing local patterns, and a novel channel-wise attention block. Another frontier to explore was the Frequency Domain, the **Attention Retractable Frequency Fusion Transformer (ARFFT)** [34] utilized the Fast Fourier Transform (FFT) for feature extraction. Both of these attention approaches focus on providing the model with more global information. But the best benchmark results in 2023 were achieved by the **Hybrid Attention Transformer (HAT)** [35]. It uses the complementary advantages of an overlapping cross-attention module and channel attention. The core goal of this hybrid design is to "activate more pixels" and to gain access to a wider context.

Researchers in 2024 pursued multiple avenues for architectural improvement. Deep architectures often suffer from an "information bottleneck", which the **Dense-Residual-Connected**

Transformer (DRCT) [36] addresses by proposing dense residual connections. They stabilize the data flow and counteract diminishing feature maps in late layers. Combining multiple different attention approaches was demonstrated by the **Multi-Attention Fusion Transformer (MAFT)** [37]. It is parallel approach fused local, global, and CNN-based attention. At the same time, a more fundamental shift occurred as research began to merge Transformers with diffusion models, a trend exemplified by the **Diffusion Transformer for SR (DiT-SR)** [38]. We look deeper into diffusion models in Section 2.3.

The **Multi-Range Attention Transformer (MAT)** [6] captures features across multiple spatial scales without significantly increasing computational cost. This allows the model to achieve state-of-the-art performance. As one of the strongest options available today, the **Progressive Focused Transformer (PFT)** [7] enhances efficiency by progressively focusing attention. We will go more in-depth into MAT and PFT in Sections 3.2.3 and 3.2.4, as we picked them as contenders for our experiments.

2.3 Evolution of Diffusion Models

Although Transformers have been a dominant force in super-resolution, **Diffusion Models** [4] have recently emerged as powerful competitors [39]. This section provides an overview of their development, from the core principle to the currently relevant application in super-resolution. A complete technical and theoretical overview of diffusion models in SR can be found in the comprehensive survey by Moser et al. [39].

2.3.1 Foundational Concepts: Denoising Diffusion Probabilistic Models

The success of most modern image generators, such as Stable Diffusion [40] and Midjourney [41], can be traced back to the foundational paper on **Denoising Diffusion Probabilistic Models (DDPM)** [4]. It described the main mechanism behind diffusion models: generating images by progressively removing noise.

The core concept consists of two processes. First, a **forward process** adds incrementally a small amount of Gaussian noise to an image. This procedure slowly degrades the image until it is reduced to pure isotropic noise. To control the rate of information loss, the amount of noise introduced at each time step is carefully managed by a specific noise schedule. Commonly, it takes hundreds to thousands of steps for this process to finish.

The generative power of the model lies in the **reverse process**, where it is attempting to remove noise iteratively. A neural network, typically a U-Net architecture, is trained to predict the noise that was added at any given timestep. By removing this predicted noise, it learns to reverse the forward process. Once trained, this denoising function can be applied iteratively, starting with a random noise sample instead of a corrupted image. After many steps, the random noise is completely transformed into a new coherent image. The results of this method look perceptually convincing, producing high-fidelity images that challenged the quality of leading GANs at the time.

2.3.2 Applying Diffusion to Super-Resolution

The challenge in using diffusion for super-resolution is guiding the generative process. Instead of creating a random image from pure noise, the model needs to generate a high-resolution (HR) output that is consistent with a given low-resolution (LR) input. One of the first and most influential works to demonstrate this was **SR3** [42]. The key innovation: concatenating the LR image to the noisy input at each step of the reverse process. The model can learn to utilize this additional input to ensure that the output remains faithful to the LR source. This conditioning strategy led to a significant leap forward, already producing highly realistic details and textures. It once again surpassed the quality of contemporary GAN based methods in perceptual quality, unfortunately with a high computational cost. Published around the same time, **SRDiff** [43] explored a different conditioning strategy. It also conditioned the denoising network on the LR image, but with one major change in methodology. The model should not predict the HR image directly. The diffusion process should be applied to the residual, the difference between the HR image and a simple bicubic upsampling of the LR image. This change forced the model to mainly focus on the high-frequency details that are missing in the bicubic upsampling.

2.3.3 Efficiency and Zero-Shot Restoration

The major factor holding early diffusion back was the immense computing power needed to train them: transformer baselines such as SwinIR can be trained on a single high-end GPU within a few days while SR3 typically requires large GPU/TPU clusters and significantly longer training times [42] [5]. The **Latent Diffusion Models (LDMs)** [40] tried to tackle this issue by not directly running the diffusion process on high-resolution images. LDMs use autoencoders to compress the image into a small and efficient latent space. The normally computationally expensive part, the diffusion process, only needed to work in this efficient latent space. Large scale models like Stable Diffusion were only made possible by this key innovation, reducing training time and inference cost significantly. The development of massive, pre-trained diffusion models raised an important question: could their powerful learned prior of the natural world be adapted for other tasks without retraining? Two papers explored this possibility: **DDRM** [44] and **DDNM** [45]. They guided a pre-trained, unconditional model using mathematical properties of the degradation process (blur and downsampling). At each denoising step, they force the model to stay consistent with the LR input, while reaping the benefits of the model's pre-trained knowledge. This "zero-shot" approach demonstrated that it was possible to achieve strong performances without requiring additional fine-tuning or retraining.

2.3.4 Tackling Inference Speed

Another drawback of diffusion models is their slow inference time, which requires hundreds or thousands of denoising steps. To address this, **ResShift** [46] fundamentally redesigned the underlying Markov chain. Instead of starting from pure noise, it learns to transition directly

from the LR image to the HR image. Subsequently, a drastically shorter number of steps (e.g. 15) is needed, as the model essentially takes a shortcut, starting its generative journey from a point already rich with information rather than from pure noise. Pushing this idea to its limit, **SinSR** [47] employs a technique known as knowledge distillation. It trains a more efficient "student" model that could mimic a pre-trained "teacher" model. The student should reproduce the teacher's output in just a single inference step. With this powerful technique, they achieved ultra-efficient high-quality SR.

2.3.5 Hybrid Architectures and Leveraging Pre-trained Models

A particularly impactful, more recent trend is the adaptation of foundation models. **StableSR** [48] leverages the powerful prior of the large, pre-trained text-to-image model Stable Diffusion for super-resolution. It freezes the pre-trained model and trains only a lightweight time-aware encoder to inject the LR image's information at each timestep. This key innovation constrains the model to accurately represent the information in the LR image, thereby limiting hallucinations. In contrast, other research does not rely on pretrained models but focuses on a novel hybrid design, combining the power of two of the most relevant architectures: Transformers and Diffusion. **HI-Diff** [49], for instance, employs a diffusion model to generate a compact latent prior, which the downstream transformer utilizes to perform the primary restoration task. The goal is, once again, to inherit the best of both worlds: diffusion's generative advantages and the more trustworthy detail preservation of transformer models. A novel reimagined latent diffusion process was introduced with **Refusion** [50]. Instead of the hard to train VAE-GAN typically used in LDM, it uses a U-Net based encoder-decoder architecture. Skip connection allows the U-Net architecture to preserve fine-grained details that are often lost in compression steps. Most diffusion models are due to computational limitations not suitable for running on edge devices. **BI-DiffSR** [51] addressed this challenge with a technique commonly referred to as binarization. Binarization is an extreme form of model compression that had not previously been successfully adapted for diffusion-based super-resolution. This innovation enables high-quality SR inference on resource-constrained hardware. This overview covered the most noteworthy general-purpose diffusion models. Diffusion models tailored to remote sensing will be discussed in the next section.

2.4 Super-Resolution in Remote Sensing

Although the fundamental mechanics of super-resolution also apply in the remote sensing space, there are some extra challenges to overcome. Satellite data can exhibit quite different properties than normal ground imagery. While standard super-resolution datasets work with RGB images, satellites have sensors that can capture a broader spectrum than the visible to the human eye. Multispectral satellites commonly can capture band counts ranging from 3 to 15 bands, while hyperspectral satellites can be equipped with hundreds of different channels. There are several strategies for handling this complex data for super-resolution, including upscaling each band individually or building models that utilize correlation between multiple

bands for improved results. Commonly, satellite data is stored in a high dynamic range format, such as 16-bit integers, which requires more profound normalization strategies (as discussed in Section 3.1.3) than typical 8-bit images. This section reviews key research in the remote sensing domain, from the development of specialized datasets to specific model designs that work on Earth observation data. For a broad overview, recent surveys by Liu et al. [52] and Qi et al. [53] provide detailed descriptions of the field.

2.4.1 The Landscape of Datasets and Benchmarks

The development of effective SR models is dependent on the availability of high-quality training and evaluation data. A still common approach is using a simple degradation model, such as bicubic, to create LR images from satellite tiles. However, to achieve the best results on real-world data, it is advisable to use datasets that model a more complex relationship between their LR and HR pairs. A recent trend is to leverage different sensors with varying spatial resolutions from multiple satellites to create a cross-sensor dataset. Creating such a dataset is extremely difficult, as it is hard to perfectly align the low-and high-resolution image pairs. For instance, a valid pair requires that the images be captured in a similar time frame, from sensors with similar angles, and with matching spectral bands. Several projects have addressed this challenge with different approaches.

To provide real-world training pairs, the **SENVENµS dataset** [54] leverages the VENµS satellite to provide 5m ground truth, pairing them with 10m Sentinel-2 patches. They utilize the substantial overlap of spectral bands between those sensors and apply the same atmospheric correction processor to both, ensuring high radiometric consistency, meaning that both sensors capture similar brightness levels for the same surfaces. The capture time difference also never exceeds 30 minutes, and most of the time, it is even under 10 minutes.

For multi-image super-resolution (MISR), two distinct needs emerged: the need for large-scale training data and the requirement for realistic evaluation metrics. Covering 10,000 km, **WorldStrat**[55] provides a massive, globally diverse training dataset by pairing Sentinel and Spot 6/7 imagery. One of its main focuses is a wide variety of different land-use types. The high ground sampling distance (GSD) **Proba-V challenge** [56] provided a realistic benchmark for upscaling from 300m to 100m. The dataset was later refined by **Proba-V-ref** [57]. Providing a comprehensive protocol for cross-sensor evaluation, **MuS2** [58] combined Sentinel-2 and WorldView-2 Imagery to address the need for realistic MISR benchmarks.

A major limitation of cross-sensor datasets is the scarcity of high-quality, overlapping image pairs. A clever hybrid solution presented by the **SEN2NAIP dataset** [59] is first learning a realistic degradation model from a small set of real Sentinel-2 and high-resolution NAIP aerial imagery. Using this degradation model, it generates a massive, realistic synthetic training set. Utilizing Sentinel-2 paired with NAIP, SPOT, and VENµS, the **OpenSR-test** [60] framework provides a selection of curated cross-sensor tests. Together with their novel evaluation protocols focused on correctness, synthesis, and consistency, they introduced a comprehensive evaluation suite for remote sensing super-resolution.

In this thesis, we decided against cross-sensor datasets, focusing instead on single-sensor data for consistency. In particular, we generated our datasets pairs for our Sentinel-2 (Section

3.1.1) and our VENµS (Section 3.1.2) datasets by degrading them at runtime.

2.4.2 Evolution of Models for Remote Sensing Super-Resolution

Early deep learning approaches for remote sensing SR often focused on adjusting existing CNN architectures, while current methods employ transformer and diffusion architectures. One of the pioneering models for multi-frame super-resolution was **HighRes-net** [61], which introduced a recursive fusion architecture that can process an arbitrary number of low-resolution views. A pure transformer architecture specifically adapted for remote sensing SR is the **Top-k Token Selective Transformer (TTST)**[62]. With the goal of selecting only the most relevant tokens, it introduces an efficient attention mechanism that thereby ignores redundant information. While transformer architectures continue to evolve into highly effective models, the field has also seen a surge of research into diffusion-based models. These have shown great promise in generating realistic, high-frequency details, representing an alternative and rapidly developing approach to the problem. A key model in this area, and one that is central to this thesis, is **EDiffSR** [8]. It was designed specifically for efficient SR of remote sensing images. To reduce the high computational cost of typical diffusion models, it replaces the standard U-Net denoiser with its own lightweight alternative. It also proposes a dedicated module to extract a more informative prior from the low-resolution image, boosting performance. The specific architecture of EDiffSR and its adaptation for our work are detailed in Section 3.2.5. Satellite imagery is known to contain diverse yet recurring ground features. The **Heterogeneous Mixture of Experts model** [63] was proposed to leverage this property. It utilizes a set of specialized sub-networks with different configurations, organized into expert groups, to focus on different types of ground objects. A two-step router adaptively selects the most suitable experts for each pixel, leading to a more effective reconstruction. The trend of leveraging large foundation models has also reached remote sensing, with **DiffusionSat** [64]. Being able to condition it on metadata, such as geolocation and timestamps, makes it powerful for a variety of downstream applications, including SR. Fusing information from different temporal images is another novel multi-image SR approach shown by **SatDiffMoE** [65]. The condition mechanism also feeds the LDM with the relative time difference between the image capture times, to consider them. The primary focus of the **Trustworthy Super-Resolution model** [66] is the correctness and reliability of its results. They adjusted the conditioning process of the LDM framework to achieve better spectral consistency. Using the probabilistic nature of DDPMs they generated pixel-wise uncertainty maps for each SR image.

Despite all these advances, super-resolution is still in its early days. Many foundational breakthroughs have only happened in the last few years, and the research in the field is just getting started. While overall image quality has improved substantially, there are still many limitations that need to be solved. Especially, super-resolution in the remote sensing domain is still quite niche, with only a few contenders developing models that are tailored to satellite data. Seeing the massive potential of this domain, this thesis contributes to the growing field by providing a detailed comparison of relevant super-resolution models and by highlighting their impact on downstream applications.

3 Methodology

3.1 Datasets

A central goal of this thesis is to develop robust super-resolution models working on real-world satellite data. To this end, we employ two distinct satellite sensors, allowing us to evaluate model performance across different spatial resolutions, spectral bands, and sensor characteristics. The main two spatial upscaling scenarios we focus on are 20m-to-10m, using Sentinel-2 imagery, and 10m-to-5m using the high-resolution VENµS satellite.

Our approach is designed to be band-agnostic, focusing on single-channel super-resolution. The models we train need to excel at upscaling single-band data by treating them as grayscale imagery. If a band is in the visible spectrum or the Short-Wave Infrared (SWIR) should be irrelevant to the models, which is why both datasets are based on a variety of different bands. Creating our High-Resolution (HR) and Low-Resolution (LR) dataset pairs on different bands prevents the model from overfitting to characteristics of a single band.

The models are also required to handle a key characteristic of scientific satellite data: 16-bit data depth. Although, a standard 8-bit dynamic range can only differ between 256 intensity levels, 16-bit data provides an expanded range of 65.536 values. This extended range presents a set of unique challenges: Standard monitors can only display 8-bit imagery, making it impossible to perceive the full depth of the image. Additionally, the majority of computer vision tools were only developed to support RGB 8-bit images. In Section 3.1.3, we will go in-depth about the challenges of normalizing 16-bit data.

3.1.1 The Sentinel Dataset: Upscaling from 20m to 10m

The primary dataset for this thesis uses imagery from the Sentinel-2 mission [2] [67], a key component of the European Union's Copernicus Programme. The core of the mission is employing two identical satellites (Sentinel-2A, Sentinel-2B) that operate in the same sun-synchronous orbit and capture high-resolution optical imagery. In 2024, a third satellite (Sentinel-2C) joined the constellation and will replace the already 10-year-old Sentinel-2A in the foreseeable future. With its high revisit frequency of 5 days at the equator, its high spatial and spectral resolution, and its free and open data policy, Sentinel-2 is widely used in remote sensing research and various applications. [68]

The mission's main objective is systematic global land monitoring, with a focus on environmental applications, including forest, agriculture, and land cover change detection [69]. To achieve this, Sentinel-2 satellites are equipped with the Multispectral Instrument (MSI), a sensor that captures data across 13 bands, in a wide spectral range [70]. Three different spatial resolutions are employed to support the different bands:

- **10 meters:** The four bands at the highest resolution include the three standard visible bands (Blue, Green, Red) and a Near-Infrared (NIR) band.
- **20 meters:** Six bands are available at this resolution: Four narrow bands track the Red-Edge portion of the spectrum, commonly used for vegetation detection and monitoring. The remaining two bands work in the Short-Wave Infrared (SWIR) bands.
- **60 meters:** Three lower resolution bands are used for atmospheric correction and cloud screening.

For our super-resolution task, we only worked with the 10m and 20m bands, disregarding the 60m options.

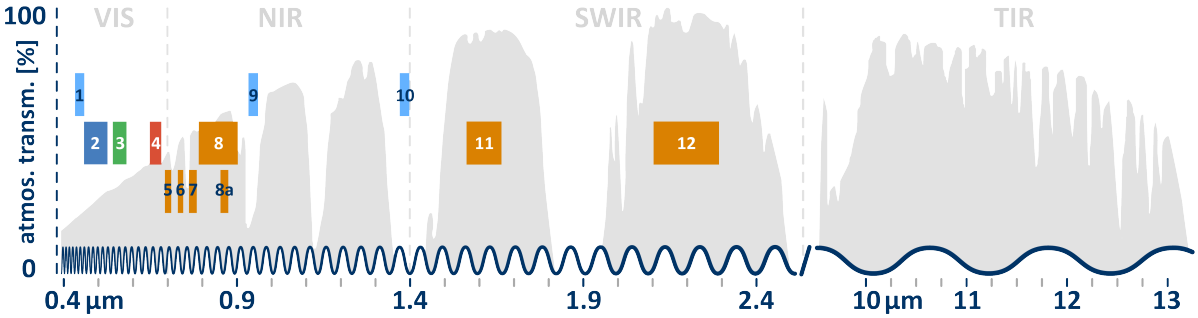

Figure 3.1: Overview over the different Sentinel 2 bands. Reproduced from [70].

Creating the Sentinel Dataset

To enable a stable training and a fair evaluation, we split our dataset into a dedicated training set and an extensive evaluation set. For training, we selected 10 Sentinel-2 tiles distributed over Europe, with the goal of capturing its diverse environments. To this end, we tried to represent a wide range of land cover types, including urban areas, forests, coasts, and agricultural areas. For correct evaluation of the models capabilities, our validation set was completely distinct from our training set, containing 10 tiles from around the world. The idea here was to capture the even bigger variety of different continents and track the model's real-world generalization capabilities over previously unseen scenarios. The geographic distribution of the selected tiles for training and validation set are illustrated in Figure 3.2.

The Sentinel tiles were selected using the Copernicus browser [71] and fetched through DLR's internal Sentinel-2 database. Only tiles with less than one percent cloud coverage were accepted, in order to minimize artifacts affecting the image quality. We selected the L2A versions of tiles since they already include atmospheric correction and remove unwanted influences such as haze, aerosols, and water vapor.

For our 2x super-resolution task, we generated High-Resolution (HR) and Low-Resolution (LR) image pairs from the four 10-meter bands of each selected tile. The original 10-meter resolution data served as the HR ground truth. The corresponding LR images were created

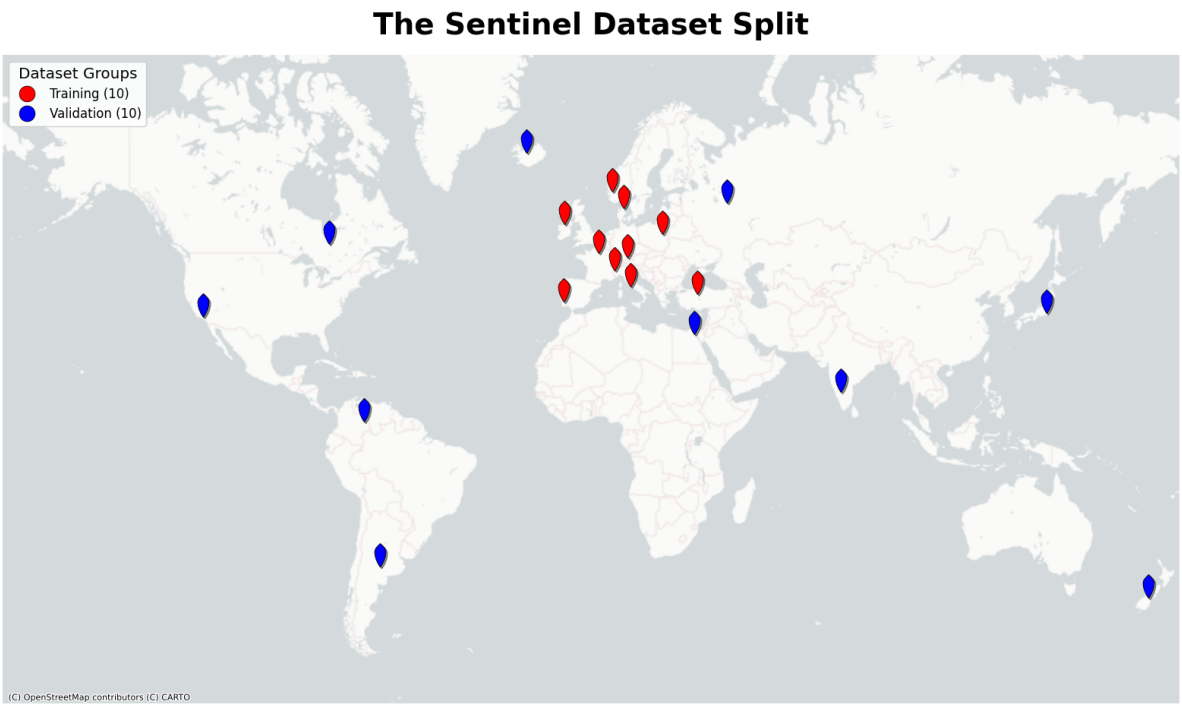

Figure 3.2: Overview over the locations of Sentinel tiles selected for the trainings and validation sets.

by 2x downscaling the HR images using Python Imaging Library's (PIL) bicubic interpolation [72]. Following the degradation, the full-tile images were divided into non-overlapping 256x256 pixel patches for the HR set and corresponding 128x128 pixel patches for the paired LR images. With a 10-meter Sentinel-2 tile having dimensions of 10980x10980 pixels, this patching strategy yields a total of 7,056 HR/LR pairs per tile (1,764 patches for each of the four bands). Consequently, both our training set and our validation set consist of 70,560 image pairs.

It is important to note that using a single, known degradation kernel like bicubic interpolation is a common but simplified approach. For future work, employing more complex and varied degradation models that better simulate real-world sensor effect, such as blur, noise, and compression artifacts could lead to models with even greater robustness [24]. However, due to the time constraints of this thesis, this was considered out of scope.

The raw 16-bit data from the Sentinel-2 tiles was normalized using a per-tile percentile clipping method. This approach was chosen to handle the sensor's wide dynamic range while mitigating the impact of extreme outliers. A more complete discussion of this strategy, its benefits, and its limitations is provided in the normalization technique Section 3.1.3.

3.1.2 The VENµS Dataset: Going from 10m to 5m

To complement the 20m-to-10m super-resolution task, and to further test model robustness and generalization, we added a second dataset to work with a spatial gap of 10m-to-5m. This dataset is based on imagery from the **Vegetation and Environment monitoring on a New Micro-Satellite (VENµS)** mission [73]. VENµS is a French-Israeli satellite launched in 2017 providing high-resolution observations over many sites worldwide with a very high revisit frequency of two days. The single sensor powering the VENµS mission, the VSSC (VENµS SuperSpectral Camera), captures 12 bands with a high resolution of 5m ground sampling distance. A key advantage of the VENµS satellite for this work is the close spectral correspondence of its bands with those of Sentinel-2 [54], an interesting property for the generalization of our models to both datasets. This spectral overlap is illustrated in Figure 3.3.

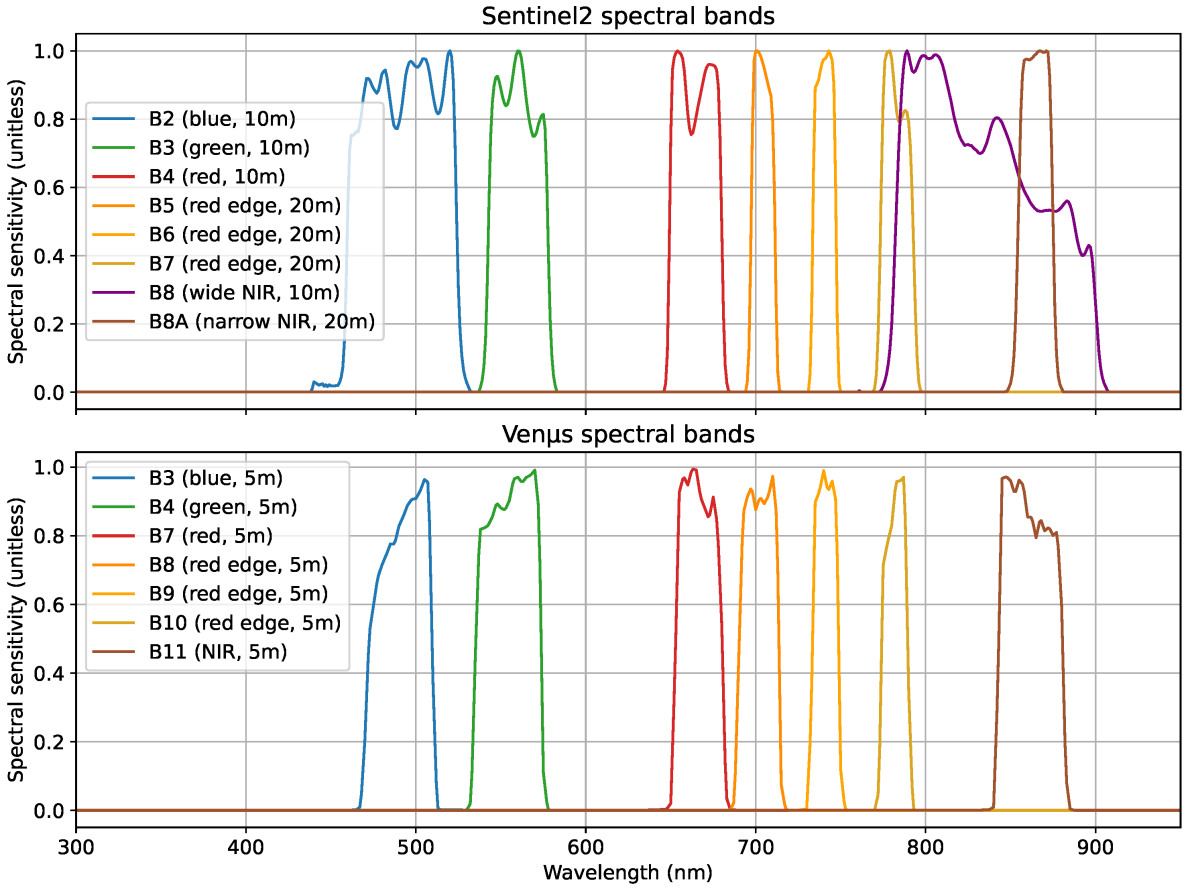

Figure 3.3: Overview of the spectral overlap of VENµS and Sentinel-2. Reproduced from [54].

Creating the VENµS Dataset

The data was sourced from the SEN2VENµS dataset, which provides co-registered Sentinel-2 and VENµS image pairs [54]. For our task, we utilized the 5-meter bands from the VENµS imagery to serve as our High-Resolution (HR) ground truth. The corresponding Low-Resolution (LR) images, at 10-meter GSD, were generated by downscaling the 5-meter HR VENµS data by a factor of two. To degrade the images, we used the same method as for the Sentinel-2 dataset: PIL bicubic [72], in order to ensure methodological consistency. The data was then partitioned into 256x256 pixel HR patches and 128x128 pixel LR patches.

Unlike the tile-based split of the Sentinel-2 data, the VENµS dataset was partitioned based on geographic location sites. The training set, comprising 849,458 HR/LR pairs, was created from a majority of the available sites. The remaining four sites were used to create a distinct validation set with 118,292 LQ/HR pairs. This site-based separation ensures an evaluation of model performance on unseen geographic locations. The distribution of these training and validation sites is shown in Figure 3.4. A representative subset of the validation pool was then sampled for use during the training epochs.

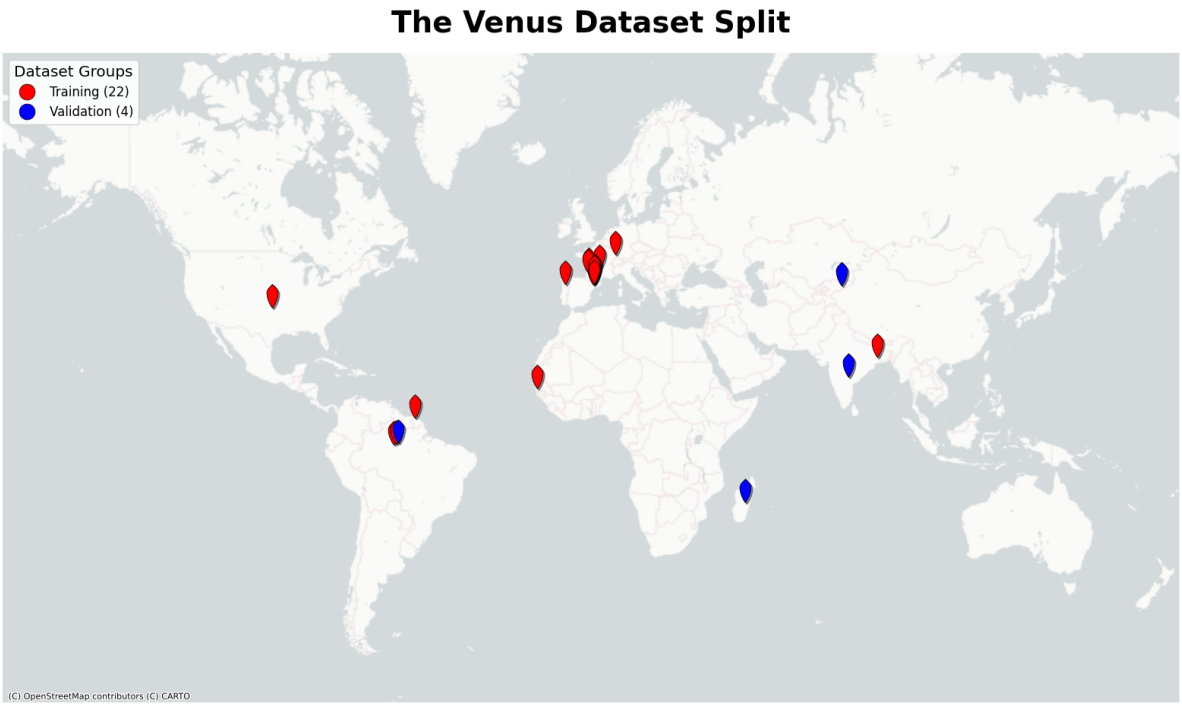

Figure 3.4: Overview over the locations of VENµS Sites selected for the trainings and validation sets. [54]

For the VENµS data, a per-patch min-max normalization was applied. An explanation for this approach, along with a detailed analysis of different normalization strategies, is presented in the following subsection.

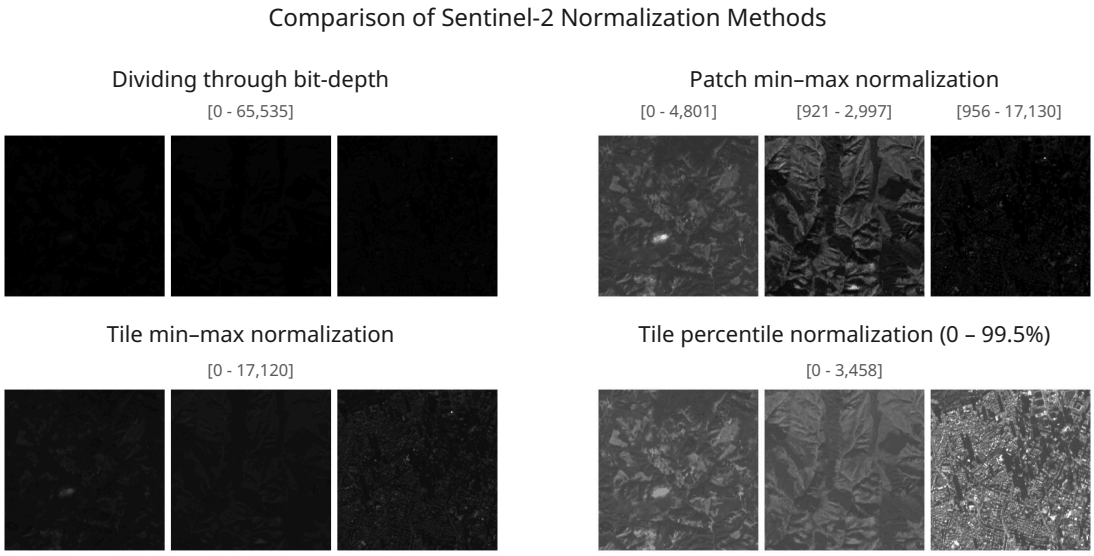

Figure 3.5: Comparison of Sentinel-2 normalization methods on three patches from the same tile, showing how different strategies affect image contrast and structural visibility. The ranges in brackets indicate the minimum and maximum pixel intensities used for each normalization.

3.1.3 The Challenges of Data Normalization for 16-bit Satellite Imagery

An essential preprocessing step for computer vision models is data normalization to a suitable format. Typically, the pixel values of an image get transformed into a standard range of [0, 1]. Figure 3.5 provides an overview of the linear normalization methods discussed in this section, using three different 256×256 patches from the same Sentinel-2 tile with varying pixel intensity ranges.

Bit-Depth Normalization by Division

While the trivial approach of dividing pixel intensities v_{in} by their dynamic range D (e.g. 255) works fine for 8-bit data, it breaks down for scientific 16-bit satellite data.

$$v_{out} = \frac{v_{in}}{D} \tag{3.1}$$

Sensors from satellites like Sentinel or VENµS capture only data with a 12-bit radiometric resolution, but store them in 16-bit numbers with a full range of [0, 65.535]. This leads to the actual data being stored in a small portion of the full range. Another challenge is that pixel intensities can appear in varying, non-uniform regions of the 12-bit range, making a fixed divisor approach ineffective. In Figure 3.5 we can clearly witness that the fixed divisor approach leads to completely dark images, as the Sentinel-2 data only lies in a small fraction of the full bit-depth range.

Min-Max Normalization on Full Satellite Tiles

Since we are working with full satellite tiles, the next naive approach would be to calculate the min (m_T) and max (M_T) pixel values and normalize all tile pixels (v_{in}) with these statistics, to achieve the output values (v_{out}):

$$v_{out} = \frac{v_{in} - m_T}{M_T - m_T} \tag{3.2}$$

While this technique is already a big improvement over the fixed divider strategy, it still struggles with a common characteristic of satellite data: outliers. Outliers can often happen due to super reflective surfaces, like snow, clouds, water, or man-made materials. This leads to extremely bright hotspots that can skew the min-max range drastically. Once again, the meaningful information of the image is compressed into a tiny portion of the resulting [0,1] range. This effect is clearly visible in Figure 3.5, where a tiny but extremely intensive highlight in the right patch skews the scaling, leading once again to dark images with no visible information in the other patches.

But what are the consequences of this narrow relevant range? It can hinder training progression and evaluation corruption. The most popular super-resolution loss function, named L1 loss, is based on the mean absolute error (MAE), which is directly affected. Meaningful pixel differences in the narrow range lead to small MAE values and are weighted less than irrelevant outliers with massive error gaps. But also other algorithms relying on MAE are affected: In preliminary tests, we observed unrealistically high PSNR values around 60 dB when employing this normalization strategy. Due to the extended range, the mean error in the relevant areas gets artificially small, leading to better but unrepresentative pixel-wise metrics.

Our Solution for Sentinel-2 Tiles: Percentile Clipping

The solution we used for our Sentinel-2 dataset was percentile-clipping, a common approach in the remote sensing community. The idea is calculating a low percentile (p_{low}) and a high percentile (p_{high}), for example, the 1st and 99th percentiles, and then normalizing the pixels values (v_{in}) using these boundaries:

$$v_{out} = \frac{v_{in} - p_{low}}{p_{high} - p_{low}} \tag{3.3}$$

Information outside of this range is clipped to the normalization range of [0, 1]:

$$v_{out} = \min\big(\max(v_{out}, 0),\, 1\big) \tag{3.4}$$

For our Sentinel-2 usecase, we opted for a more conservative range of 0th and 99.5th percentile, to retain as much valuable information as possible without skewing the range. This approach eliminates most of the harmful outliers and solves the previous mean absolute error issues. After applying percentile clipping, for the first time we can clearly see the full image in Figure 3.5. However, in the left patch the large highlight is clipped, as can be observed by comparing it to the patch min–max normalization result in the graphic.

Non-Linear Normalization Alternatives

It is worth noting that alternative, non-linear normalization methods exist, such as logarithmic scaling or histogram equalization. These techniques can minimize information loss compared to percentile clipping. However, they fundamentally alter the original data distribution, removing the originally linear relationship of pixels. This introduced transformation complexity could have unwanted effects on the model's learning abilities, which is why we went with the simple and effective technique of percentile clipping. Research by Kadunc et al. [74] has shown that linear clipping methods can outperform more complex non-linear transformations for downstream tasks on satellite data, reinforcing our choice.

Our Solution for VENµS Patches: Per-Patch Min-Max Normalization

As we didn't have access to full VENµS tiles, we had to opt for a different normalization strategy. We implemented a per-patch min–max normalization. For each HR/LR pair, the minimum (m_{HR}) and maximum (M_{HR}) pixel values were first calculated from the High-Resolution (HR) patch (256x256). These calculated statistics were then applied to normalize both the pixels (v_{in}) from the HR patch and its Low-Resolution (LR) counterpart (128x128):

$$v_{out} = \frac{v_{in} - m_{HR}}{M_{HR} - m_{HR}} \tag{3.5}$$

The main disadvantage of this approach is losing the global context of the absolute brightness values of a patch in comparison to its surrounding scene. Meaning that every patch contains data filling the entire [0, 1] range. Additionally, some rare patches might still be affected by outliers. The effect is clearly visible in 3.5, where the right patch is once again skewed by the extreme highlight, while the other image patches look decent. On the positive side, no information is lost in this normalization process.

The Perfect Normalization Approach?

Ultimately, we highlighted in this section that there is no single, perfect normalization method for 16-bit satellite data. Each strategy we highlighted here has its own strengths and limitations. We found that there is an inherent tradeoff between global radiometric context and local contrast handling.

We view the difference between the normalization strategies of our two datasets as an additional form of data augmentation. Forcing our models to train both on percentile clipped patches from Sentinel and per-patch min-max normalized VENµS patches helps our model to be more robust across different datasets. This also aligns with our overarching goal of developing versatile and robust satellite super-resolution methods, working on different spatial resolutions and with decent generalization across other satellite sensors.

3.2 Adapting the Super-Resolution Models

In this section, we look into the four super-resolution models selected for our study: SwinIR, MAT, PFT, and EDiffSR. While the first three rely on transformer architectures, EDiffSR is a diffusion-based generative model. To make these models work with our satellite data, we had to make substantial adaptations, as they were initially built for 8-bit natural RGB image datasets such as DIV2K [75].

With the goal of enabling a fair comparison, we established the same data loading, training, and validation pipeline for all models. This required different levels of modifications to the code base of each method.

In the following subsection, we first look at the core adaptations and data handling strategies shared by all models. We then discuss each method individually, highlighting both its unique functionality and any additional adaptations necessary for compatibility with our Sentinel-2 (Section 3.1.1) and VENµS datasets (Section 3.1.2).

3.2.1 Core Architectural Adaptations and Data Loading

Since SwinIR, MAT, PFT have implementations on a similar version of BasicSR [76], a popular super-resolution toolkit, most adaptations for these models were easily synced across codebases. While EDiffSR also contains parts of an older BasicSR version, it had substantial differences needing extra adjustments. BasicSR uses yml configuration files to specify trainings and validation settings. One goal of our adaptations was to make all models compatible with the dataset and validation section of these configuration files. This enabled us to run the same experiments and training procedures across all models.

To ensure that we created a unified data pipeline, we implemented our own custom satellite data loader, built for both our Sentinel and VENµS datasets and used for all trainings and experiments of this thesis. One of its main focuses was correctly handling the 16-bit single-band data of our datasets by loading and transforming the images into tensors. It could either work with simple patch directories, relevant for VENµS, or large satellite tiles by splitting them at runtime into GT patches. Sentinel tiles could also be fetched at runtime by tile-id. Working directly with full tiles allowed us to calculate the tile dependent statistics needed for normalization strategies like tile-based percentile clipping. Our dataloader also implemented the other relevant normalization techniques explained in Section 3.1.3. For training purposes, the GT patches were degraded at runtime using bicubic downsampling with PIL to generate paired LQ patches.

To handle the large image tiles and to process data directly in 16 bit, the dataloader was optimized to use Rasterio [77] together with Gdal [78] for image reading and transformations. For a faster data loading process, the loader utilized the extensive RAM capacities of Terrabyte [79] and imported the entire datasets into memory, before starting the training. The dataloader returned the required image tensors of GT and/or LQ, along with metadata necessary for correct denormalization.

Other major adaptations were setting up a 16-bit compatible validation framework, with runtime validation comparisons between GT, LQ, and SR. These were used to assess model

performances while training. We also adapted PSNR and SSIM for 16-bit, for more information about the metrics adjustments, checkout the Sections 3.3.1 and 3.3.2.

For normal inference, it was crucial to denormalize the patches correctly with the metadata provided by the dataloader. Another necessary functionality for our thesis was setting up full tile inference. The idea is splitting the full tiles into 256x256 patches in the dataloader and recombine them after upscaling all patches. We stuck to the default network structure of each model, except for adjusting the in and out channels from 3 (RGB) to 1 (Grayscale).

In the upcoming sections, we examine each super-resolution model used in this thesis and describe extra adaptations required.

3.2.2 SwinIR: An Influential Transformer Baseline

SwinIR [5] is one of the most influential transformers developed for image restoration, denoising, and super-resolution. Released in 2021, it established transformers as a competitive alternative to CNNs for super-resolution, sparking a new wave of research in the field. Although it is no longer state-of-the-art, it remains a popular baseline for newer models and continues to achieve decent results.

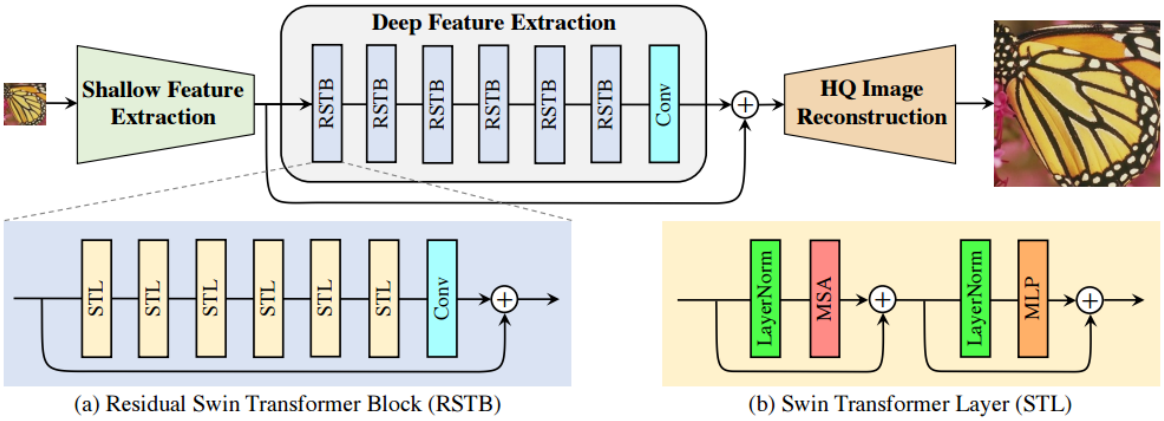

Figure 3.6: The architectural design of SwinIR. Reproduced from [5].

Architecture

Illustrated in Figure 3.6, SwinIR established the popular super-resolution architecture split into three modules: Shallow Feature Extraction, Deep Feature Extraction, and HQ Image Reconstruction.

The **Shallow Feature Extractor** is a single 3x3 convolution layer at the very front of the network, with the objective of extracting basic low-frequency features. It converts the input image into a number of feature maps, which are passed along to the next block and also get reapplied before reconstruction using a long skip.

The transformer core of SwinIR is the **Deep Feature Extraction module**, being the vital part for the model to reconstruct textures, high-frequency details, and to understand longer-range dependencies. Architecturally, it is a stack of Residual Swin Transformer Blocks (RSTBs), with a final 3x3 convolution layer for output refining. RSTB once again contain a sequence of Swin Transformer Layers (STL), these modules include the two blocks vital for any transformer: Multi-Head Self Attention (MSA) and Multi-Layer Perceptron (MLP). We explained MSA on a high level in Section 2.2.1

The real innovation of SwinIR lies in how attention is applied. Instead of calculating the quadratic scaling MSA over the entire image, it utilizes **Window-Based Multi-Head Self-Attention** to only work on 8x8 partitions. To maintain the global context, a clever mechanism was introduced called **Shifted Windows**. After every layer, the window partition is shifted by half a window size, allowing information to flow over multiple layers from one window to the next. This achieves a combination of global connectivity and efficiency. A variety of skip connections were both used in RSTBs and STL for improved data flow.

Combining features from both previous feature extractor modules into an upscaled final image is the task of the **HQ Image Reconstruction Module**. To achieve this, it empowers a sub-pixel convolution (pixel shuffle) to increase the spatial resolution and combine the feature maps.

Concepts like the three-part architectural split, the long skip connection, the implementation of the shallow feature extractor, and the image reconstruction module are still utilized by many state-of-the-art super resolution transformers.

Adaptations

We used the BasicSR version of SwinIR, which was included in the MAT repository [80]. All the major adaptations were based on the core changes detailed in Section 3.2.1. SwinIR was trained with the default options where possible. For more information on training, look into Section 3.4.

3.2.3 MAT: A Fast and Competitive Transformer

Released in November 2024, the Multi-Range Attention Transformer (MAT) [6] focuses on efficient image super-resolution, particularly with its exceptionally well-performing light model. But also their classical model with increased network size achieves state-of-the-art performances. For this thesis, we opted for the classical variant to ensure a fairer comparison with our other Transformer models.

Architecture

Following the previously introduced three-module structure, the **Shallow Feature Extractor** is once again powered by a single 3x3 convolution.

The primary building block of the **Deep Feature Extractor** is called **Residual Multi-Range Attention Group (RMAG)** and encapsulates two more important modules: LAB and MAB.

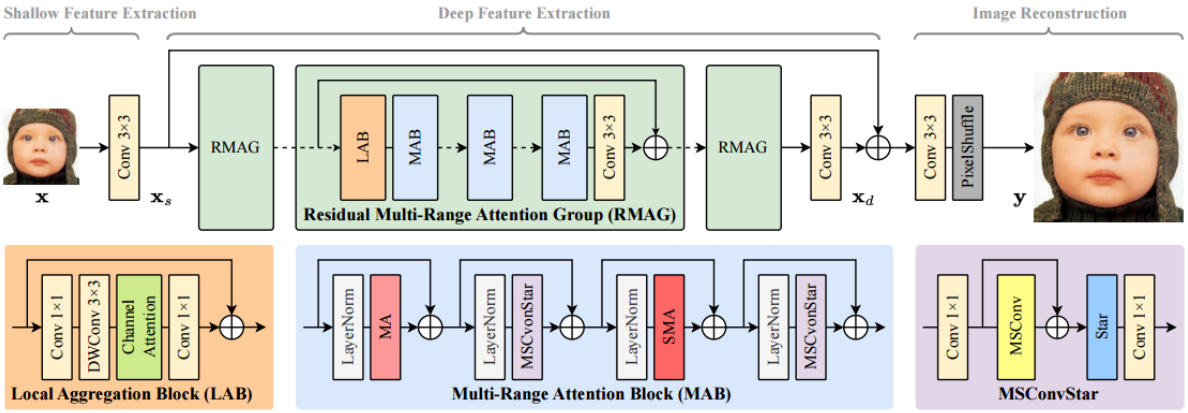

Figure 3.7: The architectural design of MAT. Reproduced from [6].

With the goal of capturing local fine patterns and textures, the **Local Aggregation Block (LAB)** combines the strength of traditional CNN-style convolution blocks with channel attention (similar to RCAN [22]). Channel attention weights previously extracted feature maps dynamically, to remove redundant or unnecessary information.

The heart of MAT is the name-giving **Multi-Range Attention Block (MAB)**, which has the goal to capture dependencies across different ranges, from local to global. To achieve this, it introduces two more attention algorithms: **Multi-Range Attention (MA)**, which calculates local attention with multiple window sizes (e.g. 7x7, 9x9, 11x11) at the same time, capturing details from fine-textures to broader patterns. Global patterns are captured by the **Sparse Multi-Range Attention (SMA)**, building on MA by using larger windows, but only capturing every, e.g 3rd pixel for attention calculations. This sampling using strides keeps performance efficient while still capturing context over large areas. The Multi-Range Attention Block also replaces the typical MLP with its own implementation called **MsConvStar**, which uses differently sized convolutions to account for the captured patterns with varying sizes and learns from them. Additionally, nearly every introduced block utilizes residual connections to enhance data flow.

The **Image Reconstruction Stage** is implemented in a similar fashion to SwinIR, by empowering a pixel shuffle to fuse and upscale the outputs of the deep feature and shallow feature extraction modules.

Adaptations

We already mentioned the MAT [6] repository, which is based on BasicSR [76]. Similarly to SwinIR, all major adaptations have already been discussed in Section 3.2.1. We trained the classical MAT version with the default network parameters and settings. For more information about training parameters, look into Section 3.4.

3.2.4 PFT: A State-of-the-Art Transformer

Published in March 2025, the Progressive Focused Transformer (PFT) [7] represents a recent state-of-the-art model with strong benchmark results.

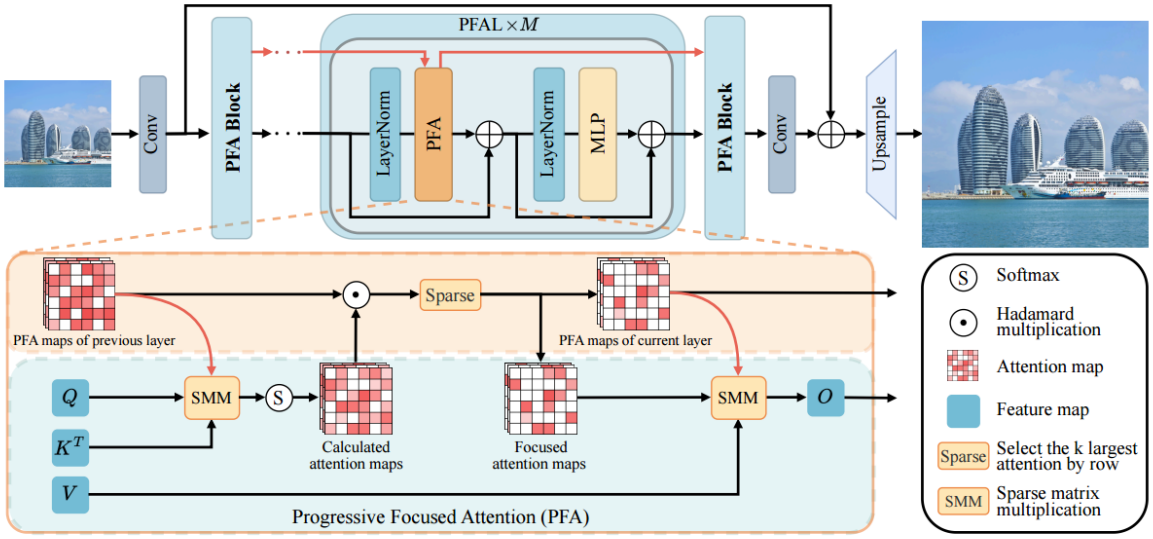

Figure 3.8: The architectural design of PFT. Reproduced from [7].

Architecture

It once again implements the previously introduced three-module structure, with a single 3x3 convolution acting as **Shallow Feature Extractor**, and a pixel shuffle upsampling the image in the **Image Reconstruction Stage**.

The **Deep Feature Extractor** leverages a simple structure nearly identical to SwinIR, containing a sequence of **Progressive Focused Attention Blocks (PFA Blocks)** that encapsulate several **Progressive Focused Attention Layers (PFAL)**. The key innovation is the **Progressive Focused Attention (PFA)**, which replaces the standard MSA block. The problem it aims to address is limiting computation on irrelevant tokens by introducing Sparse Matrix Multiplication (SMM), which allows the model to control which tokens to calculate.

How does the model decide which tokens are relevant enough to be computed? It links attention blocks between layers and only calculates tokens that were important in the preceding layer. At the first layer, attention is calculated on all pixels in a larger window (32x32), which is computationally heavy due to the quadratic nature of attention. The resulting attention map is forwarded to the next layer, where it serves as a guide for determining which tokens to recalculate. With a focus ratio of $\alpha = 0.5$ (common value), only the highest half of tokens in the map get refined. The newly calculated attention map is multiplied by the previous map using the Hadamard product to generate the attention map for the following

layer. Repeating this process across layers progressively sharpens the focus, as irrelevant tokens get removed early, while meaningful dependencies are reinforced. This enables PFT to leverage larger window sizes while keeping the computational cost low. PFT also borrows the shifted window concept from SwinIR, implementing it for every second layer, to allow communication between windows.

Adaptations

Even though the PFT repository [81] is built on BasicSR [76], it uses a slightly different version than our other models. Only minor extra adjustments were made to ensure compatibility with our data pipeline. All major changes were already discussed in Section 3.2.1. We trained the PFT model with the default network parameters, where possible. A more detailed overview of training settings can be found in Section 3.4.

3.2.5 EDiffSR: An Efficient Diffusion Model Designed for Remote Sensing

The final model included in our comparison is a diffusion-based approach. We were particularly interested in evaluating how such a model would perform relative to our transformer models. To this end, we searched for a recent and efficient model so that training times would be feasible and more comparable to our Transformer alternatives. Our choice fell onto EDiffSR [8], introduced in October 2023, it was particularly designed for remote sensing, with exactly the right balance of efficiency and performance for our comparison.

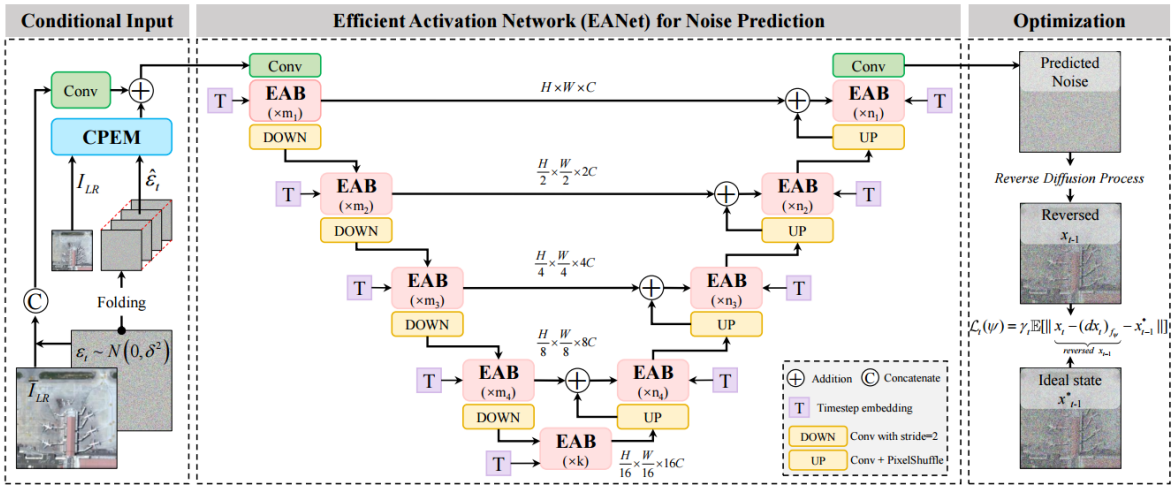

Figure 3.9: The architectural design of EDiffSR. Reproduced from [8].

Architecture

EDiffSR is a Denoising Diffusion Probabilistic Model (DDPM) [4], meaning it improves resolution through an iterative denoising process. We quickly summarized this process in

Section 2.3.1. The three main components of EDiffSR's architecture are the in Figure 3.9 illustrated Conditional Prior Enhancement module (CPEM), the Efficient Activation Network (EANet), and the Optimization/Reverse Diffusion Process.

Most famous super-resolution diffusion models (such as SR3 [42]) condition the denoiser only on a bicubic-upsampled LR image, which provides blurry and low-detail guidance. EDiffSR addresses this limitation with its **Conditional Prior Enhancement Module (CPEM)**, which fuses the original LR image, its bicubic upsampling, and additional noise through convolutions and channel attention. The resulting feature maps preserve structural cues and textures, providing the diffusion process with higher-level information than just a blurry image.

Commonly, a heavy U-Net plays the role of the denoising backend for many diffusion SR models (such as SR3 [42]), but EDiffSR replaces it with a much lighter, faster network, which it names the **Efficient Activation Network (EANet)**. Although the EANet inherits many properties from its big U-Net brother, such as the encoder-decoder architecture with skip connections, it introduces its lightweight **Efficient Activation Block (EAB)** as an alternative to the typical heavy residual/attention blocks. EAB empowers Multi-Scale Depthwise Convolutions, parallel convolution with different kernel sizes (e.g. 3x3, 5x5, 7x7), to capture fine details and broad regional context. Additionally, it uses Simple Channel Attention (SCA), a lightweight version of RCAN [22] attention, to weight the relevance of resulting feature maps. A simple gating operation ensures that only meaningful signals are preserved before the output data gets fused using a pointwise convolution.

In the **Optimization and Reverse Diffusion Process** stage, EDiffSR introduces several key improvements over prior diffusion-based SR models. It employs a mean-reverting SDE to stabilize the reverse process and is particularly well-suited for super-resolution tasks. The model is trained not only with a standard noise-prediction loss but also with a maximum likelihood objective, improving stability and fidelity. Additionally, conditional information from CPEM is injected throughout the denoising process, forcing the models to stay consistent with its input. EDiffSR also commonly employs a low amount of 100 diffusion steps.

These three main components improve the efficiency and performance of EDiffSR significantly: In the EDiffSR paper [8], the authors report that their Efficient Activation Network (EANet) contains only 26M parameters compared to 137M in IRSDE's [82] U-Net, and achieves up to 7 times faster inference than SR3 [42]. Notably, EDiffSR still outperforms both SR3 and IRSDE in perceptual quality, achieving superior FID scores across multiple benchmark datasets. The parameter counts and inference speeds of our adapted models are displayed in Table 4.1, and deviate from their original RGB parent.

Adaptations

While EDiffSR [8][83] utilizes some components of the BasicSR framework [76], we had to make significant adaptations across the entire codebase to provide the same features and functionality as our other models. Some of the most significant adjustments were rewriting the training script to be compatible with multiple datasets, connecting our custom metrics and tracking them correctly to TensorBoard, reintroducing our visualizations, and adding extra

settings for inferring full satellite tiles. We also had to make adjustments to our dataloader and to the way the YAML options files were loaded. For a detailed overview of the settings and learning rates we used for training EDiffSR, look into Section 3.4 and Section 3.4.1.

3.3 Evaluation Methods on Super-Resolution

Super-resolution models aim to generate visually plausible high-resolution images from low-resolution inputs. But how can we reliably assess the quality of these upscaled images? Visual inspection is highly subjective and time-consuming. We need quantitative metrics that can efficiently evaluate large numbers of upscaled images and allow objective comparison between different SR techniques. Evaluating image quality with a single number is impossible, as different models emphasize different strengths. One model might prioritize staying true to the original low-resolution input, while another competitor shifts the focus towards generating perceptually realistic outputs. This is why multiple metrics are needed, each with distinct purposes, strengths, and limitations. Improving one metric often comes at the cost of another. **Pixel-wise metrics** such as PSNR [9] or SSIM [10] require a ground truth for each upscaled image. By comparing these pairs, we can measure the reconstruction fidelity of a model. **Perceptual metrics** have the advantage of not needing a ground truth, and often try to assess the perceptual quality or naturalness of an image. To achieve this, they rely on reference images, precalculated statistics, or pretrained models. These metrics are particularly important for evaluating real-world super-resolution tasks, where a ground truth reference image is typically unavailable: Otherwise, upscaling would have no purpose. In this thesis we used FID [11] and NIQE [12] as perceptual metrics.

A common theme is to combine pixel-wise and perceptual metrics to obtain a comprehensive evaluation of super-resolution models. In the following subsections, we describe the metrics relevant to this thesis and discuss how we adapted them for our remote sensing use case.

3.3.1 Peak Signal-to-Noise Ratio (PSNR)

The most widespread method for evaluating imagery in the field of super-resolution is the PSNR [9]. It directly compares each pixel value of a super-resolved image (SR) with its ground truth (GT) using the mean square error (MSE).

$$\text{MSE} = \frac{1}{WH} \sum_{i=1}^{W} \sum_{j=1}^{H} (I_{GT}(i,j) - I_{SR}(i,j))^2 \tag{3.6}$$

$I(i,j)$ returns the pixel intensity of the respective image at position i and j. In this thesis, W (image width) and H (image height) are always 256, as we work with 256x256 GT images. The PSNR is computed using the MSE and D, which stands for the maximum pixel intensity. For 16 bit images D would be 65,535.

$$\text{PSNR} = 10 \cdot \log_{10} \left(\frac{D^2}{\text{MSE}} \right) \tag{3.7}$$

Higher PSNR values correspond to super-resolved images that are closer to the ground truth. PSNR is computationally cheap, making it attractive for validation during training. Some limitations of PSNR are that it is very sensitive to small misalignments, which can lead it to favor overly smooth or blurry images.

For our special data, we **adapted** PSNR to work directly on tensors ranging from 0 to 1, ensuring that it works as intended on 16-bit single-band data.

3.3.2 Structural Similarity Index Measure (SSIM)

While PSNR provides a straightforward measure of pixel-wise differences, it often fails to capture perceptual quality. To address this, PSNR is commonly paired with SSIM [10], a metric that evaluates structural similarity between images. The idea is that visual perception is more sensitive to structural changes rather than absolute pixel differences. The SSIM is calculated locally, on small windows and then gets averaged across the whole image. These windows (typical size 11x11) are introduced to capture local structural differences rather than global averages. SSIM compares three different components of image patch x with the GT patch y: luminance $l(x,y)$, contrast $c(x,y)$, and structure $s(x,y)$. These three comparisons are weighted with α, β, γ and multiplied together.

$$\text{SSIM}(x,y) = [l(x,y)]^{\alpha} \cdot [c(x,y)]^{\beta} \cdot [s(x,y)]^{\gamma} \tag{3.8}$$

Commonly, all parameters are weighted equally $\alpha = \beta = \gamma = 1$. $C1 = (k_1 L)^2, C2 = (k_2 L)^2, C3 = C_2/2$ with constants k_1, k_2 and the dynamic pixel intensity range L, are small variables that are used to prevent division by zero. The **luminance** component compares the average pixel intensities of both patches: μ_x, μ_y

$$l(x,y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \tag{3.9}$$

Using the standard deviation of the pixel intensities: σ_x, σ_y SSIM checks if the **contrast** is preserved.

$$c(x,y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \tag{3.10}$$

The **structural** component utilizes the covariance of the two patches σ_{xy} to measure whether the patterns and edges align between the two images.

$$s(x,y) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3} \tag{3.11}$$

SSIM values range from 0 to 1, with 1 being a perfect match. Although it aligns better with human perception, it can still be fooled by unnatural artifacts and is more computationally expensive than PSNR.

Our **adaptation** worked once again directly on normalized tensors with pixel range $L = 1.0$, to ensure compatibility with 16-bit. The SSIM constants were set to standard values: $k_1 = 0.01, k_2 = 0.03$, and also a window size of 11x11 was implemented.

3.3.3 Fréchet Inception Distance (FID)

Pixel-wise metrics like PSNR and SSIM focus on reconstruction fidelity and are reliant on a ground truth to function. In contrast, FID [11] evaluates images without requiring a paired ground truth. It does this by comparing the statistical distribution of features extracted from generated and real images, focusing on perceptual realism rather than exact pixel accuracy.

In order to obtain these features, FID utilizes a pretrained feature extractor neural network, commonly Inception-V3 [84]. After passing through the network, the output of a high-level layer is extracted, resulting in a **feature vector** representing the semantic and structural content of the image.

For a set of N real images, we compute the mean μ_r and the covariance Σ_r of their feature vectors.

$$\mu_r = \frac{1}{N}\sum_{i=1}^{N} f(x_i), \tag{3.12}$$

$$\Sigma_r = \frac{1}{N-1}\sum_{i=1}^{N}(f(x_i)-\mu_r)(f(x_i)-\mu_r)^\top \tag{3.13}$$

For the generated image set μ_g, Σ_g are calculated in the same manner. The Fréchet Inception Distance (FID) is then computed as follows:

$$\text{FID} = \|\mu_r - \mu_g\|_2^2 + \text{Tr}\Big(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{1/2}\Big) \tag{3.14}$$

Lower FID values indicate that the generated images are closer to the distribution of real images, and thus appear more perceptually realistic. While no paired ground truth is required, a set of realistic images, is still needed to use FID. Additionally, for the best results it is advised to use a symmetric number of images for the realistic and generated set.

Adaptations

FID was developed for RGB 8-bit ground images, which have fundamentally different properties to our 16-bit single band satellite images. So we had to implement some major adaptations to BasicSR [76] FID implementation. We decided against retraining our own satellite feature extractor and stuck to the petrained InceptionV3 model [84], as retraining is beyond the scope of this thesis and is not commonly done in related works. To transform our 16 bit single band data into the correct format that InceptionV3 expects, we've duplicated our grayscale image to 3 channels, normalized the 16 bit data into the expected range of $[-1, 1]$. Additionally we activated InceptionsV3 resizing feature, which bicubic upscales our 256x256 patches to 299x299, the dimensions the model expects. We worked with floating-point precision where possible to preserve the 16-bit image information. To run FID on large amount of images, we build our own data loading, and execution pipeline.

Despite the adaptations we had to make, FID remains a valuable metric for comparing the perceptual quality of different models.

3.3.4 Naturalness Image Quality Evaluator (NIQE)

Although FID compares generated images to a set of reference images, NIQE [12] is a no-reference image quality metric that does not require real images or a paired ground truth for evaluation. Its goal is to measure the naturalness of an image by comparing its statistical properties to those of a model trained on "pristine" images: images that exhibit natural statistical properties.

The main steps are summarized below. First, the image is normalized using the Mean Subtracted Contrast Normalized (MSCN) transform:

$$\hat{I}(i,j) = \frac{I(i,j) - \mu(i,j)}{\sigma(i,j) + C}, \tag{3.15}$$

where $I(i,j)$ denotes the pixel intensity at location (i,j), $\mu(i,j)$ and $\sigma(i,j)$ are the local mean and standard deviation computed over a local window, and C is a small constant to avoid division by zero. In simple terms, this step removes the local brightness and contrast variations in the image as a preprocessing step, so that the following analysis does not get distracted by global lighting differences.

The distribution of the MSCN coefficients is modeled using the Asymmetric Generalized Gaussian Distribution (AGGD). The key parameters are estimated as follows:

$$\gamma = \frac{\sigma_L}{\sigma_R}, \quad r = \frac{\left(\frac{1}{N}\sum_{i=1}^{N} |\hat{I}_i|\right)^2}{\frac{1}{N}\sum_{i=1}^{N} \hat{I}_i^2}, \tag{3.16}$$

σ_L and σ_R are the standard deviations of the negative and positive coefficients and N is the number of coefficients in the patch. These parameters summarize the local statistical behavior of the image: γ captures the asymmetry between bright and dark regions, and r helps to describe the "peakedness" or spread of the distribution. These parameters provide a compact representation of the statistical structure of image patches. For each image, we collect these parameters from multiple overlapping patches and concatenate them into a single feature vector.

Finally, the NIQE score is computed as the Mahalanobis distance between the feature vector of the test image and the reference pristine model:

$$\text{NIQE} = \sqrt{(\mu_s - \mu_r)^T \left(\frac{\Sigma_s + \Sigma_r}{2}\right)^{-1} (\mu_s - \mu_r)}, \tag{3.17}$$

where (μ_s, Σ_s) are the mean and covariance of the test image features, and (μ_r, Σ_r) correspond to the values of the pre-trained reference model.

This step measures the difference between a test image and the pristine images that were used to train the model. Smaller NIQE values indicate that the image's local statistics are closer to those of natural images, so should have a higher perceived quality.

Adaptation

To make NIQE work for our domain, we build our own Python implementation using a repository from nuniniyujin [85] (fixed version of guptaprafal [86]) and the BasicSR version [76]. We implemented two pipelines, one for evaluating images with a pretrained model, and a pipeline to train our own custom satellite optimized models. In experiment Section 4.2 we will compare the pretrained NIQE models with our custom trained versions. A primary focus was working directly on floats with the goal of reaping the benefits of the full 16-bit quality range. In order to stay compatible with other pretrained models we stuck to a float value range from 0.0 to 255.0. Input images, for both training and evaluation, were divided into non-overlapping patches of size 96, a requirement by many NIQE implementations. The evaluation pipeline accepted the same dataloading process as FID, and aggregated a simple average NIQE score over the model, using a trained model. The training pipeline generates a Matlab file, containing the trained models feature vectors, as expected from other NIQE implementations.

3.4 Training Details

All experiments and training were computed on the Terrabyte servers [79], a service built by LRZ [87] in collaboration with the DLR [88]. It provides computing nodes equipped with NVIDIA A100-SXM4-80GB GPUs, enabling us to train multiple models in parallel. Additionally, we were able to leverage the large RAM Capacity of the cluster by loading our entire datasets into memory, significantly improving data processing speeds. To support the models' codebases, we used Mamba environments [89]. The PFT model was developed with Python 3.9 and PyTorch 2.5 [90], while the other models were compatible with Python 3.10 and PyTorch 2.2.2.

As one of our goals was to evaluate the models on different spatial resolutions, we trained them on three distinct dataset mixes. The first set of models focused exclusively on the 20m $\rightarrow$ 10m upscaling task using the pure Sentinel dataset (discussed in 3.1.1). In contrast, the models trained on VENµS (Section 3.1.2) optimized for the 10m $\rightarrow$ 5m domain. The final training setup combines these two approaches by using the previously trained Sentinel model and fine-tuning them on VENµS.

While training, we equipped the models with generous validation and test packs. A Sentinel subset of 5,000 images was used in all training configurations to evaluate performance on metrics like PSNR and SSIM. For models trained or finetuned on VENµS, a validation pack of the same size was added for 10m to 5m performance tracking. Additive small test sets for native sentinel upscaling were introduced for visual assessment of model performance. Due to the slower inference of EDiffSR, we restricted runtime validation to 500 images per pack.

We used **TensorBoard** to plot PSNR and SSIM over the course of training, helping us identify when the models had clearly converged. After training, the best-performing checkpoints (based on PSNR and SSIM) were selected for further experiments and tests. Figure 3.1 shows an overview of the trainings iterations employed for each model.

Dataset	Model	Warmup Iterations	Total Iteration	Experiment Iterations
Sentinel	SwinIR	0	500k	240k
	MAT	0	500k	380k
	PFT	20k	500k	250k
	EDiffSR	0	530k	370k
VENµS	SwinIR	0	1M	890k
	MAT	0	1M	980k
	PFT	20k	1M	990k
	EDiffSR	0	390k	190k
Finetuned on VENµS	SwinIR	0	920k	910k
	MAT	0	910k	900k
	PFT	10k	960k	870k
	EDiffSR	0	420k	60k

Table 3.1: Overview of Warmup Iterations, the Total Iterations the model was trained on, and the selected model iteration used for further experiments.

To ensure a fair comparison between different models, we tried to use their default settings where possible. Our datasets used 256×256 images for ground truth, and due to technical instabilities observed with larger batch sizes for PFT, we opted for a conservative batch size of 2. We maintained the same batch size across all models to keep the trainings comparable. As our datasets were both already quite large, there was no need for data augmentation. We stuck to the default optimizer settings of each model and also used the default learning rates. All models were trained on L1 loss. For an overview of Optimizer and Learning Rate settings, see Figure 3.2.

Dataset	Model	Optimizer ($\beta_1 = 0.9, \beta_2 = 0.99$)	Learning-Rate	LR Scheduler
Sentinel	SwinIR	Adam	2e-4	MultiStepLR
	MAT	AdamW	2e-4	MultiStepLR
	PFT	AdamW	2e-4	MultiStepLR
	EDiffSR	AdamW	4e-5	TrueCosineAnnealingLR
VENµS	SwinIR	Adam	2e-4	MultiStepLR
	MAT	AdamW	2e-4	MultiStepLR
	PFT	AdamW	2e-4	MultiStepLR
	EDiffSR	AdamW	4e-5	TrueCosineAnnealingLR
Finetuned on VENµS	SwinIR	Adam	1e-4	MultiStepLR
	MAT	AdamW	1e-4	MultiStepLR
	PFT	AdamW	1e-4	MultiStepLR
	EDiffSR	AdamW	2e-5	TrueCosineAnnealingLR

Table 3.2: Summary of optimizer, learning rate, and learning rate scheduler for all models.

3.4.1 EDiffSR Learning Rate Ablation Study

During training, EDiffSR (Section 3.2.5) exhibited signs of instability, with fluctuations in validation metrics (e.g. PSNR/SSIM) between training steps. We investigated further by training our models with different learning rates.

For the **Sentinel training** phase, the default learning rate of 4e-5 resulted in good results but instabilities while training, prompting further experimentation. We tested a range of values, including 1e-4, 1e-5, 5e-6, 2e-6, and 1e-6. Although we got the most stable convergence with 1e-6, the lower learning rates produced significantly worse results. Additionally, we retrained 4e-5 once again to test if our initial training was a fluke, but once again, it outperformed other learning rates significantly.

Similar to Sentinel, the default learning rate for **VENµS training** was 4e-5. As we already got a better understanding of sensible learning rates, we tried out 1e-6 and 9e-7, and we once again ran into the issue of low learning rates underperforming. For **fine-tuning on VENµS**, it is common to use lower learning rates to allow the model to adjust more gently to the new data. Our starting learning rate of 2e-5 was quickly surpassed by lower rates in term of stability, such as 1e-7, 5e-7, 5e-8, and 1e-8, while higher values, like 1e-6 or 5e-6, led to unstable training. But the trend stayed consistent, while the lower learning rates led to a more stable training, they also peaked way lower on validation metrics like (PSNR/SSIM) and perceptual metrics (FID/NIQE). Additionally, their inference images looked noisier. We believe the training instabilities are caused by the low batch size, which amplifies the noisy fluctuations when using high learning rates.

We decided to stick with the default learning rates due to their superior results, and to keep the fairness towards the other models, where we also maintained the default values.

4 Experiments and Results

In this chapter, we thoroughly evaluate the performance of our models and highlight their strengths and weaknesses across different domains. We begin with an analysis of model properties such as inference speed and reference-based metrics (PSNR, SSIM), complemented by visual comparisons against ground truth images. In the second Section 4.2 we'll go deeper on perceptual quality, tracking our models performances across different spatial resolution gaps on NIQE and FID. Here we look into the native upscaling capabilities of our models without relying on any ground truth. In the last Section 4.3, we demonstrate the value and real-world practicality of our models on a field boundary detection task, and compare their performance against simple bicubic upsampling.

4.1 Quantitative Evaluation

This section presents the quantitative evaluation of our models in three different parts: inference speed, reference-based validation metrics (PSNR and SSIM) across both Sentinel and VENμS datasets, and visual comparisons of the twelve trained models against their corresponding ground truth images.

4.1.1 Model Inference Speed Evaluation

Especially in the remote sensing space, real-world super-resolution applications often require processing massive amounts of data, making fast inference speeds and computational efficiency essential for the usability of a model. This creates an inherent trade-off between a model's computational demands and the quality of its results. To evaluate inference speeds for our four models, we set up a simple experiment using 10,000 images from our Sentinel validation pack (Section 3.1.1). Each model ran inference on the subset under identical conditions, using a batch size of one, on a Terrabyte [79] compute node equipped with an NVIDIA A100-SXM4-80GB GPU. Our goal was to only measure the time of the forward pass, excluding the data-loading and processing pipeline. Using CUDA Events with synchronization, we ensured accurate tracking of GPU latency. To remove GPU startup bias, the first 10 iterations were discarded as a warm-up, leaving 9,900 256x256 images for the metric calculations. The metrics calculated for each model are the simple mean latency, the median latency, and the 95th percentile latency. The overview of the model's parameters and latencies is visible in Figure 4.1.

Across all models, the differences between mean, median, and 95th-percentile latencies were relatively small, indicating stable computational performance. SwinIR and PFT showed almost no relevant variation (<1%), while MAT and EDiffSR exhibited small ~5% differences between

Model	Architecture	Parameter Count	Mean Latency (ms/image)	Median Latency (ms/image)	95th-pct Latency (ms/image)
SwinIR	Transformer	11,748,093	54.120	54.098	54.273
MAT	Transformer	9,592,005	87.384	86.936	91.650
PFT	Transformer	19,618,053	313.931	313.392	314.915
EDiffSR	Diffusion	20,400,261	2,163.721	2,137.074	2,256.431

Table 4.1: Comparison of models by parameter count and inference speed measured in ms per inferred image.

their median and 95th percentile values. Although EDiffSR was designed as an efficient DDPM, it still cannot compete with the speed of transformer models, being up to 40 times slower than our fastest model: SwinIR. Despite having a comparable number of parameters to PFT, EDiffSR reliance on 100 denoising steps explains the speed gap. Although inference speed is often assumed to correlate with model size, our experiments show SwinIR running faster than MAT despite its larger parameter count. The reason lies in architectural design: SwinIR's shifted-window attention is lightweight, while MAT relies on multi-range and sparse attention modules that increase computational cost. Our largest model, PFT, contains roughly twice the parameters of MAT and was about ~3.5x slower due to its progressive focused attention mechanism which initialy computes attention on larger 32x32 windows. Overall, the relative runtime differences among all models were consistent with expectations.

4.1.2 Quantitive Metrics: PSNR/SSIM

To evaluate reconstruction fidelity, we compared our twelve trained models using the metrics PSNR [9] and SSIM [10]. To track model performance on the 20 m $\rightarrow$ 10 m and 10 m $\rightarrow$ 5 m super-resolution tasks, we employed two validation sets: Sentinel (Section 3.1.1) and VENµS (Section 3.1.2). 10,000 images were used for both validation packs to calculate our 16-bit compatible PSNR and SSIM variants, discussed in Section 3.3.1 and Section 3.3.2. The twelve models consist of the four base architectures, each trained in three ways: on Sentinel-only, on VENµS-only, and on Sentinel with subsequent fine-tuning on VENµS. For more information about these trainings we recommend exploring Section 3.4. Table 4.2 displays an overview of the experiment results.

Results

Before diving into the specifics of the results, it is important to note that the absolute metric ranges for Sentinel and VENµS differ considerably. Some factors contributing to this metric gap are satellite-specific sensor properties and different normalization strategies. Benchmark performance can only be evaluated relative to other models on the same dataset and never across datasets.

Looking at the results of the experiment, a few trends emerge: As expected, models trained for 20m $\rightarrow$ 10m or 10m $\rightarrow$ 5m excel on their respective validation sets. For 10m $\rightarrow$ 5m,

Model Configuration		Sentinel Validation		VENµS Validation	
Dataset (GSD)	**Model**	**PSNR ↑**	**SSIM ↑**	**PSNR ↑**	**SSIM ↑**
Sentinel 2 (20m → 10m)	SwinIR	37.9632	0.9415	40.1498	0.9569
	MAT	<u>37.9904</u>	<u>0.9418</u>	39.4787	0.9425
	PFT	**38.0119**	**0.9420**	40.2166	0.9572
	EDiffSR	34.0745	0.8825	36.1214	0.9377
VENµS (10m → 5m)	SwinIR	37.4958	0.9388	42.3371	0.9795
	MAT	37.6951	0.9398	**42.5773**	<u>0.9796</u>
	PFT	37.6233	0.9395	42.4817	<u>0.9796</u>
	EDiffSR	34.2691	0.8879	36.9904	0.9469
Finetuned on VENµS (10m → 5m)	SwinIR	37.3789	0.9383	42.3107	<u>0.9796</u>
	MAT	37.5253	0.9395	42.4621	<u>0.9796</u>
	PFT	37.3590	0.9384	<u>42.5581</u>	**0.9797**
	EDiffSR	34.1313	0.8866	37.0319	0.9451

Table 4.2: PSNR and SSIM of our 12 super-resolution models on the Sentinel and VENµS validation sets. Best values per column are **bold**, second best are <u>underlined</u>.

VENµS-trained models generally achieve the highest values, but an exception is the fine-tuned PFT model, which reaches the overall best SSIM score, even surpassing its VENµS-trained counterparts.

Our two state-of-the-art transformer models generally dominate the benchmarks with the highest scores, across all categories. PFT outperforms MAT on Sentinel-2 and finetuned VENµS, while MAT shows a strong performance on the pure VENµS results. SwinIR seems to be only trailing behind slightly, while EdiffSR yields the worst results. This was expected since pixel-wise metrics like PSNR and SSIM generally favor smoother reconstructions. At the same time, they heavily penalize small misalignments, which disadvantages diffusion models that often generate sharper details [39].

Another interesting trade-off can be observed for the finetuned models. Fine-tuning naturally shifts the models towards the VENµS domain. As a result, models that perform extremely well on the VENµS benchmarks show reduced performance on Sentinel. This effect is so strong that the finetuned models even achieve worse Sentinel results than the models trained exclusively on VENµS.

These results underline the clear advantage of transformers in pixel-accurate metrics, though the strengths of diffusion models are highlighted in following experiments.

4.1.3 Image Comparisons with GT

Quantitative metrics such as PSNR and SSIM provide useful benchmarks, but often fail to capture distinct characteristics of a model that are only visible through visual inspection. This is why we generated comparison images of our 12 models against bicubic upsampeling and

the GT. Figure 4.1, Figure 4.2 and Figure 4.3 show imagery from the 20m → 10m Sentinel Validation set, while Figure 4.4 and Figure 4.5 were generated from the 10m → 5m VENµS validation pack.

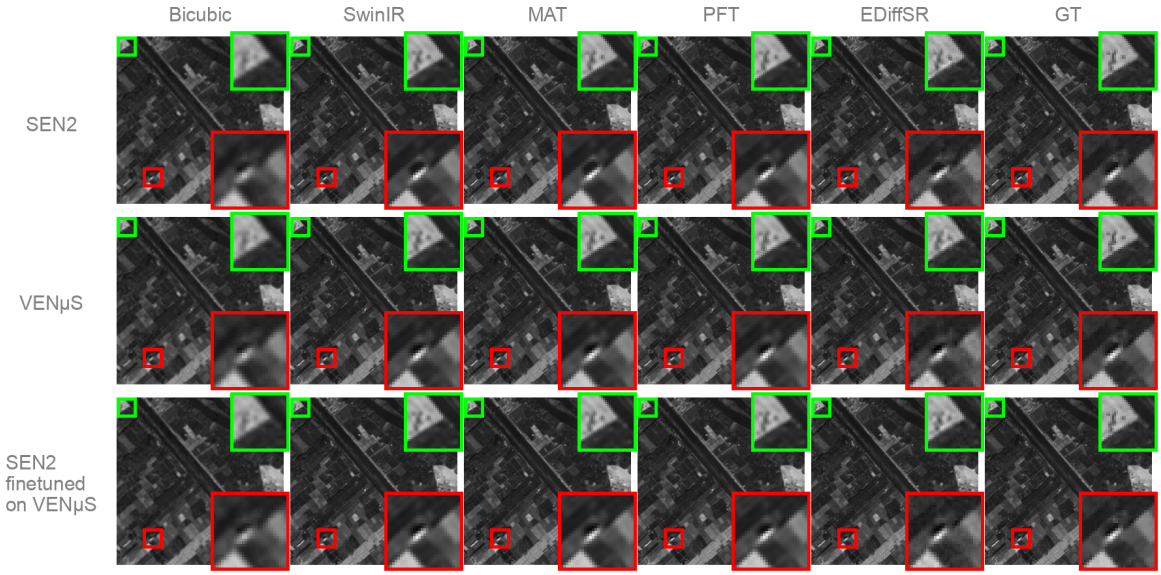

Figure 4.1: Visual comparison of bicubic interpolation, our 12 super-resolution models, and ground truth on a Sentinel-2 validation image.(I)

Before looking into individual model performances, it is worth noting that the ground-truth images from the VENµS validation set generally appear slightly smoother than those from Sentinel. This highlights the impact of different satellite sensors and why comparison across datasets can lead to flawed conclusions.

A general trend across all comparisons is how the images of each model, from Bicubic, SwinIR, MAT, and PFT to finally EDiffSR, become progressively sharper. However, it is still clearly visible that the transformer architectures favor generating smooth images, while EDiffSR tries to recover fine details and textures.

These additional details from EDiffSR can make the images appear more realistic, but they also introduce noticeable noise, as shown in Figure 4.2. For generating these extra details the model must hallucinate subpixel patterns, which can deviate from the actual ground truth. This is visible in Figure 4.3, where EdiffSR introduces way more pattern variety than visible in the GT.

The generated images of MAT and PFT look percepually nearly indistinguishable, while we can notice that SwinIR produces slightly blurrier results.

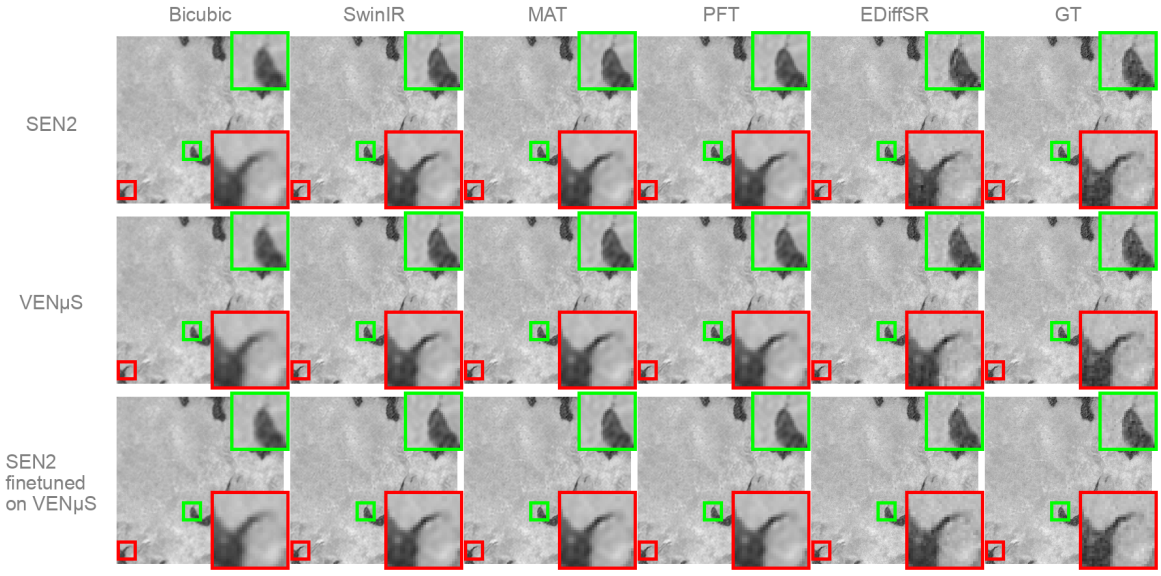

Figure 4.2: Visual comparison of bicubic interpolation, our 12 super-resolution models, and ground truth on a Sentinel-2 validation image. (II)

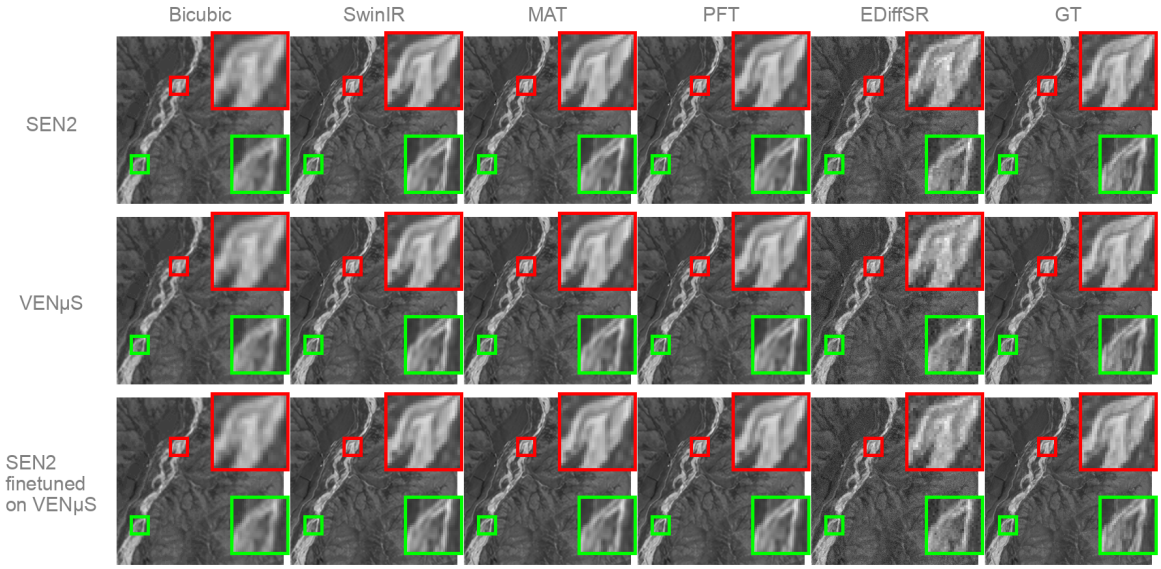

Figure 4.3: Visual comparison of bicubic interpolation, our 12 super-resolution models, and ground truth on a Sentinel-2 validation image. (III)

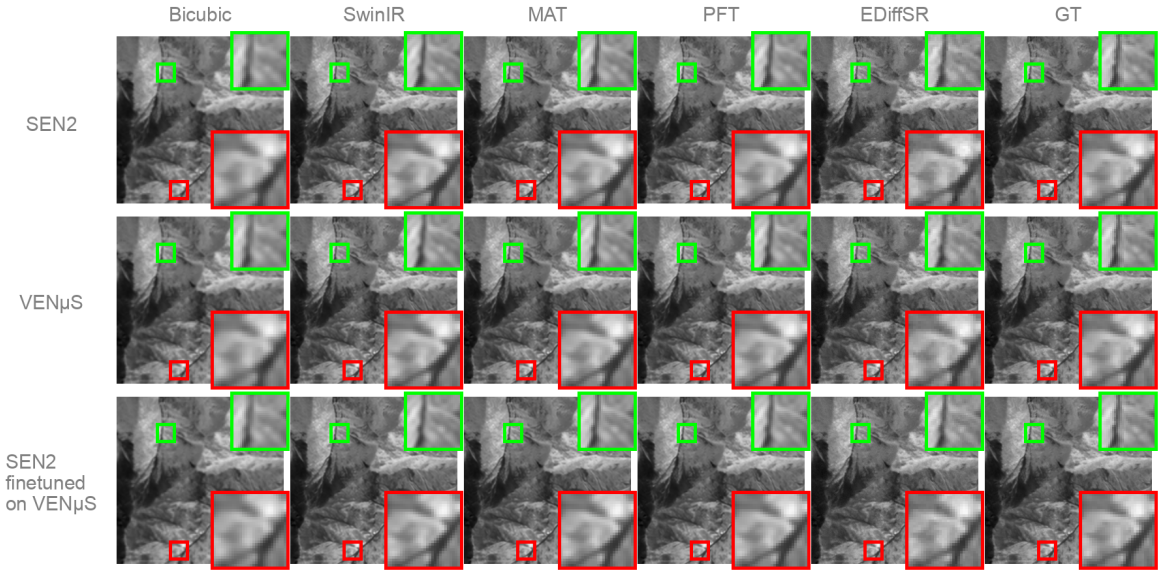

Figure 4.4: Visual comparison of bicubic interpolation, our 12 super-resolution models, and ground truth on a VENµS validation image. (I)

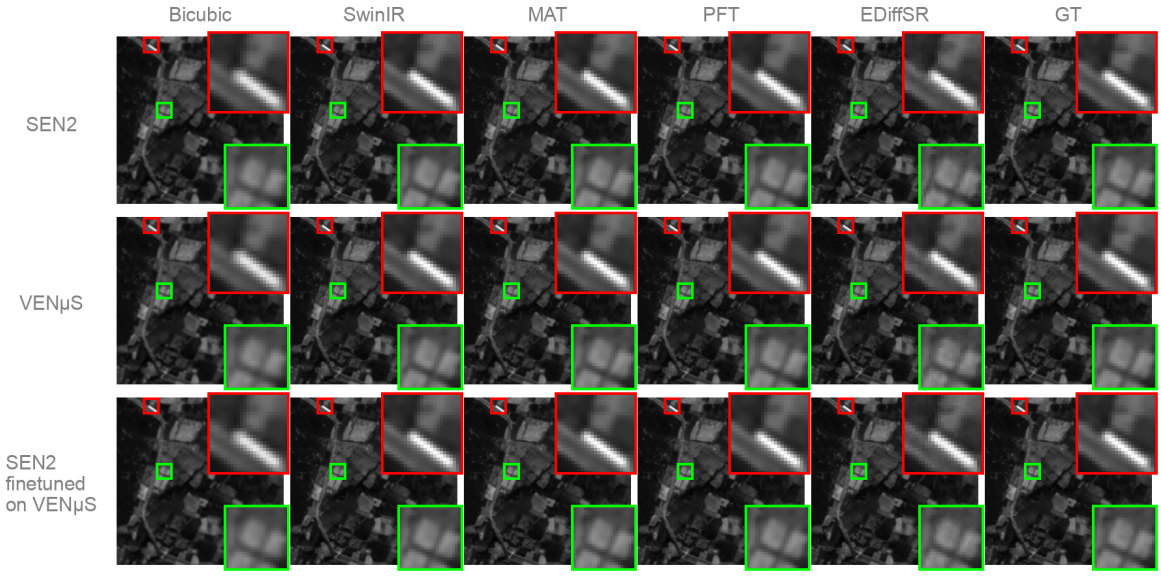

Figure 4.5: Visual comparison of bicubic interpolation, our 12 super-resolution models, and ground truth on a VENµS validation image. (II)

4.2 No-Reference Quality Assessment using NIQE and FID

For most practical tasks, no ground truth data is available for evaluation. This motivates the use of no-reference alternatives to PSNR and SSIM. In this section, we employ our adapted versions of FID (Section 3.3.3) and NIQE (Section 3.3.4) to assess the native upscaling capabilities of our models, while additionally capturing perceptual quality. For NIQE, we used two variants in each subexperiment: a pre-trained model on natural ground imagery from [86] and a custom-trained model fitted to the respective dataset.

In our setup, the **NIQE (custom)** models are the most reliable evaluators, as they were trained on our satellite data and best reflect whether the results look like real images at the target resolution. **FID** looks at the outputs as a group and checks if, overall, they resemble the real reference set, giving a good sense of realism, but can sometimes hide errors in single images. **NIQE (pretrained)** comes from natural photos and mainly shows whether the outputs look smooth and free of artifacts, but it is not suitable for judging satellite quality. For all three, lower values mean better quality.

In the following sections, we compare the performance of our twelve trained models against bicubic upsampling and a corresponding reference set. The three sub-experiments are: (1) a perceptual validation with our Sentinel validation pack (20m $\rightarrow$ 10m), (2) native 20m $\rightarrow$ 10m upscaling using Sentinel-2's 20m bands with 10m bands as reference, and (3) a 10m $\rightarrow$ 5m task starting from Sentinel-2's 10m bands and using a subset of VENµS as 5m reference. Together, these tests cover both spatial gaps and provide insights into the real-world applicability of our models. To achieve a fair comparison between all our models, we have leveraged a new custom dataloader that ensured consistent normalizations across all models by utilizing precomputed statistics. This was necessary as we needed to remap the complete Sentinel tile statistics to our already inferred patches. We also employed image identifier lists so that all machine learning models and the bicubic operation infer the same 10,000 images per experiment. Finally, we ensured that FID was always computed symmetrically, with 10,000 images in both the reference and generated sets.

4.2.1 Perceptual Evaluation on Sentinel-2 (20m $\rightarrow$ 10m) with Ground Truth

The first NIQE and FID experiment evaluates perceptual quality and naturalness on the same task our Sentinel-2 models were trained on, super-resolving from 20m to 10m GSD using our Sentinel-2 Validation Pack (Section 3.1.1). Although the main purpose of this experiment was tracking the perceptual quality of our models under the conditions they were trained on, it additionally provided a sanity check for our adapted FID and NIQE implementations, ensuring they produce reasonable and consistent values.

While the custom NIQE model (S2 10m) was trained on the full 70k images of our Sentinel-2 Validation, we chose a subsample of 10k images for symmetrically calculating FID. The Bicubic Baseline was generated by up-sampling the degraded LQ images from the validation pack using Rasterio's bicubic interpolation [77]. The results (Table 4.3) of this experiment provide a direct comparison of perceptual quality on the standard 20m $\rightarrow$ 10m task for all our 12 model configurations, the reference sets, and bicubic as a baseline.

Dataset (GSD)	Model	FID ↓	NIQE (pretrained)↓	NIQE (S2 10m)↓
	Bicubic	23.4393	5.9932	8.4191
Sentinel 2 (20m → 10m)	SwinIR	16.3962	6.7250	9.4056
	MAT	16.0830	6.7303	9.4222
	PFT	16.1416	6.7112	9.3485
	EDiffSR	**8.0020**	**4.7260**	**5.6492**
VENµS (10m → 5m)	SwinIR	15.1960	6.4821	9.9797
	MAT	15.3941	6.5272	9.9559
	PFT	15.0179	6.5230	9.9640
	EDiffSR	12.6340	<u>5.1119</u>	<u>6.5807</u>
Finetuned on VENµS (10m → 5m)	SwinIR	14.9111	6.4914	10.0020
	MAT	14.9444	6.5689	9.9613
	PFT	14.6423	6.5326	9.9916
	EDiffSR	<u>10.7600</u>	5.4254	8.2896
	Reference	0.0	4.4380	4.8794

Table 4.3: Comparison of FID, NIQE (pre-trained), and NIQE (custom-trained on Sentinel-2 validation data at 10 m) across the 12 super-resolution models, Bicubic upsampling, and the Reference set. For each metric, the best result (excluding Bicubic and Reference) is shown in **bold**, and the second-best is <u>underlined</u>.

Metric Sanity Check

Before discussing model performance, it is important to validate that the metrics behave as expected. When applying the reference set against itself, the **FID** must yield a score of 0.0, which confirms correctness in our results. Additionally, Bicubic consistently performs worse than all other super-resolution techniques, which is an expected behavior. All other FID values also fall within a reasonable expected range, indicating a reliable performance. Importantly, FID's feature extraction backend InceptionV3 was trained on RGB 8-bit imagery [84] [11], a quite different domain from our single-band 16-bit satellite data. As the metric compares our models against the satellite data references, we think it can still serve as a valuable metric.

For **NIQE (pre-trained)**, results can be more difficult to interpret as the model was originally trained on pristine ground images instead of our satellite imagery. This metric is included with the goal of providing a perspective of how natural model images might appear to the human eye. It is not unusual for Bicubic to outperform some SR models with NIQE, as it is particularly sensitive to artifacts and can penalize hallucinated details harshly [12].

The **custom NIQE models** are the most relevant and reliable measures for our experiments, as they were directly optimized and trained on our satellite domain. The reference set achieves very low scores, as expected, confirming that the model is well aligned. In theory, NIQE

can be gamed, meaning higher scores than the reference are possible [12]. Compared to the pre-trained NIQE version, the values are generally higher/stricter because the satellite-specific distribution is narrower than the one for diverse ground imagery.

Model Performance Overview

The **EDiffSR models** strongly outperform all transformer models across all metrics. In line with our expectations, the model trained on Sentinel-2 (20m $\rightarrow$ 10m) achieves the overall best results, while the Venus-trained and fine-tuned variants take the second-best places: the Venus-trained model performs better on NIQE, while the finetuned model performs better on FID. The three **transformer models** achieve very similar results, with only minor fluctuations in ranking. Models trained or fine-tuned on Venus generally perform better on FID and NIQE (pre-trained), suggesting that their outputs share characteristics more closely related to natural ground imagery. In contrast, the Sentinel-trained models dominate on the custom-trained NIQE metric, which is consistent with our expectations. While the gap between transformer models is small, PFT and MAT show a slight edge over SwinIR, which only secures some wins on the pre-trained NIQE metric. For visual model comparisons on the Sentinel-2 validation set, see Section 4.1.3.

4.2.2 Native Blind Super-Resolution on Sentinel-2 (20m $\rightarrow$ 10m)

Although the reference set and spatial gap of 20m $\rightarrow$ 10m remain consistent with the first experiment, the previous LQ images generated by bicubic degradation (Section 3.1.1) were replaced with 10,000 native Sentinel-2 20m band patches. These patches were directly super-resolved to 10m resolution by the twelve models and bicubic interpolation. For evaluation, they were compared against the corresponding 10m reference patches. The goal of this setup is to assess how well the models generalize to real sensor data, where the underlying degradation process is not explicitly defined. This provides a measure of robustness to blind degradation kernels that may differ from the synthetic bicubic downsampling used during training. The NIQE model used here is identical to the previous experiment, trained on the full Sentinel-2 10m validation set.

Model Metric Performance Overview

In this section we discuss the metric results (Table 4.4) of the 20m $\rightarrow$ 10m native super-resolution experiment. As this experiment deviates from the training setup of our models, they achieve higher (worse) FID values compared to the previous test. The sanity checks stay consistent, applying the reference set to FID still yields zero, bicubic still achieves the worst FID scores, and the reference sets secure the best scores on NIQE.

Notably, EDiffSR shows very consistent behavior, once again achieving the best scores across all metrics and even reproducing the same best and second best ranking pattern observed in the previous experiment. The transformer models also remain close to each another, with only marginal differences. PFT and MAT slightly outperform SwinIR overall,

Dataset (GSD)	Model	FID ↓	NIQE (pretrained)↓	NIQE (S2 10m)↓
	Bicubic	34.0979	6.4450	8.9070
Sentinel 2 (20m → 10m)	SwinIR	21.8058	6.7661	9.9579
	MAT	21.5178	6.7412	10.0000
	PFT	21.5094	6.7090	9.9083
	EDiffSR	**17.1899**	**4.9288**	**5.6667**
VENµS (10m → 5m)	SwinIR	21.2752	6.5870	10.5191
	MAT	21.3294	6.5903	10.5054
	PFT	21.1027	6.6008	10.5308
	EDiffSR	19.7959	<u>5.3150</u>	<u>6.8834</u>
Finetuned on VENµS (10m → 5m)	SwinIR	21.0822	6.5769	10.5175
	MAT	21.0619	6.6106	10.5163
	PFT	20.7453	6.5817	10.5015
	EDiffSR	<u>18.5627</u>	5.5773	8.0558
	Reference	0.0	4.4380	4.8794

Table 4.4: Perceptual evaluation on native Sentinel-2 20 m → 10 m super-resolution. Comparison of FID, NIQE (pre-trained), and NIQE (custom-trained on Sentinel-2 validation data at 10 m) across the 12 super-resolution models, Bicubic upsampling, and the Reference set. For each metric, the best result (excluding Bicubic and Reference) is shown in **bold**, and the second-best is <u>underlined</u>.

except in the NIQE (pretrained) column, where SwinIR retains a small advantage, particularly on the VENµS-trained and finetuned models.

The models trained on Sentinel-2 (20m → 10m) still perform best on the custom trained NIQE metric, which aligns with expectations, but even here the gaps between training setups are minor. This suggests that our models generalize reasonably well across datasets, maintaining stable native perceptual performance despite differences in training domain.

Visual Comparison of Model Outputs (native 20m → 10m)

For the native Sentinel-2 20m → 10m task, we provide two comparison visualizations (Figure 4.6 and Figure 4.7) against bicubic upsampling. As we are working with native super-resolving, there is no ground truth to compare with. For this native setting, the transformer models produce sharper boundaries and more stable structures, with only minor differences between them. EDiffSR generates the most fine-grained detail and higher contrast, but sometimes adds extra texture in uniform areas (see Figure 4.6). Interestingly enough, the differences between the same models trained on different datasets are also barely noticeable. Overall, the visual trends are consistent with the quantitative metrics: transformers provide clean and faithful reconstructions, while EDiffSR pushes more detail at the cost of introducing

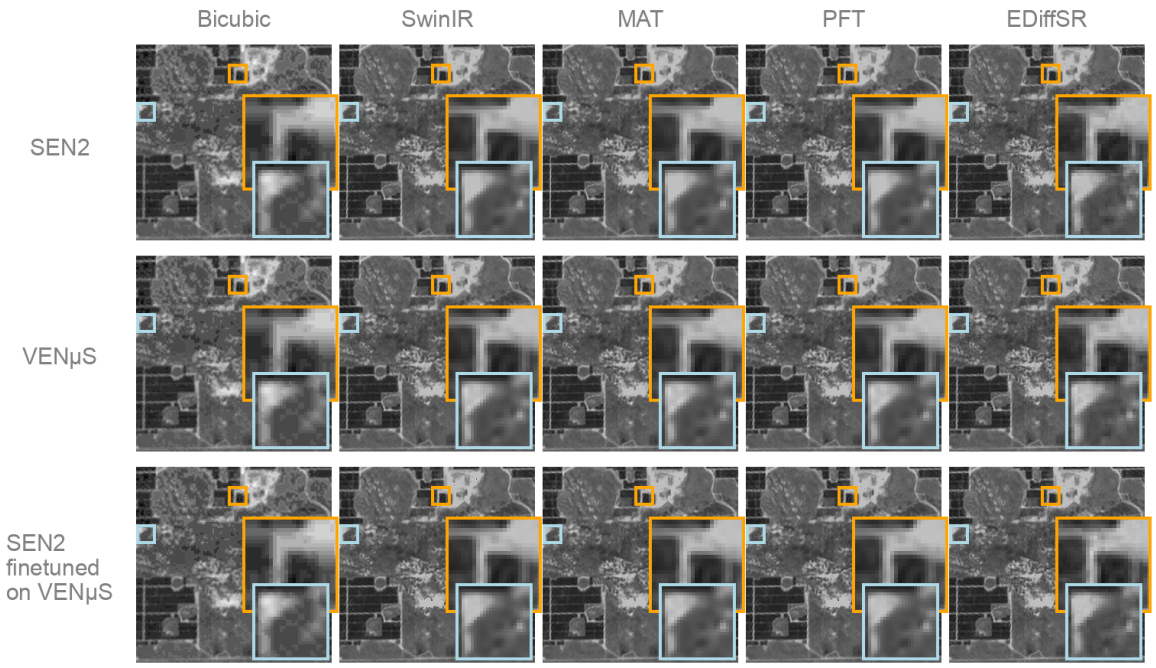

Figure 4.6: Visual comparison of SR outputs for the native Sentinel-2 20m → 10m task. (I)

noise.

4.2.3 Cross-Dataset Super-Resolution (Sentinel-2 10m → VENµS 5m)

The final experiment goes a step further by evaluating across different satellite datasets. Sentinel-2 10m bands are natively upscaled to 5m and compared against the VENµS 5m bands of the validation set, which contains 118,256 patches. For NIQE, we trained a new model on the full VENµS validation dataset (Section 3.1.2) to capture the characteristics of 5m imagery, while for FID, a subsample of 10,000 images was used as reference.

All super-resolution methods, including the bicubic baseline, were applied to the Sentinel-2 10m images. This setup allows us to test the native upscaling capabilities of the VENµS trained and finetuned models in particular, which should excel on the 10m → 5m spatial gap. In contrast, for Sentinel-2 trained models, this is the first test evaluating their native upscaling capabilities on a different spatial gap.

Additionally, this experiment is especially challenging because our Sentinel-2 and VENµS data sets rely on different normalization approaches, as discussed in Section 3.1.3.

Model Metric Performance Overview

With the large domain shift caused by the cross-dataset approach, the FID results (see Table 4.5) got inflated, but the sanity checks stayed consistent. For the first time, NIQE (pretrained)

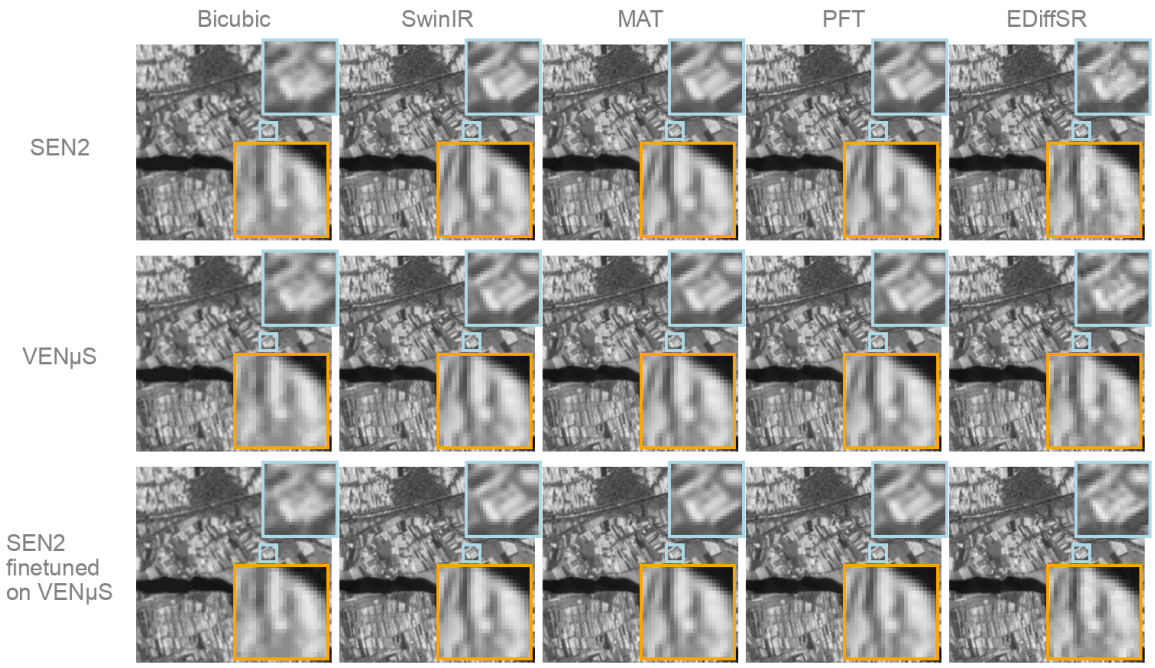

Figure 4.7: Visual comparison of SR outputs for the native Sentinel-2 20m→ 10m task. (II)

scores the reference set lower than some EDiffSR performances, hinting that VENµS imagery is less aligned with the natural image statistics than Sentinel-2. Consequently, this could mean that models that generate images closer to the VENµS domain can be punished by this metric.

In the cross-dataset setting, we observe a clear shift in model ranking compared to the previous experiments. While the diffusion-based EDiffSR excelled in the earlier tests, the transformer models now achieve the best results on both NIQE (custom) and FID, indicating that their outputs are more faithful to the VENµS 5 m reference distribution and local statistics. EDiffSR still performs strongly on NIQE (pretrained), hinting that EDiffSR still produces images perceptually closer to natural ground imagery. In this test, the performance gap of the transformer models is even smaller, with no real winner emerging. Another interesting trend is that some Sentinel-2 trained variants score better on FID, whereas VENµS-trained/finetuned variants lead on the NIQE (custom) metric. This can happen because FID uses InceptionV3 features and assesses the set globally, while NIQE (custom) is tuned to VENµS-specific local statistics. Between the two, NIQE (custom) is the more reliable signal of cross-sensor faithfulness in this experiment.

In short, transformers adapt better to the VENµS domain, while EDiffSR still produces perceptually clean results but does not capture the VENµS style as well. Among the transformers, the differences are small, so there is no clear winner.

Dataset (GSD)	Model	FID ↓	NIQE (pretrained)↓	NIQE (VENµS 5m)↓
	Bicubic	64.1758	5.9855	9.5205
Sentinel 2 (20m → 10m)	SwinIR	**58.4494**	6.2597	9.2336
	MAT	58.7443	6.2472	9.1720
	PFT	<u>58.7150</u>	6.2323	9.1658
	EDiffSR	62.1256	**5.0281**	12.5796
VENµS (10m → 5m)	SwinIR	59.8736	6.1745	<u>8.6983</u>
	MAT	59.9421	6.1825	8.7010
	PFT	60.0405	6.1833	**8.6838**
	EDiffSR	64.1453	<u>5.1510</u>	10.4690
Finetuned on VENµS (10m → 5m)	SwinIR	59.9298	6.1765	8.7359
	MAT	59.8633	6.1818	8.7456
	PFT	59.7854	6.1636	8.7587
	EDiffSR	62.7515	5.6922	11.9561
	Reference	0.0	5.8562	4.8466

Table 4.5: Perceptual evaluation on cross-dataset super-resolution from Sentinel-2 10m → VENµS 5m. Comparison of FID, NIQE (pre-trained), and NIQE (custom-trained on VENµS validation data at 5m) across the 12 super-resolution models, Bicubic upsampling, and the Reference set. For each metric, the best result (excluding Bicubic and Reference) is shown in **bold**, and the second-best is <u>underlined</u>.

Visual Comparison of Model Outputs (native 10m → 5m)

For the cross-dataset Sentinel-2 10m → VENµS 5m task, we provide two qualitative comparisons (see Figure 4.9 and Figure 4.8). Without ground truth, we focus on perceptual trends. Here, EDiffSR struggles more clearly than in previous experiments, introducing noise (see Figure 4.9) and overly sharp artifacts (see Figure 4.8). In contrast, the transformer models produce smoother yet coherent structures that stay closer to the bicubic upsampling. This matches the quantitative results, where the transformers adapt better to the VENµS domain, while EDiffSR produces less stable outputs.

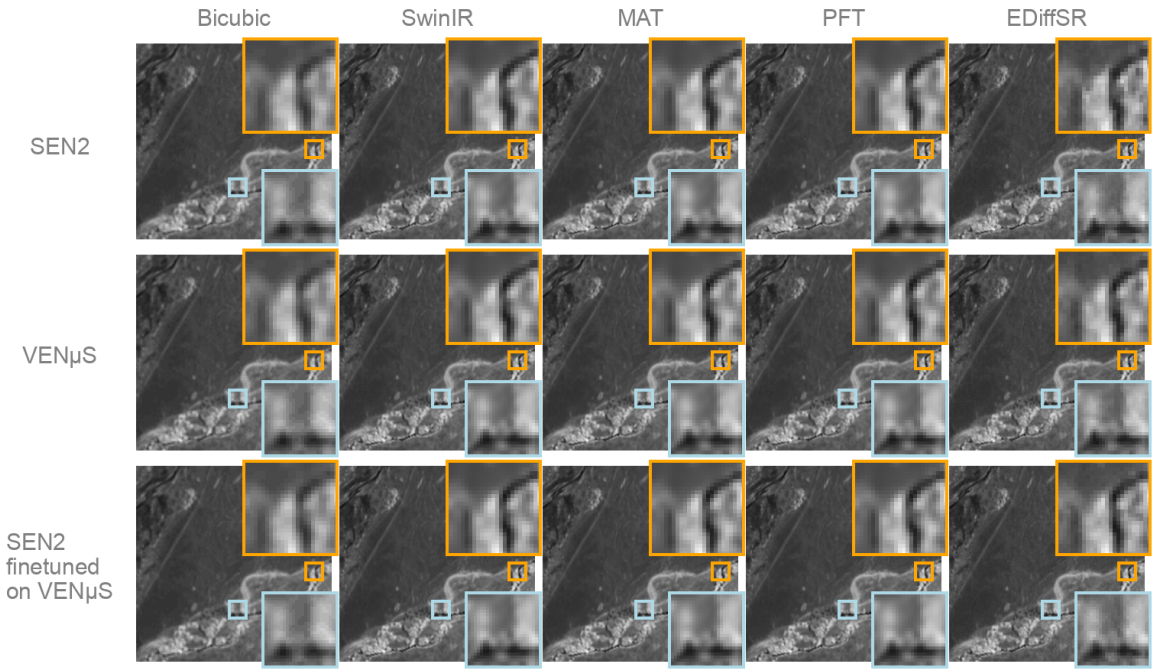

Figure 4.8: Visual comparison of SR outputs for the native Sentinel-2 10m→ 5m task. (I)

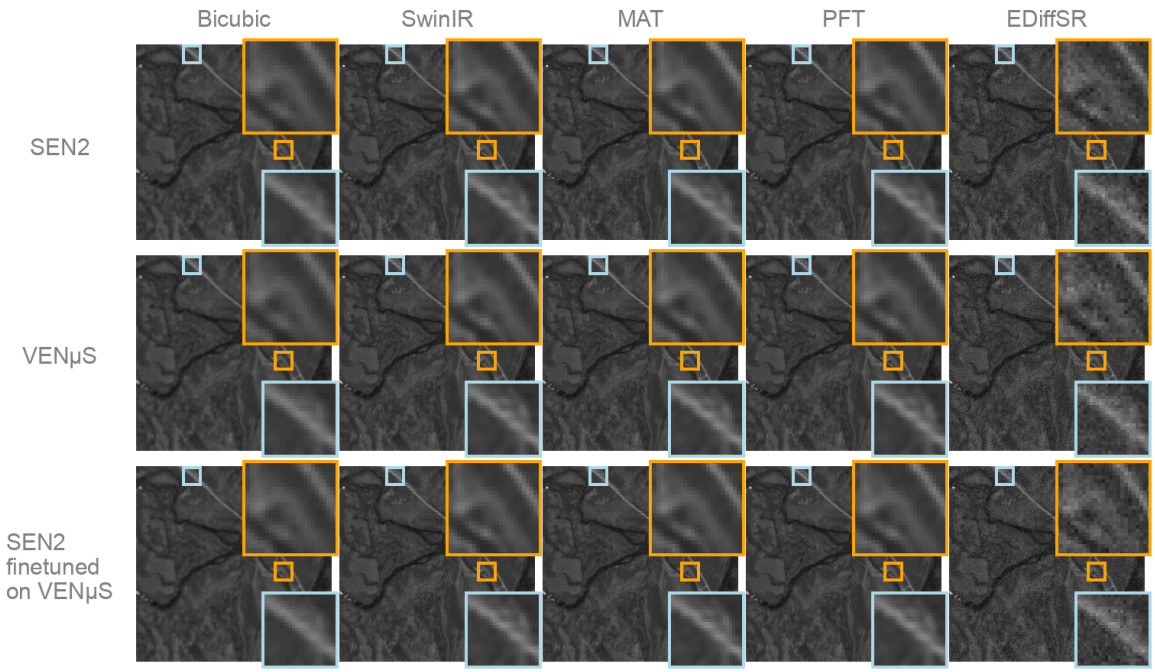

Figure 4.9: Visual comparison of SR outputs for the native Sentinel-2 10m → 5m task. (II)

4.3 Evaluation on Field Detection Downstream Task

While previous experiments focused on quantitative and perceptual metrics, they cannot prove the utility of super-resolution for real-world applications. To address this, we introduce a practical downstream task to evaluate our models against bicubic interpolation. Our choice fell onto a field boundary detection task, where we compare the trustworthiness of super-resolved Sentinel-2 imagery by generating soil masks and comparing their bounds against a high-resolution ground truth. We first explored standard evaluation metrics such as AUROC [91] and then turned to the boundary-focused F1 score [92], which better captures the strengths of super-resolution. By assessing the accuracy of the detected parcel outlines, we can directly test whether and how super-resolving improves the result of our downstream task. If our models outperform bicubic interpolation, this provides strong evidence that super-resolution can be a valuable preprocessing step for real-world applications.

4.3.1 The Setup: Sentinel-2 Data and PlanetScope Ground Truth

For our downstream experiment, we worked with two Sentinel-2 tiles located near Munich, captured on 8 March 2025 [1] and March 18, 2025 [2]. They were chosen as they were captured in a similar time frame to imagery from PlanetScope, and contain barely any clouds.

PlanetScope is a commercial Earth observation constellation operated by Planet Labs [93][94], consisting of hundreds of small "Dove" satellites. It provides high-quality daily global coverage at 3m spatial resolution, making it a strong pick as ground truth for our lower resolution Sentinel-2 data.

Through DLR [88] and its Department of Imaging Spectroscopy [95], we got access to PlanetScopet [94] tiles and to a pre-computed mask of field parcels, essentially polygons outlining individual agricultural fields. These parcels were generated by DLR utilizing 23 PlanetScope scenes acquired between March and May 2025. The individual fields were computed by grouping areas that change together over this time series. Additionally, urban areas and forests were manually removed so that only the agricultural fields are left over. This trustworthy, high-quality mask with embedded parcel IDs was our primary GT used in the following experiments. Figure 4.10 shows one of Planetscope tiles over the bigger Sentinel-2 tile and the GT parcelmask.

4.3.2 Preprocessing: From Sentinel-2 Bands to 5 m Inputs

In order to compare our super-resolution models against the PlanetScope ground truth, we first needed to generate soil masks from Sentinel-2 imagery at 5m resolution. For our index-based soil classification algorithms, we require three bands: one red band (B04), one band in the NIR (B08) and one band for SWIR (B12), upscaled to 5m.

While B04 and B08 are provided at Sentinel-2's highest resolution of 10m, B12 requires an additional step to go from 20m $\rightarrow$ 5m. Initially, we upscale the B12 Band from 20m

[1] Sentinel Tile: S2B_MSIL2A_20250308T100749_N0511_R022_T32UPU_20250308T122841
[2] Sentinel Tile: S2A_MSIL2A_20250318T101751_N0511_R065_T32UPU_20250318T160400

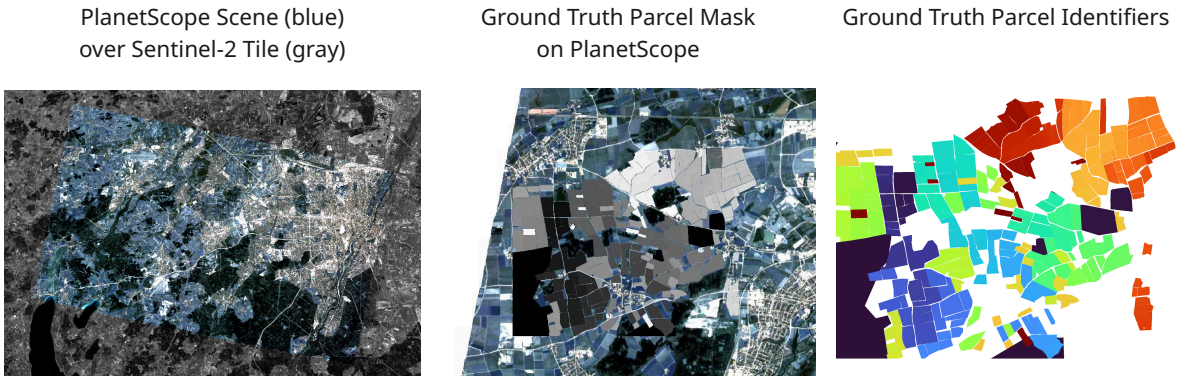

Figure 4.10: Illustration of our data. Left: PlanetScope [94] scene overlaid on a Sentinel-2 tile. Middle: Ground truth parcel mask aligned with the PlanetScope crop. Right: Parcel mask visualization with unique identifiers shown in different colors.

to 10m using our Sentinel-2 trained model configurations. Subsequently, all three bands (B04, B08, and the super-resolved B12) were further upscaled from 10m to 5m using our VENµS-finetuned models. This two-step approach allows us to use the strength of our model specializations: Sentinel-trained versions handle their natural scenario of 20m $\rightarrow$ 10m, while VENµS-finetuned models can close their trained 10m $\rightarrow$ 5m gap.

With this process, we upscaled the three described bands using our five contenders: SwinIR, MAT, PFT, EDiffSR, and Bicubic (utilized as a baseline). These bands then served as input for soil mask generation using the NDVI and NBR indices, which we describe in the following subsection.

4.3.3 Soil Mask Generation with NDVI+NBR

To classify bare fields from the super-resolved Sentinel-2 imagery, we relied on a combination of two well-known vegetation indices: NDVI and NBR. The Normalized Difference Vegetation Index (NDVI) [96] tracks the contrast between red (B04) and near-infrared (B08) reflectance to recognize vegetation:

$$NDVI = \frac{B08 - B04}{B08 + B04} \tag{4.1}$$

In contrast, the Normalized Burn Ration (NBR) [97] can be used to track soil and relies on near-infrared (NIR)(B08) and shortwave-infrared (SWIR)(B12) data.

$$NBR = \frac{B08 - B12}{B08 + B12} \tag{4.2}$$

The combination of both, called NDVI+BNR [98], emphasizes bare soil pixels by suppressing vegetation signals (NDVI) and highlighting non-vegetated surfaces (NBR), leading to a more robust result.

$$NDVI + NBR = \left(\frac{B08 - B04}{B08 + B04}\right) + \left(\frac{B08 - B12}{B08 + B12}\right) \tag{4.3}$$

To generate the mask, an image only containing values of 1 and 0, the Department of Imaging Spectroscopy applied the NDVI+NBR on each pixel of our super-resolved Sentinel tiles and classified them using thresholding. The thresholds for each pixel were derived from a precomputed threshold map of Europe. For our study, the mean threshold was 0.194. For more information about the process, look into these papers [99][98]. To further improve the mask, urban areas were filtered using ESA WorldCover 2021 data [100], while clouds, shadows, and cirrus were removed using the Sentinel-2 Level-2A scene classification layer (SCL) [101]. After applying this procedure to all our models, we received binary soil masks at 5m resolution for our two Sentinel tiles. These masks are the basis for all the following experiments.

4.3.4 Visual Change Analysis Against Bicubic Interpolation

To better understand how our super-resolution models alter soil masks compared to the bicubic baseline, we visualized their pixel-wise differences (Figure 4.11). Red pixels indicate areas that only bicubic classifed as soil, while blue pixels mark only model-specific classifications. The darker gray shows where the masks of bicubic and our super-resolution models overlap.

Before we talk about the characteristics of specific models, there are some general trends we can observe over all models. Most differences are visible at the edges of field parcels: Sharper edges generated by the SR models differ from the smoother outlines produced by bicubic interpolation. Another interesting observation is that mainly red pixels appear inside of field parcels, while blue pixels show up on the borders or even outside of classified soil areas. This is due to the blurring effect of bicubic, which plugs holes in uniform areas. At the same time, our super-resolution models introduce sharper textures and stronger local contrast, which can lead to pixels dropping under the NDVI+NBR thresholds. Outside of parcels, the super-resolution models' contrast can lead to more variation and a noisy mask. This effect is way more substantial for EDiffSR, as diffusion models try to reproduce sharp edges and strong textures, which introduces a lot of noise into the mask. While the transformer models (SwinIR, MAT, and PFT) exhibit far more consistent behavior, with fewer spurious changes and more precise boundaries, they also experience the same issues. Importantly, we cannot be certain whether the models are hallucinating details or just capturing natural variability, as real fields are inherently messy and often contain small gaps or irregularities. This was the first foreshadowing of why a simple area comparison against our GT parcels could punish our super-resolution models (the AUROC Section 4.3.6), in contrast to bicubic, which just smoothes out the image. As the most significant impact of our super-resolution models was witnessed at the edges of our field parcels, this motivated our primary evaluation using a boundary-focused metric such as Boundary F1 (section: 4.3.7).

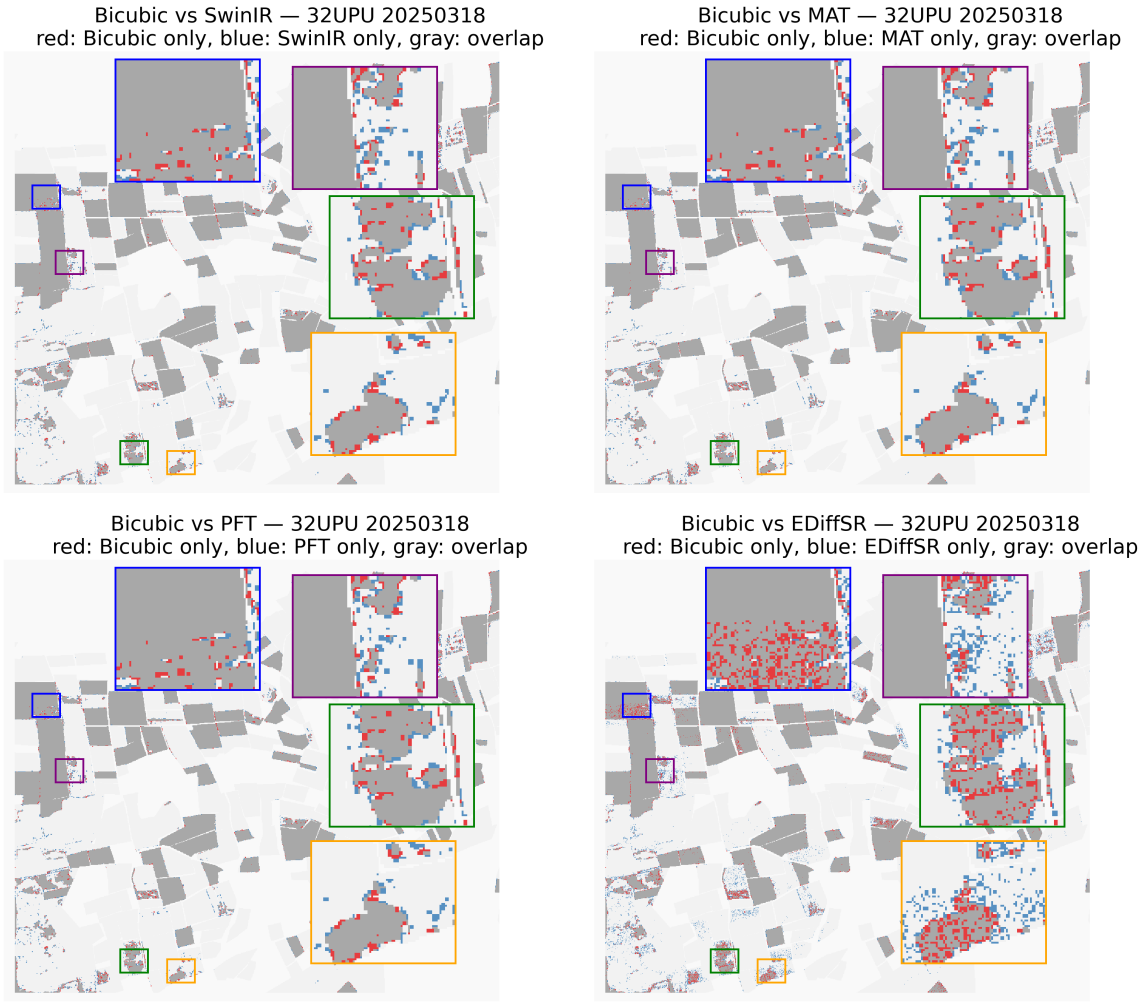

Figure 4.11: Pixel-wise comparison of soil masks generated by our super-resolution models against the bicubic baseline (tile 18 March 2025). Red indicates soil pixels detected only by bicubic, blue pixels detected only by the super-resolution model, and gray pixels where both methods overlap.

4.3.5 Ground Truth Parcel Selection

The PlanetScope [94] ground truth mask contains all agricultural parcels in the study area, while our Sentinel-based soil masks only capture fields that are bare at the specific acquisition dates. With the goal of ensuring a fair evaluation, we filtered the ground truth parcels so that only relevant fields are utilized for further metric calculations. This was achieved by leveraging the embedded parcel identifiers, which allowed us to compute per-parcel statistics on the fraction of soil pixels.

Before applying any filtering strategies, we removed all parcels that contain less than 80 pixels in area, as they can behave noisy in metrics down the line. The two filtering strategies

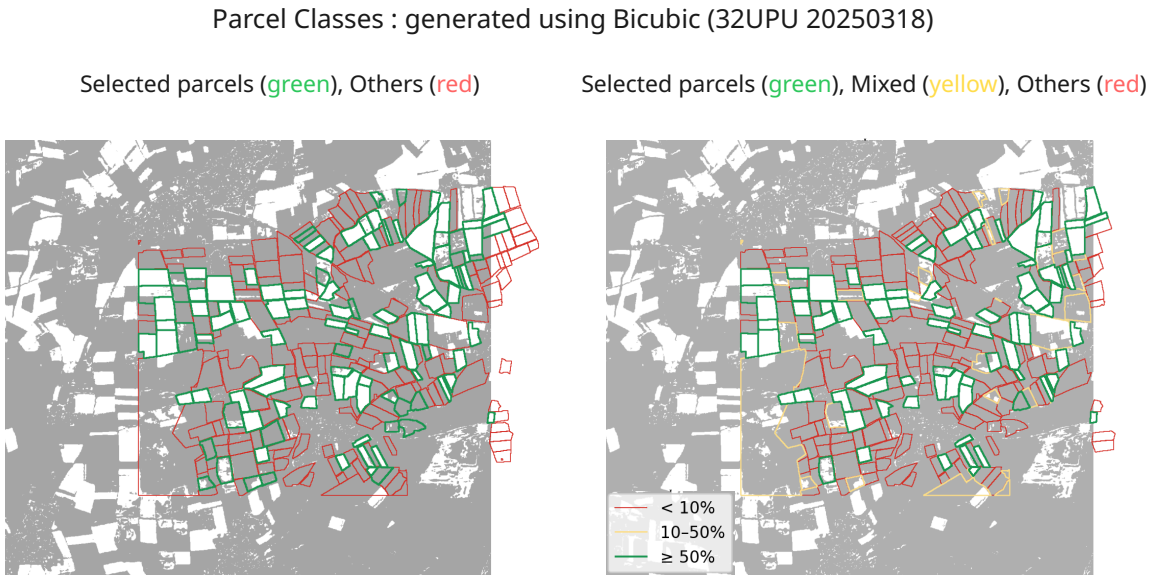

Figure 4.12: Filtering of PlanetScope ground truth parcels based on soil coverage derived from the bicubic mask (tile 32UPU, 18 March 2025). Left: binary filtering / Right: three-class filtering

we used are visualized in Figure 4.12 on tile 32UPU (18 March 2025). To classify the parcels, we used the bicubic soil mask and calculated the parcel soil coverage for all ground truth parcels. In the first strategy, parcels with at least 50% soil coverage were selected (green), while all others were marked negative (red). This binary filtering was later utilized for the Boundary F1 evaluation in Section 4.3.7. For the second filter strategy, we introduced a third mixed class for partially bare parcels, which happens when fractions of these field parcels are in different agricultural states. Parcels with less than 10% were marked red, those with $10 - 50\%$ were assigned the new mixed class (yellow), and those with more than 50% coverage remained selected (green). This technique was used for our AUROC experiments, discussed in the next section.

4.3.6 AUROC as a First Attempt and Its Pitfalls

As foreshadowed in Section 4.11, the evaluation attempt using AUROC produced misleading results in which some super-resolution models appeared to perform worse than bicubic interpolation. In the following, we explain our initial setup of the AUROC experiment, look at metric results and visualizations. At the end of this section, it becomes clear that the metric fails in our setting due to limitations in the ground truth, not due to shortcomings of our super-resolution models.

AUROC Metric and Experimental Setup

The more intuitive approach to validate our soil masks would be the Intersection-over-Union (IoU) [102], which measures the overlap between predicted and GT parcels. However, IoU requires fixing a threshold on the predicted soil fraction (e.g. parcel is bare if at least 50% of its pixels are classified as soil). However, a slow shift in thresholds can lead to entirely different results, making the metric highly sensitive to noise or our partially bare parcels. This is why we selected the **Area Under the Receiver Operating Characteristic Curve (AUROC)**[91], as an alternative, better approach. AUROC avoids this problem by generating a ranking of parcels, not bound to any thresholds, by asking this question: if we randomly pick one bare parcel and one non-bare parcel, does the model assign the bare parcel a higher soil fraction score? If a super-resolution model consistently ranks bare parcels above non-bare parcels, it receives a high AUROC score, making the algorithm more robust for our setup.

In our implementation, we utilize the in Section 4.3.5, introduced mixed class, excluding such labeled parcels from the evaluation, as they represent ambiguous cases where only part of the field is bare. By ignoring them, AUROC is calculated only between clearly bare and clearly non-bare parcels. This reduces label noise and makes the comparison more fair. In the following results, we report AUROC in both macro and micro variants, as explained in the next section.

AUROC Results and Limitations

Table 4.6 reports the AUROC scores for both Sentinel-2 tiles (18.03.2025, 08.03.2025), evaluated in macro and micro variants. Macro averages all parcel scores with equal weight, while micro averages parcels weighted by their area. This provides two different perspectives: macro AUROC reflects how models perform across the population of fields, while micro AUROC highlights their behavior on larger parcels and across all pixels.

Model	Tile 1 (08/03/2025)		Tile 2 (18/03/2025)	
	Macro ↑	Micro ↑	Macro ↑	Micro ↑
Bicubic	**0.9831**	**0.9854**	<u>0.9745</u>	<u>0.9745</u>
SwinIR	0.9773	0.9821	0.9756	0.9758
MAT	0.9779	0.9825	0.9676	0.9682
PFT	<u>0.9784</u>	<u>0.9828</u>	0.9677	0.9684
EDiffSR	0.9771	0.9824	**0.9856**	**0.9856**

Table 4.6: AUROC results for field detection on two Sentinel-2 tiles. Best and second-best results per column are marked in **bold** and <u>underline</u>.

Looking at the results, we can immediately identify some misleading values. Before addressing the unexpectedly strong bicubic performance, let us first analyze the other models.

EdiffSR yields contradictory results, achieving the highest scores on the 8 March tile and some of the worst results on the second tile. Additionally, visualized in Figure 4.13, we can

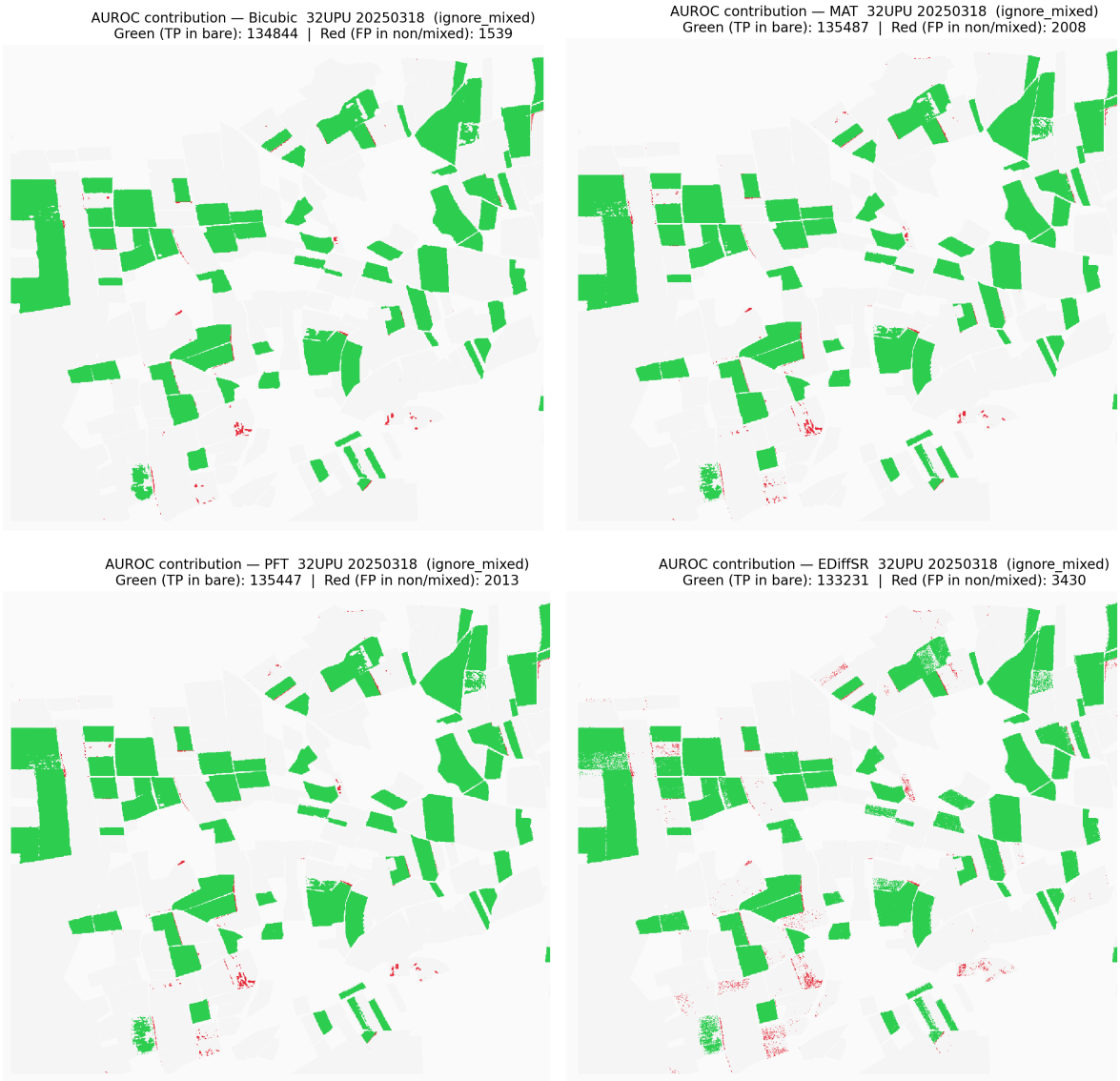

Figure 4.13: AUROC contribution maps for selected Models: Bicubic, MAT, PFT, EDiffSR (Tile 18 March 2025). Green pixels indicate true positives inside bare parcels, red pixels false positives inside non-bare parcels.

observe that EDiffSR produces noisy images, with many false positives and holes inside bare parcels, displaying way more inconsistent results than our transformers. Nevertheless, because AUROC only measures whether bare parcels receive higher average soil fractions than non-bare ones, these artifacts do not hurt its ranking performance. The transformer models perform more stably, with only minor variations between them.

Bicubic interpolation reaches the best AUROC scores on the 18 March tile, and second-best on the other date. However, this interpretation is deceptive. Bicubic benefits from its inherent smoothing, which reduces false positives and fills small gaps inside parcels, unfairly boosting AUROC. We came to similar conclusions in Section 4.3.4, where we found that bicubic unifies areas by lowering the contrast. While this effect is rewarded here, it does not reflect reality. In comparison, our super-resolution models introduce sharper boundaries and more realistic within-field variation, which AUROC penalizes due to our simplified ground truth.

And this is where the real issue lies: our ground truth wasn not made to correctly represent variations inside a field. Nature is diverse, agricultural fields are not unified areas, like our gt parcel masks, but they contain holes and even some bare soil outside of fields. Moreover, fields change, and while our ground truth captures data from a long period of time, our Sentinel imagery stems from a single acquisition.

This raises the central question: What is a suitable metric for our ground-truth and super-resolution setup? Revisiting the bicubic comparison experiment, we observed that most discrepancies between bicubic and our models occur at field borders. Naturally, super-resolution can refine the blurry edges left by bicubic interpolation. Furthermore, the meaningful information our ground truth field parcels describe is not the data inside the fields, but in the boundaries they delineate. This is why in the next chapter, we choose F1-Boundary Detection as a reliable evaluation metric, playing both into the strength of our super-resolution model and working with the limitations of our ground truth.

4.3.7 Measuring Boundary Quality with BF1

The previous AUROC evaluation highlighted the limitations of overlap- and ranking-based metrics for our task. Now we will try out a better fit for our setup: The **Boundary F1 (BF1)**[92] directly evaluates how well the predicted parcel outlines align with the ground truth boundaries, utilizing the real strength of both our super-resolution models and our field parcel ground truth mask. In this section, we will calculate the BF1 on both our tiles and also visually inspect the results.

Boundary F1 Metric and Setup

The Boundary F1 (BF1) score measures how well the predicted parcel edges align with the reference boundaries. Instead of evaluating the overlap of entire areas, BF1 compares boundary pixels within a tolerance radius τ. A predicted boundary pixel is counted as a true positive (TP) if there exists a ground-truth boundary pixel within τ pixels. If not, it is marked as a false positive. Additionally, if a ground-truth boundary pixel has no nearby predicted counterpart, it counts as a false negative (FN).

From these counts, we compute Precision, Recall, and the F1 score as follows:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4.4}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4.5}$$

$$BF1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4.6}$$

In our experiment, we will work with two tolerance radii, $\tau = 1$ and $\tau = 2$ pixels. Smaller tolerances require very sharp edge alignments, whereas larger tolerances are more forgiving, allowing minor discrepancies. With $\tau = 1$, even if a point is predicted one pixel away from the ground truth, it would be marked as a miss, while $\tau = 2$ would approve it.

Similarly to AUROC, we will report both macro and micro values for our two Sentinel-2 tiles. **Macro BF1** averages the per-parcel scores equally, while **Micro BF1** aggregates TP, FP, and FN counts across all parcels before computing the score, which effectively weights by parcel size.

We will be working with the binary class system introduced in Section 4.3.5, excluding parcels with less than 50% bare coverage and parcels that contain less than 80 pixels. After applying this filtering, the tile of the 8 March 2025 preserves 105 parcels while the tile captured on 18 March 2025 works with 81 parcels.

The Results for $\tau = 1$

Model	08/03/2025 (105 parcels)		18/03/2025 (81 parcels)	
	Macro-BF1 ↑	Micro-BF1 ↑	Macro-BF1 ↑	Micro-BF1 ↑
Bicubic	0.8297	0.8319	0.7523	0.7555
SwinIR	**0.8412**	**0.8331**	<u>0.7778</u>	<u>0.7643</u>
MAT	0.8405	0.8326	**0.7779**	**0.7648**
PFT	<u>0.8410</u>	<u>0.8328</u>	0.7773	0.7644
EDiffSR	0.7929	0.7333	0.7041	0.6618

Table 4.7: Boundary F1 results at $\tau = 1$ pixel tolerance for field detection on two Sentinel-2 acquisitions. Both macro (equal-weighted parcels) and micro (area-weighted parcels) variants are reported. Best and second-best values per column are marked in **bold** and <u>underline</u>.

Table 4.7 reports Boundary F1 results at $\tau = 1$ pixel tolerance for both Sentinel-2 acquisitions. On both dates, the transformer models clearly outperform bicubic interpolation. They achieve sharper and more consistent boundaries. While SwinIR and MAT achieve the best performances on both dates, PFT reaches similar scores, underlining the overall stability of the transformer family.

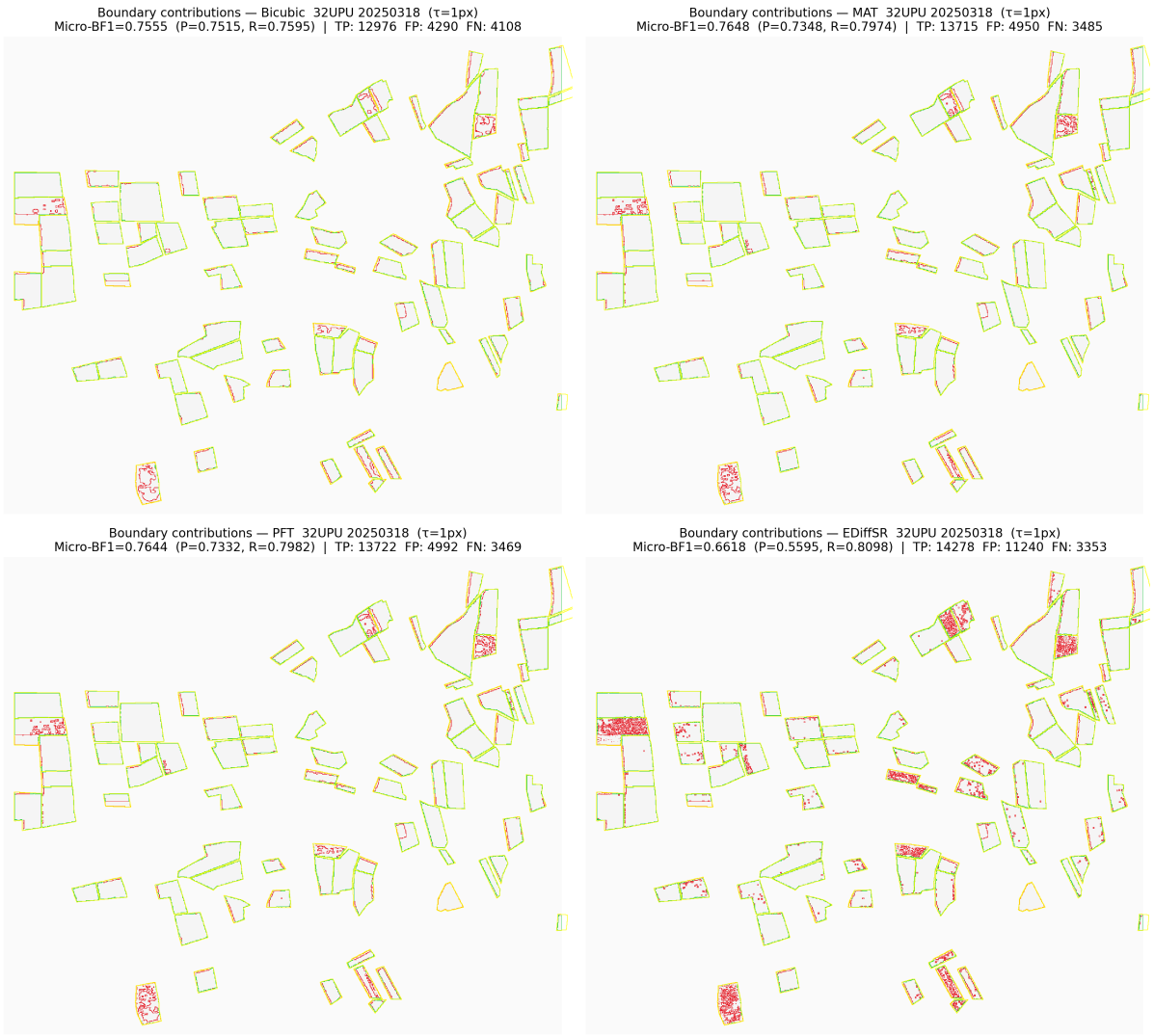

Figure 4.14: Enter Caption

Figure 4.15: Boundary F1 maps at $\tau = 1$ pixel (tile 32UPU, 18 March 2025) for selected models (Bicubic, MAT, PFT, EDiffSR). Green = correct boundaries, red = false positives, orange = false negatives.

In contrast, EDiffSR produces the worst results across all metrics, with especially low peaks for Micro-BF1 values. Looking at the visualizations in Figure 4.15, we can observe the cause of it. EDiffSR produces noisy and unstable edges, with boundary fragments scattered within the parcels. Bicubic also achieves worse scores than the transformer family, as it produces overly smoothed outlines, which BF1 ($\tau = 1$) punishes.

Overall, the $\tau = 1$ evaluation confirms that super-resolution preprocessing improves parcel boundary detection. Unlike AUROC, Boundary F1 captures the sharper outlines produced by the transformer models and exposes the noisy artifacts of EDiffSR, providing a more faithful assessment of super-resolution quality.

The Results for $\tau = 2$

Model	08/03/2025 (105 parcels)		18/03/2025 (81 parcels)	
	Macro-BF1 ↑	Micro-BF1 ↑	Macro-BF1 ↑	Micro-BF1 ↑
Bicubic	0.9068	**0.9039**	0.8706	0.8666
SwinIR	<u>0.9125</u>	<u>0.8994</u>	0.8855	0.8662
MAT	0.9120	0.8989	**0.8863**	**0.8673**
PFT	**0.9126**	<u>0.8994</u>	<u>0.8858</u>	<u>0.8667</u>
EDiffSR	0.8668	0.8059	0.8177	0.7696

Table 4.8: Boundary F1 results at $\tau = 2$ pixel tolerance for field detection on two Sentinel-2 acquisitions. Both macro (equal-weighted parcels) and micro (area-weighted parcels) variants are reported. Best and second-best values per column are marked in **bold** and <u>underline</u>.

Table 4.8 summarizes Boundary F1 results at a tolerance of $\tau = 2$ pixels. Compared to the stricter $\tau = 1$ results, all models improved significantly due to the higher pixel tolerance. Small boundary shifts and smoother edges were now recognized as positives, which strongly benefits bicubic interpolation. On both dates, bicubic catches up to the transformer models and even achieves the highest micro-BF1 on the first tile.

The transformer models (SwinIR, MAT, PFT) stay consistently strong, with similar scores and only marginally above bicubic. While MAT and PFT slightly lead on the second date, SwinIR outperforms them on the first, achieving the best macro score. Even though EdiffSR scores improve relative to $\tau = 1$, it still reaches the worst results and also shows the same signs of noise and instability in its boundaries.

This can be clearly observed in Figure 4.16. Furthermore, bicubic's oversmoothed edges now align more within the tolerance. Overall, increasing τ reduces the relative advantage of super-resolution, as the metric begins to favor blurred outlines. This highlights that the stricter $\tau = 1$ setting better reflects the improvements achieved by super-resolution.

In Conclusion, the Boundary F1 experiments demonstrate that super-resolution, using transformer-based models, consistently improves field boundary detection over bicubic

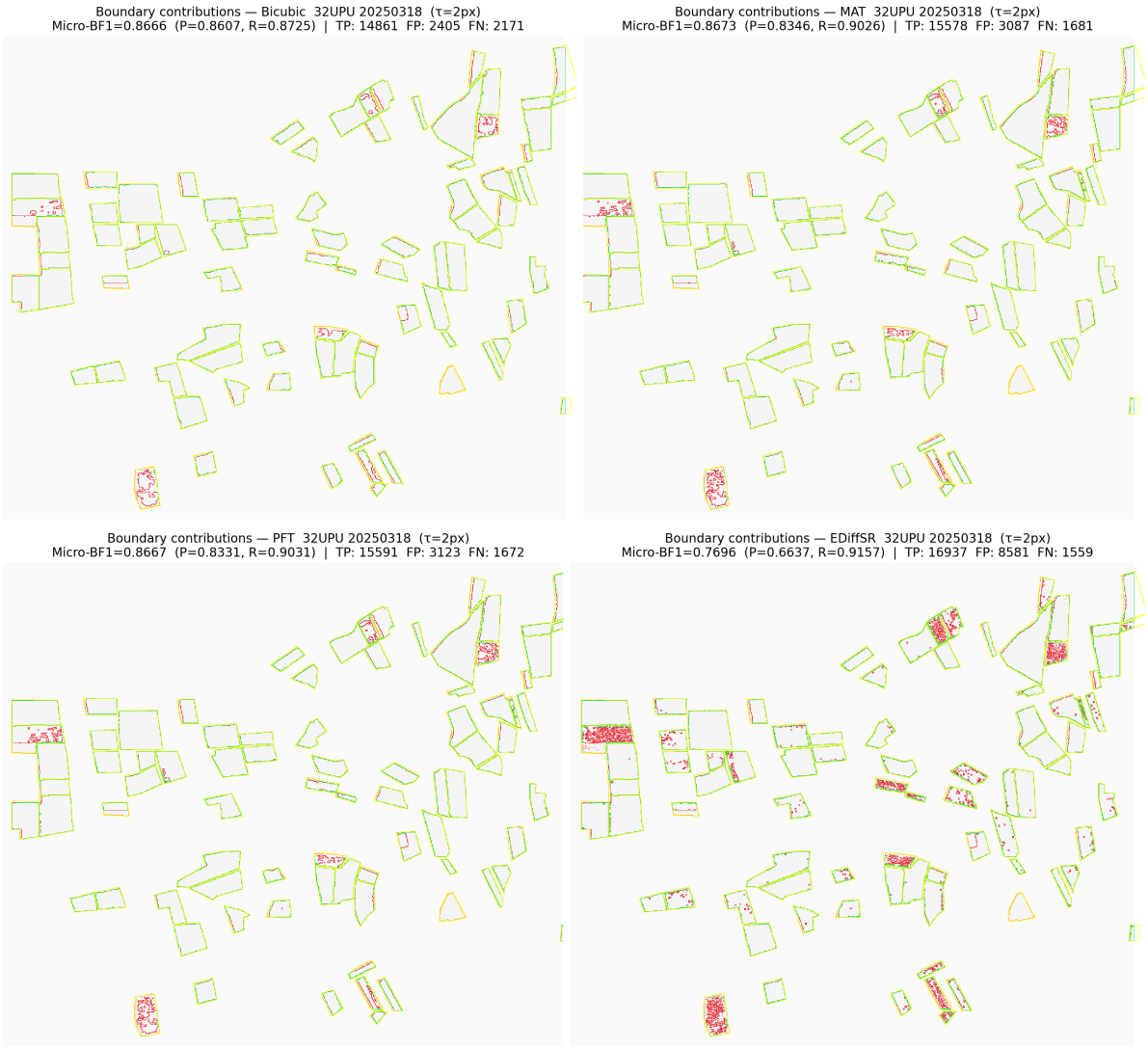

Figure 4.16: Boundary F1 maps at $\tau = 2$ pixel (tile 32UPU, 18 March 2025) for selected models (Bicubic, MAT, PFT, EDiffSR). Green = correct boundaries, red = false positives, orange = false negatives.

interpolation, providing sharper and more reliable parcel outlines.

5 Analysis

While the previous chapter presented the experimental results in detail, this chapter takes a step further by interpreting and contextualizing them. Here, we don't want to reiterate exact numbers, but to identify overarching trends across all experiments. Our focus is on asking why the models behaved as they did and what it really means for the remote-sensing space. First, we will go deeper on model performances across all our experiments, highlighting interesting trends and giving a simple guide to picking the correct super-resolution model for your applications. Continuing with reflecting on what makes super-resolution a valuable preprocessing step for downstream applications, using our field-detection downstream experiments as a case study of the do's and don'ts. Finally, we will highlight the limitations of our study, including methodological compromises and open challenges for future work.

5.1 Analysis of Model Performance

This section provides a comprehensive overview of the super-resolution models' performances across various experiments, domains, and training configurations. We will highlight their robustness across spatial domains, their performances relative to training datasets, and we finish with a practical guide to pick the right model for your use cases.

5.1.1 Transformer Family vs. Diffusion-Based EDiffSR

Across all our experiments, we observed the trend of separating the results between EDiffSR and our transformer models: SwinIR, MAT, and PFT. In this section, we highlight the strengths and limitations of both architectures relative to the remote sensing domain.

As expected, we saw a performance gap between the transformer family and EDiffSR on the PSNR and SSIM experiments (Section 4.1.2). These metrics reward pixel-level fidelity and penalize small misalignments harshly. Here, we can highlight the first characteristic of transformers: they remain conservative, producing stable imagery that closely follows the low-resolution input while sharpening its structures. In contrast, EDiffSR tries to generate sharper textures and increase contrast. It takes more risks, predicting details that were not there before, but look perceptually closer to high-quality images.

This is why the diffusion architecture scores well on FID, NIQE (pretrained), our simple artifact detector, and our custom-trained NIQE models. For both the synthetic and native 20m $\rightarrow$ 10m experiments (Section 4.3 and Section 4.4), it outperforms all our models significantly across all three metrics. Only on the cross dataset test (Section 4.5), the ranking for NIQE (custom) and FID flip. The reason here is the domain gap that all our models were not trained on, going from a sentinel sensor to VENµS imagery. Transformer models have the edge

here, as they focus more on preserving simple shapes, instead of trying to predict sensor characteristic patterns. EDiffSR generates domain-specific textures that fail to generalize to VENµS. This also highlights the limitation of the cross-dataset evaluation itself: the VENµS ground truth does not perfectly represent what 5m Sentinel imagery would look like.

Visual inspection of our comparison imagery reveals additional insights: Although we can notice some noise in uniform areas in our other graphical comparisons (Figure 4.2), it is particularly prevalent in the cross-dataset experiment (Figure 4.9). This noise can be a concerning issue for downstream tasks, which we first noticed in our comparison against bicubic for our field soil masks (Section 4.11). The extra local contrast introduced by EDiffSR led to many artifacts in the final mask and presumably also hurt the final field boundary prediction task (Section 4.3.7). Again, visual inspection brought extra insights (see Figure 4.15), some boundaries appeared fragmented with no uniform lines, inconsistent with the LQ image. This propagated into the scoring, being outperformed by bicubic interpolation.

Inference time is another critical factor for the practicality of a model. While EDiffSR is already an efficient diffusion model, it remains significantly slower than transformers due to its iterative processing. Especially in the remote sensing space, where billions of square kilometers of imagery can require processing, efficiency can be the deciding factor.

The results suggest one fundamental distinction: Transformers can be characterized as careful, faster and more trustworthy, prioritizing consistency over perceptual appearance. They focus on preserving the general shapes and lines of bicubic, but making them sharper and more distinct. Diffusion is confident, creating more complex patterns that look realistic to the human eye and perceptual metrics. But they are prone to hallucinating details or patterns that should not be added: They trade trustworthiness for visual appeal. This is well-suited for domains where perceptual realism is prioritized over fidelity and can feel magical, like super-resolution applied to our phone cameras. But for the rigorous scientific remote sensing space, trustworthiness is crucial. Satellite images are measurements of reality and do not need to look perceptually pleasing. This is where transformer models are a safer and more reliable option.

5.1.2 Transformer Models Compared: SwinIR, MAT, and PFT

Across all experiments, the three transformer models behaved similarly, with only small variations in performance. PFT and MAT were indistinguishable in many experiments: their images appeared nearly identical, and their scores on FID and NIQE were extremely close, with no consistent pattern of one outperforming the other. PFT might have a slight edge over MAT in terms of PSNR and SSIM, but even then there are fluctuations in ranking order. SwinIR more consistently trails by a small margin, yet it also achieves top results in some experiments. Looking at visualizations (see Figure 4.2), it becomes easier to distinguish SwinIR from its competitors: MAT and PFT produce sharper results, while SwinIR is marginally softer, yet way sharper and more stable than bicubic. Importantly, SwinIR is substantially faster than PFT, which has a mean latency 6× higher (Section 4.1.1). However, MAT delivers the best performance-to-efficiency ratio, while being only slightly more computationally expensive than SwinIR.

5.1.3 Comparison of Training Approaches: VENµS Trained vs. VENµS Fine-Tuned

One of the early decisions we made for this thesis was to train the models on three different data set mixes, with two focusing on the 10m → 5m domain. In this section, we compare the performance of both approaches, utilizing VENµS for training and finetuning.

Comparing their performance on their natural (10m → 5m) VENµS domain leads to varied results with no clear winner. Some models like PFT finetuned outperform their trained counterparts, while for MAT, it is exactly the other way around. The margins between the different models are tiny, and they can be due to training run-specific factors, such as initialization weights, rather than a better training method. This also shows that the VENµS trainings capture the domain pretty well, not leaving much room for improvement to the finetuned models.

When comparing the results of the models on the Sentinel-2 (20m → 10m) spatial gap, an unexpected clear trend occurs. VENµS trained models outperform their finetuned counterparts across all 20m → 10m experiments. This behavior also stays consistent across the NIQE (custom) experiments. This indicates that fine-tuning, while starting from Sentinel-trained weights, does not preserve cross-domain generalization. The lower learning rate of finetuning leads the model to adapt more cautiously and narrowly to VENµS characteristics. Sentinel-2 pretraining likely shortened the training path toward the VENµS optimum, yet any Sentinel-specific bias was lost, with the models ultimately optimized entirely for VENµS. This leads us to think that VENµS trained models have stronger capabilities to generalize across domains. This hypothesis was strengthened by the cross-dataset NIQE (custom) experiment (Section 4.5), which is part of the 10m → 5m domain and tries to test model robustness.

5.1.4 A Practical Guide to Choosing the Right Super-Resolution Model

This subsection summarizes the findings of the previous analyses and provides a practical perspective on selecting the most suitable model based on resources, priorities, and application goals.

The first question to ask is: How much data do you need to process and what resources are available? If the task needs processing of very large amounts of data, diffusion-based approaches such as EDiffSR can be ruled out for most applications. However, when computational efficiency is not the main priority and the end result needs to have the best visual quality, looking realistic and rich with high-fidelity textures, diffusion-based approaches like EDiffSR are perfect for that. If the focus switches to a more trustworthy scientific approach, transformers are a safer choice. Once again, the next step in choosing the right method is to ask the performance question. Is a method with high throughput and speed the priority? Then the best pick out of our options is SwinIR. Realistically, looking into the lights options from MAT or PFT might be a more optimal step, as they will provide better performance for their speed. If even those methods are too slow, CNN-based methods are the next domain to explore. For the best trade-off between performance and speed, MAT is our best pick: It provides state-of-the-art performance across all our experiments while retaining fast inference speeds. This makes it the default choice for most practical projects. If squeezing out maximum

performance is the goal, PFT can marginally surpass MAT on fidelity metrics. However, for most projects, the extra performance cost will not warrant this slight improvement.

This overview highlights the trade-offs between architectures and offers concrete guidance on when each model can be most effectively applied.

5.2 What Makes Super-Resolution a Valuable Preprocessing Step for Downstream Tasks?

One of the central objectives of this thesis was to investigate whether super-resolution can provide benefit for downstream applications. To this end, we tested the performance of our super-resolution models on a field boundary detection task, comparing their performance to bicubic interpolation. The results provided us with insights into when and why super-resolution can be a valuable preprocessing step, as well as where its limitations lie.

Our first AUROC (Section 4.3.6) experiment led to concerning results: bicubic outperformed our super-resolution methods. However, on further inspection, we found the real cause of the issue: The failure was not caused by the models themselves, but by the design of the experiment. We were using a simple per-parcel field mask as ground truth, which was generated over many PlanetScope [94] acquisitions. This led to it being a uniform mask, which did not correctly represent the real world for our test. While our super-resolution techniques were sharpening natural deviations, such as spots in fields that were not uniform, bicubic was rewarded for averaging out these spots. The super-resolution models have most likely been closer to reality, but our AUROC setup failed to capture this. This highlights an important point: When evaluating downstream task performance, we need a ground truth that is fine-grained enough to assess its benefits correctly.

When we thought deeper about what our ground truth actually captures, we switched to evaluating with the boundary-focused F1 score. Our ground truth contained high-quality field edges, which allowed us to test the real performance of our super-resolution models. The result: a great performance by our transformer family, clearly outperforming bicubic. Importantly, the test played into the strength of super-resolution, it encourages sharp edges over uncertain blurry borders. When we increased our tolerance for edge misalignments, we once again observed bicubic interpolation catching up to our techniques, as sharp edges were not as necessary anymore.

These findings have now raised the question: For what downstream tasks can super-resolution improve performance in remote sensing? The obvious answer we came up with is: Tasks that benefit from higher resolution satellite data and are built to handle it. Especially tasks that depend on spatial details, such as land-cover classification, urban mapping, and object detection, can be great examples. However, it can be more nuanced. Thinking of an object detection algorithm that was previously trained on the original super-resolution data might actually perform worse with higher spatial resolutions. The entire downstream application of it was built, with its spatial limitations in mind. To achieve better performance on a downstream task after applying super resolution, requires compatibility of the following algorithms with higher resolution data. There are also lots of examples where the gain of

super-resolution might be negligible. Applications that average image information over large quantities can statistically remove the gains of super-resolution entirely. Finally, we can't forget about hallucinations and artifacts generated by super-resolution. There is always room for error when working with algorithms that predict uncertain elements. This is why it is crucial to thoroughly benchmark models for trustworthiness before applying them to critical real-world applications. It is also important to manage expectations and to stay critical: Super resolution will not magically bring your 10m Sentinel data to the subpixel level, at least not without heavy hallucinating.

However, for applications that work better with higher-resolution satellite data, super-resolution is a valuable extra step to take. And we should keep in mind: With the rapid research and the constantly improving techniques, this is the worst super-resolution will ever be.

5.3 Limitations of the Study

In this section, we discuss the main limitations of our study, including potential areas for improvement and methodological compromises we had to make.

The most relevant limitation of this thesis was the reliance on a simple bicubic degradation kernel for generating our low-resolution training image pairs. While bicubic is widely used for this purpose in super-resolution literature, this simplified approach does not accurately capture the complexity of real-world sensors and can hurt model performance on native upscaling tasks. For models used in real-world applications, a more complex degradation model with blur, aliasing, noise, or compression artifacts should be prioritized [24]. For the limited duration of this thesis, it was out of scope.

Another limitation affecting our two datasets is the difference in their normalization strategies. While the Sentinel-2 dataset percentile-clipping normalization relied on full satellite tiles, we didn't have access to the full tiles for VENµS, forcing us to adapt using a min-max per patch normalization. For a detailed explanation of the different normalization methods and our approach, look into Section 3.1.3. While we considered it an interesting experiment, how models will generalize across this normalization gap, utilizing the same normalization strategy, could be another factor to maximize super-resolution model performance.

Our training setup was also relatively simple, working with the default hyperparameters of the models. To achieve maximum performance on our special single-band data, ablation studies could be conducted to find the optimal hyperparameters. Additionally, we worked with a conservative batch size of two, due to training issues with higher batch sizes using PFT. This could also be a factor that was limiting the stability of our EDiffSR training. For more information about the EDiffSR training problems, read Section 3.4.1. Exploring different loss functions (other than L1) would be another suggestion for future work.

Now, let us dive into the limitations of the evaluation pipeline: For our 16-bit single-band data, we were forced to adapt metrics such as FID and NIQE to match our domain. Although we trained custom NIQE models for our experiments (Section 4.2), FID was still reliant on the InceptionV3 [84] feature extractor, which was originally trained for ground data. Retraining

was out of scope for this thesis and is uncommon in the remote sensing space.

The robustness of the downstream task could also be improved. Due to the ground truth parcel mask, we were limited to one location and utilized two Sentinel-2 tiles with different acquisition dates. Creating more ground-truth masks for diverse fields around the world enables a more comprehensive evaluation of model generalizability and improves statistical significance.

6 Conclusion

6.1 Summary

One of the major factors holding research back in the remote sensing space is the limited spatial resolution of open and free satellite imagery, such as the data collected by Sentinel-2. A primary objective of this thesis was to explore whether super-resolution can serve as a preprocessing step to address this problem, and to identify which models are best suited for it.

To this end, we adapted four super-resolution models, containing three transformers: SwinIR (Section 3.2.2), MAT (Section 3.2.3), PFT (Section 3.2.4), and one diffusion architecture EDiffSR (Section 3.2.5), to work on single-band 16-bit satellite data. These models were trained in three different configurations on our synthetic Sentinel-2 (Section 3.1.1)(20m $\rightarrow$ 10m) and VENµS (Section 3.1.2) (10m $\rightarrow$ 5m) datasets. With the goal of finding the strengths and weaknesses of each model, we assessed them thoroughly on our comprehensive evaluation pipeline. Testing them on simple quantitative metrics like PSNR and SSIM (Section 4.1), and on FID and custom-trained NIQE models for native evaluation across different spatial gaps (Section 4.2). Most importantly, we set up a field boundary detection downstream task to track their real-world practicability (Section 4.3).

Although our diffusion approach (EDiffSR) performed well on the perceptual metrics (FID, NIQE) and visual inspections, it couldn't compete with the transformer models on PSNR, SSIM, and the downstream application. The transformer models convincingly outperformed bicubic on the field boundary detection downstream task, demonstrating that super-resolution can serve as a practical and valuable preprocessing step for real-world remote sensing tasks.

6.2 Final Thoughts

It is inevitable that super-resolution and image enhancement algorithms will become increasingly more popular in computer vision, and particularly in the remote sensing space. With so many new techniques and innovations emerging, super-resolution approaches will only become easier to train, deploy, and reach new performance heights. Models will learn to adapt to diverse settings, lowering the barrier to entry. The overarching promise of super-resolution improving performance across many downstream applications by simply adding a single preprocessing step will fuel further research in this niche. Super-resolution has the chance to become a versatile tool, applied like a "Swiss army knife" for computer vision.

For remote sensing in particular, mass adoption will depend on two main factors: efficiency and accessibility. In remote sensing, it is common to work with massive amounts of imagery,

making faster inference time a necessity. But accessibility is equally important: Applying super-resolution needs to get way easier, without requiring setting up complex codebases. The goal should be that baseline models get integrated into popular computer vision libraries, working as simply as selecting a different method than bicubic for changing image sizes. But also projects like BasicSR [76] lower the barrier to entry for new researchers, providing a comprehensive toolkit with the most essential features for training and evaluating models. With more and more models being built on such a modular codebase, it will become easier to switch out models, metrics, and loss functions, further accelerating research progress.

Finally, the development of meaningful benchmarks will play a central role. Similar to our evaluation pipeline, they need to rely on many different metrics and evaluation approaches. Assessing model performance on real-world downstream applications should be the gold standard for benchmarking models, directly showcasing their real-world practicability. Building comprehensive benchmarks will encourage the process of developing meaningful models that do not focus on maxing out one metric, but on real-world utility.

6.3 Future Works

This thesis provides a foundation for multiple directions of future research and development in satellite image super-resolution. One of the main limitations of our study was the simple bicubic downsampling operation used for our synthetic data sets. With the goal of making our models perform stronger on real-world applications, a natural next step would be building a realistic degradation model with sensor blur, aliasing, noise, and compression artifacts.

Another avenue would be utilizing our comprehensive super-resolution evaluation pipeline and testing various new models and techniques. A meaningful focus could be investigating more performant models, starting with the already adapted light versions of MAT and PFT, which look very promising. Additionally, comparing them to more traditional CNN architectures could provide insight into the usability of transformers in large-scale remote sensing operations.

At the same time, the evaluation pipeline still leaves room for plenty of improvements, such as working with more Sentinel tiles in the downstream application, creating higher-quality ground truth images, and adding additional downstream application tests. Following this direction, the pipeline could be built as a comprehensive real-world benchmark for new super-resolution techniques, testing the performance of models on a variety of real-world applications. To reach this goal, the codebase would need to be streamlined so that one command executes all tests and provides a detailed model performance overview.

If the goal is to further improve our models or develop novel methods, testing our different loss functions could be another interesting pathway. Although our experiments relied on a standard L1 loss, a radiometric loss [60] could prove beneficial, as it would help preserve the physical consistency of surface reflectance in the super-resolved images.

Overall, these directions show that there is plenty of room to build on the work started here. Remote-sensing super-resolution is still a young field, with rapid innovations on the horizon that could bring it closer to reliable real-world use.

List of Figures

List of Tables

Bibliography

[1] Nova Space, *Earth observation satellites set to triple over the next decade*, Press Release, [Online; accessed 6-September-2025], 2023. [Online]. Available: `https://nova.space/press-release/earth-observation-satellites-set-to-triple-over-the-next-decade/`.

[2] M. Drusch, U. Del Bello, S. Carlier, O. Colin, V. Fernandez, F. Gascon, B. Hoersch, C. Isola, P. Laberinti, P. Martimort, *et al.*, "Sentinel-2: Esa's optical high-resolution mission for gmes operational services", *Remote sensing of Environment*, vol. 120, pp. 25–36, 2012.

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

[4] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models", *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[5] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "Swinir: Image restoration using swin transformer", in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 1833–1844.

[6] C. Xie, X. Zhang, L. Li, Y. Fu, B. Gong, T. Li, and K. Zhang, "Mat: Multi-range attention transformer for efficient image super-resolution", *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

[7] W. Long, X. Zhou, L. Zhang, and S. Gu, "Progressive focused transformer for single image super-resolution", in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 2279–2288.

[8] Y. Xiao, Q. Yuan, K. Jiang, J. He, X. Jin, and L. Zhang, "Ediffsr: An efficient diffusion probabilistic model for remote sensing image super-resolution", *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–14, 2023.

[9] R. C. Gonzalez, *Digital image processing*. Pearson education india, 2009.

[10] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity", *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[11] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium", *Advances in neural information processing systems*, vol. 30, 2017.

[12] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "completely blind" image quality analyzer", *IEEE Signal processing letters*, vol. 20, no. 3, pp. 209–212, 2012.

[13] P. Lab, *Components of the electromagnetic spectrum*, Blog Post, [Online; accessed 6-September-2025], 2023. [Online]. Available: `https://www.primalucelab.com/blog/components-of-electromagnetic-spectrum/`.

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

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

[16] J. Kim, J. K. Lee, and 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.

[17] J. Kim, J. K. Lee, and K. M. Lee, "Deeply-recursive convolutional network for image super-resolution", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1637–1645.

[18] 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.

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

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

[21] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2472–2481.

[22] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks", in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 286–301.

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

[24] X. Wang, L. Xie, C. Dong, and Y. Shan, "Real-esrgan: Training real-world blind super-resolution with pure synthetic data", in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 1905–1914.

[25] S. K. Jangir, "Aerial and satellite image enhancement with super resolution using deep learning", This master thesis was supervised by Seyed Majid Azimi and Dr. Reza Bahmanyar., M.S. thesis, Technical University of Munich, 2020. [Online]. Available: https://elib.dlr.de/138151/.

[26] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.*, "Language models are few-shot learners", *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[27] D. Dutta, D. Chetia, N. Sonowal, and S. K. Kalita, "State-of-the-art transformer models for image super-resolution: Techniques, challenges, and applications", *arXiv preprint arXiv:2501.07855*, 2025.

[28] G. Kalra, *Attention networks: A simple way to understand self-attention*, https://medium.com/, Accessed: 6-September-2025. Available at: https://medium.com/@geetkal67/attention-networks-a-simple-way-to-understand-self-attention-f5fb363c736d, 2022.

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

[30] Z. Lu, J. Li, H. Liu, C. Huang, L. Zhang, and T. Zeng, "Transformer for single image super-resolution", in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 457–466.

[31] J. Cao, J. Liang, K. Zhang, Y. Li, Y. Zhang, W. Wang, and L. V. Gool, "Reference-based image super-resolution with deformable attention transformer", in *European conference on computer vision*, Springer, 2022, pp. 325–342.

[32] X. Pan, T. Ye, Z. Xia, S. Song, and G. Huang, "Slide-transformer: Hierarchical vision transformer with local self-attention", in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 2082–2091.

[33] Z. Chen, Y. Zhang, J. Gu, L. Kong, X. Yang, and F. Yu, "Dual aggregation transformer for image super-resolution", in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 12 312–12 321.

[34] Q. Zhu, P. Li, and Q. Li, "Attention retractable frequency fusion transformer for image super resolution", in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2023, pp. 1756–1763. DOI: 10.1109/CVPRW59228.2023.00176.

[35] X. Chen, X. Wang, W. Zhang, X. Kong, Y. Qiao, J. Zhou, and C. Dong, "Hat: Hybrid attention transformer for image restoration", *arXiv preprint arXiv:2309.05239*, 2023.

[36] C.-C. Hsu, C.-M. Lee, and Y.-S. Chou, "Drct: Saving image super-resolution away from information bottleneck", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 6133–6142.

[37] G. Li, Z. Cui, M. Li, Y. Han, and T. Li, "Multi-attention fusion transformer for single-image super-resolution", *Scientific Reports*, vol. 14, no. 1, p. 10 222, 2024.

[38] K. Cheng, L. Yu, Z. Tu, X. He, L. Chen, Y. Guo, M. Zhu, N. Wang, X. Gao, and J. Hu, "Effective diffusion transformer architecture for image super-resolution", in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, 2025, pp. 2455–2463.

[39] B. B. Moser, A. S. Shanbhag, F. Raue, S. Frolov, S. Palacio, and A. Dengel, "Diffusion models, image super-resolution, and everything: A survey", *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

[40] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and 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. 10 684–10 695.

[41] Midjourney, Inc., *Midjourney (version 6)*, `https://www.midjourney.com`, 2025.

[42] C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, and M. Norouzi, "Image super-resolution via iterative refinement", *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 4, pp. 4713–4726, 2022.

[43] H. Li, Y. Yang, M. Chang, S. Chen, H. Feng, Z. Xu, Q. Li, and Y. Chen, "Srdiff: Single image super-resolution with diffusion probabilistic models", *Neurocomputing*, vol. 479, pp. 47–59, 2022.

[44] B. Kawar, M. Elad, S. Ermon, and J. Song, "Denoising diffusion restoration models", *Advances in neural information processing systems*, vol. 35, pp. 23 593–23 606, 2022.

[45] Y. Wang, J. Yu, and J. Zhang, "Zero-shot image restoration using denoising diffusion null-space model", *arXiv preprint arXiv:2212.00490*, 2022.

[46] Z. Yue, J. Wang, and C. C. Loy, "Resshift: Efficient diffusion model for image super-resolution by residual shifting", *Advances in Neural Information Processing Systems*, vol. 36, pp. 13 294–13 307, 2023.

[47] Y. Wang, W. Yang, X. Chen, Y. Wang, L. Guo, L.-P. Chau, Z. Liu, Y. Qiao, A. C. Kot, and B. Wen, "Sinsr: Diffusion-based image super-resolution in a single step", in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 25 796–25 805.

[48] J. Wang, Z. Yue, S. Zhou, K. C. Chan, and C. C. Loy, "Exploiting diffusion prior for real-world image super-resolution", *International Journal of Computer Vision*, vol. 132, no. 12, pp. 5929–5949, 2024.

[49] Z. Chen, Y. Zhang, D. Liu, J. Gu, L. Kong, X. Yuan, *et al.*, "Hierarchical integration diffusion model for realistic image deblurring", *Advances in neural information processing systems*, vol. 36, 2024.

[50] Z. Luo, F. K. Gustafsson, Z. Zhao, J. Sjölund, and T. B. Schön, "Refusion: Enabling large-size realistic image restoration with latent-space diffusion models", in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 1680–1691.

[51] Z. Chen, H. Qin, Y. Guo, X. Su, X. Yuan, L. Kong, and Y. Zhang, "Binarized diffusion model for image super-resolution", *Advances in Neural Information Processing Systems*, vol. 37, pp. 30 651–30 669, 2024.

[52] Y. Liu, J. Yue, S. Xia, P. Ghamisi, W. Xie, and L. Fang, "Diffusion models meet remote sensing: Principles, methods, and perspectives", *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[53] Y. Qi, M. Lou, Y. Liu, L. Li, Z. Yang, and W. Nie, "Advancing image super-resolution techniques in remote sensing: A comprehensive survey", *arXiv preprint arXiv:2505.23248*, 2025.

[54] J. Michel, J. Vinasco-Salinas, J. Inglada, and O. Hagolle, "Sen2venμs, a dataset for the training of sentinel-2 super-resolution algorithms", *Data*, vol. 7, no. 7, p. 96, 2022.

[55] J. Cornebise, I. Oršolić, and F. Kalaitzis, "Open high-resolution satellite imagery: The worldstrat dataset–with application to super-resolution", *Advances in Neural Information Processing Systems*, vol. 35, pp. 25 979–25 991, 2022.

[56] M. Märtens, D. Izzo, A. Krzic, and D. Cox, "Super-resolution of proba-v images using convolutional neural networks", *Astrodynamics*, vol. 3, no. 4, pp. 387–402, 2019.

[57] N. L. Nguyen, J. Anger, A. Davy, P. Arias, and G. Facciolo, "Proba-v-ref: Repurposing the proba-v challenge for reference-aware super resolution", in *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, IEEE, 2021, pp. 3881–3884.

[58] P. Kowaleczko, T. Tarasiewicz, M. Ziaja, D. Kostrzewa, J. Nalepa, P. Rokita, and M. Kawulok, "A real-world benchmark for sentinel-2 multi-image super-resolution", *Scientific Data*, vol. 10, no. 1, p. 644, 2023.

[59] C. Aybar, D. Montero, J. Contreras, S. Donike, F. Kalaitzis, and L. Gómez-Chova, "Sen2naip: A large-scale dataset for sentinel-2 image super-resolution", *Scientific Data*, vol. 11, no. 1, p. 1389, 2024.

[60] C. Aybar, D. Montero, S. Donike, F. Kalaitzis, and L. Gómez-Chova, "A comprehensive benchmark for optical remote sensing image super-resolution", *IEEE Geoscience and Remote Sensing Letters*, vol. 21, pp. 1–5, 2024. DOI: 10.1109/LGRS.2024.3401394.

[61] M. Deudon, A. Kalaitzis, I. Goytom, M. R. Arefin, Z. Lin, K. Sankaran, V. Michalski, S. E. Kahou, J. Cornebise, and Y. Bengio, "Highres-net: Recursive fusion for multi-frame super-resolution of satellite imagery", *arXiv preprint arXiv:2002.06460*, 2020.

[62] Y. Xiao, Q. Yuan, K. Jiang, J. He, C.-W. Lin, and L. Zhang, "Ttst: A top-k token selective transformer for remote sensing image super-resolution", *IEEE Transactions on Image Processing*, vol. 33, pp. 738–752, 2024.

[63] B. Chen, K. Chen, M. Yang, Z. Zou, and Z. Shi, "Heterogeneous mixture of experts for remote sensing image super-resolution", *IEEE Geoscience and Remote Sensing Letters*, 2025.

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

[65] Z. Luo, B. Song, and L. Shen, "Satdiffmoe: A mixture of estimation method for satellite image super-resolution with latent diffusion models", *arXiv preprint arXiv:2406.10225*, 2024.

[66] S. Donike, C. Aybar, L. Gómez-Chova, and F. Kalaitzis, "Trustworthy super-resolution of multispectral sentinel-2 imagery with latent diffusion", *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 18, pp. 6940–6952, 2025. DOI: 10.1109/JSTARS.2025.3542220.

[67] Deutsches Copernicus-Büro, *Sentinel-2*, Accessed: 2025-09-02, 2025. [Online]. Available: https://www.d-copernicus.de/daten/satelliten/satelliten-details/news/sentinel-2/.

[68] C. Gualersi, *Growing impact of copernicus sentinel data revealed*, Accessed: 2025-09-02, 2022. [Online]. Available: https://sentinels.copernicus.eu/-/growing-impact-of-copernicus-sentinel-data-revealed.

[69] European Union Copernicus, *Sentinel-2 data collection*, Accessed: 2025-09-02, 2025. [Online]. Available: https://dataspace.copernicus.eu/data-collections/copernicus-sentinel-data/sentinel-2.

[70] Freie Universität Berlin, *Sentinel-2*, Accessed: 2025-09-02, 2025. [Online]. Available: https://blogs.fu-berlin.de/reseda/sentinel-2/.

[71] European Union Copernicus, *Copernicus browser*, Accessed: 2025-09-02, 2025. [Online]. Available: https://browser.dataspace.copernicus.eu/.

[72] Pillow Contributors, *Pil.image module — pillow (pil fork) documentation*, Accessed: 2025-09-02, 2025. [Online]. Available: https://pillow.readthedocs.io/en/stable/reference/Image.html.

[73] ESA/European Space Agency, *Venμs – vegetation and environment monitoring on a new micro-satellite*, Accessed: 2025-09-02, 2024. [Online]. Available: https://www.eoportal.org/satellite-missions/venus#ven%C2%B5s-vegetation-and-environment-monitoring-on-a-new-microsatellite.

[74] N. O. Kadunc, D. Peressutti, N. Vesel, M. Batič, S. Verbič, Ž. Lukšič, and M. Aleksandrov, *How to normalize satellite images for deep learning*, Accessed: 2025-09-02, 2022. [Online]. Available: https://medium.com/sentinel-hub/how-to-normalize-satellite-images-for-deep-learning-d5b668c885af.

[75] E. Agustsson and R. Timofte, "Ntire 2017 challenge on single image super-resolution: Dataset and study", in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, Jul. 2017.

[76] X. Wang, L. Xie, K. Yu, K. C. Chan, C. C. Loy, and C. Dong, *BasicSR: Open source image and video restoration toolbox*, https://github.com/XPixelGroup/BasicSR, 2022.

[77] S. Gillies *et al.*, *Rasterio: Geospatial raster i/o for Python programmers*, Mapbox, 2013. [Online]. Available: https://github.com/rasterio/rasterio.

[78] GDAL/OGR contributors, *GDAL/OGR geospatial data abstraction software library*, Open Source Geospatial Foundation, 2025. DOI: 10.5281/zenodo.5884351. [Online]. Available: https://gdal.org.

[79] German Aerospace Center (DLR), *Terrabyte high-performance cluster*, https://terrabyte.dlr.de, Accessed: 2025-09-02.

[80] C. Xie, X. Zhang, L. Li, Y. Fu, B. Gong, T. Li, and K. Zhang, *Mat: Multi-range attention transformer for efficient image super-resolution – code repository*, https://github.com/stella-von/MAT, GitHub repository, accessed: 2025-09-02, 2025.

[81] W. Long, X. Zhou, L. Zhang, and S. Gu, *Progressive focused transformer for single image super-resolution – code repository*, https://github.com/LabShuHangGU/PFT-SR, GitHub repository, accessed: 2025-09-02, 2025.

[82] Z. Luo, F. K. Gustafsson, Z. Zhao, J. Sjölund, and T. B. Schön, "Image restoration with mean-reverting stochastic differential equations", *arXiv preprint arXiv:2301.11699*, 2023.

[83] Y. Xiao, Q. Yuan, K. Jiang, J. He, X. Jin, and L. Zhang, *Ediffsr: An efficient diffusion probabilistic model for remote sensing image super-resolution – code repository*, https://github.com/XY-boy/EDiffSR, GitHub repository, accessed: 2025-09-02, 2024.

[84] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[85] nuniniyujin, *Niqe for iqa in python (modified fork) – code repository*, https://github.com/nuniniyujin/niqe, GitHub repository (fork), accessed: 2025-09-02, 2025.

[86] P. Gupta, *Niqe for iqa in python – code repository*, https://github.com/guptapraful/niqe, GitHub repository, accessed: 2025-09-02, 2025.

[87] Leibniz Supercomputing Centre (LRZ), *Leibniz supercomputing centre*, https://www.lrz.de, Accessed: 2025-09-02.

[88] German Aerospace Center (DLR), *German aerospace center*, https://www.dlr.de, Accessed: 2025-09-02.

[89] QuantStack and mamba contributors, *Mamba: Fast drop-in replacement for conda*, Documentation, Read the Docs, Available online; accessed: 2025-09-10, 2024. [Online]. Available: https://mamba.readthedocs.io/.

[90] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, *et al.*, "Pytorch: An imperative style, high-performance deep learning library", *Advances in neural information processing systems*, vol. 32, 2019.

[91] T. Fawcett, "An introduction to roc analysis", *Pattern recognition letters*, vol. 27, no. 8, pp. 861–874, 2006.

[92] D. R. Martin, C. C. Fowlkes, and J. Malik, "Learning to detect natural image boundaries using local brightness, color, and texture cues", *IEEE transactions on pattern analysis and machine intelligence*, vol. 26, no. 5, pp. 530–549, 2004.

[93] Planet Labs PBC, *Planet labs: Earth observation company*, `https://www.planet.com`, Accessed: 2025-09-10, 2025.

[94] S. Marta, "Planet imagery product specifications", *Planet Labs: San Francisco, CA, USA*, vol. 91, p. 170, 2018.

[95] DLR – German Aerospace Center, Department of Imaging Spectroscopy, *Department of imaging spectroscopy, institute of remote sensing methods*, `https://www.dlr.de/en/eoc/about-us/remote-sensing-technology-institute/imaging-spectroscopy`, Accessed: 2025-09-10, 2025.

[96] J. W. Rouse, R. H. Haas, J. A. Schell, and D. W. Deering, "Monitoring vegetation systems in the great plains with ERTS", in *Third Earth Resources Technology Satellite-1 Symposium*, NASA SP-351, vol. 1, Washington, D.C.: NASA Goddard Space Flight Center, 1974, pp. 309–317.

[97] C. H. Key and N. C. Benson, "Landscape assessment: Ground measure of severity, the composite burn index; and remote sensing of severity, the normalized burn ratio", in *FIREMON: Fire Effects Monitoring and Inventory System*, ser. General Technical Report RMRS-GTR-164-CD, Ogden, UT, USA: USDA Forest Service, Rocky Mountain Research Station, 2006.

[98] P. Karlshoefer, P. d'Angelo, J. Eberle, and U. Heiden, "Evaluation framework for the generation of continental bare surface reflectance composites", *Geoderma*, vol. 459, p. 117 340, 2025.

[99] U. Heiden, P. d'Angelo, P. Schwind, P. Karlshöfer, R. Müller, S. Zepp, M. Wiesmeier, and P. Reinartz, "Soil reflectance composites—improved thresholding and performance evaluation", *Remote Sensing*, vol. 14, no. 18, p. 4526, 2022.

[100] D. Zanaga, R. Van De Kerchove, D. Daems, W. De Keersmaecker, C. Brockmann, G. Kirches, J. Wevers, O. Cartus, M. Santoro, S. Fritz, *et al.*, *ESA WorldCover 10 m 2021 v200*, [Online; accessed 10-September-2025], 2022. DOI: `10.5281/zenodo.7254221`. [Online]. Available: `https://doi.org/10.5281/zenodo.7254221`.

[101] M. Main-Knorn, B. Pflug, J. Louis, V. Debaecker, U. Müller-Wilm, and F. Gascon, "Sen2cor for sentinel-2", in *Image and signal processing for remote sensing XXIII*, SPIE, vol. 10427, 2017, pp. 37–48.

[102] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge", *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010.