Evaluating Super-Resolution Models for Real-World Sentinel-2 Applications: A Case Study

Ron Mühlhaus[1,2], Sandeep Kumar Jangir[2], Cecilia Curreli[1,3], Paul Karlshöfer[2], Daniel Cremers[1,3]

[1] Technical University of Munich (TUM)
[2] German Aerospace Center (DLR)
[3] Munich Center for Machine Learning (MCML)

Keywords: Sentinel-2, Super-Resolution, Deep Learning, Satellite Imagery, Remote Sensing, Field Boundary Detection

Abstract

High-resolution Earth observation data are crucial for applications such as agriculture, urban planning, and environmental monitoring. Although commercial satellites provide sub-meter imagery, open-access alternatives like Sentinel-2 are limited to resolutions around 10 m ground sampling distance, which is insufficient for many tasks. In this work, we investigate image super-resolution as a method to bridge this gap, enhancing downstream performance on freely available satellite data. We leverage two 16-bit single-band datasets, consisting of Sentinel-2 (20 m→10 m) and VENμS (10 m→5 m) images, to train and benchmark state-of-the-art SR methods, including transformer- and diffusion-based approaches, across multiple dataset mixes. These models are evaluated quantitatively using reference-based metrics (PSNR, SSIM) using ground-truth and no-reference scores (FID, NIQE) for native upscaling from 20 m→10 m and 10 m→5 m. We observe that different SR architectures present trade-offs between standard quantitative metrics and perceptual image quality. We further assess their impact on a practical downstream task: field boundary detection from Sentinel-2 imagery. Our experiments demonstrate that SR pre-processing improves quantitative fidelity and downstream task performance, enabling low-resolution satellites to compete more effectively with commercial imagery.

1. Introduction

High-resolution satellite imagery is vital for several applications from sustainable agriculture and urban planning to environmental monitoring and disaster response (Ma et al., 2019). While commercial satellites provide sub-meter imagery, their high cost creates a divide, limiting widespread research and application. In contrast, open-access programs like the European Union’s Sentinel-2 (Drusch et al., 2012) provide data for free, but at coarser spatial resolutions (e.g., 10 m, 20 m and 60 m), which are insufficient for many fine-scale tasks. This raises a critical question: How can we bridge the resolution gap to unlock the full potential of open-access satellite data?

Image super-resolution (SR), particularly modern deep learning approaches, presents a promising solution. By learning to reconstruct high-resolution (HR) images from low-resolution (LR) inputs, SR can enhance the effective spatial resolution of open-access data. This work investigates the practical utility of state-of-the-art SR methods, including transformer- and diffusion-based architectures, to enhance satellite imagery. As shown by the authors of (Jangir et al., 2025)—which demonstrated that $\times 2$ SR provides an effective balance between quality enhancement and artifact minimization—we too focus on two operationally relevant scenarios: natively upscaling Sentinel-2 data from 20 m→10 m and 10 m→5 m.

However, applying SR to remote sensing data introduces unique challenges not found in standard computer vision tasks. First, scientific data demands trustworthiness; the SR model must preserve radiometric accuracy and avoid ”hallucinating” artifacts. A perceptually pleasing image is insufficient if it does not reflect reality. Second, satellite data is fundamentally different: it consists of single-band, 16-bit files measuring physical surface reflectance, not 8-bit RGB display values. This requires significant adaptation of existing SR models, metrics, and training pipelines.

This work directly confronts these specific challenges. Unlike 8-bit RGB images, satellite data consists of multiple spectral bands (e.g., NIR, SWIR) capturing 16-bit surface reflectance values. This high dynamic range is critical for scientific analysis and must not be corrupted by the SR process. We, therefore, develop a band-agnostic approach, adapting state-of-the-art SR models to process 16-bit single-band imagery individually and handle patch-based processing of large tiles. This paper details our methodology for adapting these models, metrics, and data pipelines for the 16-bit domain. We then provide a comprehensive benchmark, evaluating models not only on quantitative fidelity (PSNR, SSIM) and perceptual quality (FID, NIQE) but also on their real-world impact on a downstream field boundary detection task, thereby validating SR as a trustworthy pre-processing step.

2. Related Work

Early single-image super-resolution (SISR) methods evolved from simple algorithms like bicubic interpolation to deep learning approaches. The pioneering Super-Resolution Convolutional Neural Network (SRCNN) (Dong et al., 2014) sparked a wave of deeper CNN architectures, such as VDSR, SRResNet, and EDSR (Kim et al., 2016, Ledig et al., 2017, Lim et al., 2017), that achieved high fidelity by optimizing for pixel-wise metrics like PSNR. In parallel, Generative Adversarial Networks (GANs), such as SRGAN (Ledig et al., 2017) and ESRGAN (Wang et al., 2018), introduced perceptual losses to create more realistic and detailed textures, establishing a key trade-off between pixel-level accuracy and perceptual quality.

A paradigm shift occurred with the Transformer (Vaswani et al., 2017), which uses self-attention to model long-range dependencies. While Vision Transformer (ViT) (Dosovitskiy et al., 2020) adapted this for image classification, its global attention mechanism was computationally prohibitive for SR. The SwinIR model

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-3-2026-215-2026

(Liang et al., 2021) provided a critical breakthrough by using efficient, shifted window-based self-attention. This architecture became a foundational baseline, inspiring subsequent models like MAT (Xie et al., 2025) and PFT (Long et al., 2025), which further refined attention mechanisms and feature aggregation to achieve state-of-the-art, fidelity-focused results.

More recently, Denoising Diffusion Probabilistic Models (Ho et al., 2020) have demonstrated unparalleled success in high-fidelity image generation. This was adapted for SR in models like SR3 (Saharia et al., 2022), which learn to reverse a noise-driven degradation process, conditioned on the low-resolution image. Diffusion models excel at generating sharp, perceptually convincing details but typically suffer from high computational costs. This has spurred research into more efficient methods, such as latent diffusion models (Rombach et al., 2022), which operate in a compressed latent space.

The application of these methods to remote sensing has led to a distinct sub-field with unique approaches. On the data front, specialized datasets have been created to provide realistic training and evaluation pairs, which are difficult to obtain due to varying sensor properties and acquisition times. Key examples include SEN2VENμS (Michel et al., 2022) for cross-sensor learning and benchmarks like the Proba-V challenge (Maertens et al., 2019) and WorldStrat (Cornebise et al., 2022) for multi-image super-resolution (MISR). Leveraging high temporal frequency, MISR models like HighRes-net (Deudon et al., 2020) represent a popular RS-specific strategy, fusing multiple LR views of the same scene. Other works focus on leveraging correlations between different spectral bands to enhance a target band. Concurrently, state-of-the-art architectures are being adapted; this includes lightweight diffusion models like EDiffSR (Xiao et al., 2023) and transformer-based models like TTST (Xiao et al., 2024). Some diffusion-based works have specifically focused on reliability, attempting to generate uncertainty maps to gauge model trustworthiness (Donike et al., 2025). However, this prior work has often focused on the complex tasks of multi-image or multispectral fusion. A foundational benchmark comparing the latest transformer and diffusion paradigms on the fundamental 16-bit, single-band data format remains less explored. This leaves a critical gap in understanding the performance trade-offs of these state-of-the-art architectures for scientific applications and their true impact on downstream analytical tasks.

3. Methodology

3.1 Datasets

Our methodology is built around two 16-bit, single-band datasets designed to evaluate SR performance on two operationally relevant tasks: 20 m→10 m and 10 m→5 m. All models are trained to be band-agnostic, treating each spectral band as an individual grayscale channel.

The first dataset uses Sentinel-2 imagery for the 20 m→10 m task. We use the 10 m bands (RGB, NIR) from 10 diverse European region tiles as the high-resolution ground-truth, creating 256×256 patches. The corresponding low-resolution 128×128 patches, simulating 20 m data, are generated using 2x bicubic downsampling. A separate set of 10 global tiles outside Europe is reserved for validation.

The second dataset focuses on the 10 m→5 m task using the VENμS satellite. This dataset is derived from the SEN2VENμS collection (Michel et al., 2022), which provides aligned Sentinel-2 and VENμS image pairs. However, to maintain a consistent single-sensor training methodology, we only utilize the 5 m VENμS imagery for this dataset. This approach was chosen for two key reasons. First, it allows us to test model generalization on a different sensor and a finer resolution gap. Second, the VENμS sensor has a close spectral correspondence with Sentinel-2, making it an ideal candidate for transfer learning experiments. The 5 m HR patches are likewise downsampled by 2x bicubic interpolation to create the 10 m LR pairs.

3.2 The Challenge of 16-bit Normalization

A significant challenge in this work is the normalization of 16-bit satellite data. Standard 8-bit normalization methods are ineffective. For instance, a simple division by the maximum value (65,535) compresses the data, as satellite sensors (e.g., 12-bit) often use only a small, variable portion of the full 16-bit range. A more common approach, tile-wide min-max scaling, suffers heavily from outliers. A few extremely bright pixels (e.g., clouds, sun glint) can skew the tile's statistics, compressing all meaningful surface information into a very narrow range (e.g., [0.1, 0.3]). This dramatically impairs model training, as the L1 loss function would treat meaningful errors and outlier errors disproportionately. To solve this, we employed two distinct linear scaling strategies:

1. For Sentinel-2, we use tile-based percentile clipping (clipping at the 0th and 99.5th percentiles). This effectively removes extreme outliers while preserving a consistent radiometric context across all patches from a single tile.

2. For the VENμS data, we applied a per-patch min-max normalization, scaling each pair based on the HR patch's statistics. This maximizes local contrast at the expense of global radiometric consistency.

Training on these two different normalization schemes also serves as a form of data augmentation, forcing models to become more robust to different input distributions.

3.3 SR Models and Adaptations

We adapt and benchmark four state-of-the-art SR models to represent the two dominant architectures in the field.

For the transformer-based approach, we selected three models: SwinIR (Liang et al., 2021), an influential baseline renowned for its efficient shifted-window attention mechanism; MAT (Xie et al., 2025), a more recent model utilizing a multi-range attention strategy; and PFT (Long et al., 2025), a state-of-the-art model that uses a progressive focusing strategy for high performance.

For the generative approach, we chose EDiffSR (Xiao et al., 2023), an efficient diffusion-based model designed specifically for remote sensing. It was selected for its lighter architecture, which replaces the standard U-Net denoiser, making its training and inference times more comparable to the transformer models.

The primary adaptation for all four models was modifying their architecture from the default 8-bit, 3-channel (RGB) input and output to accept 16-bit, single-channel (grayscale) data. We developed a unified data pipeline using Rasterio (Gillies et al., 2013) to properly load, normalize, and process the 16-bit patches and enable patch-based inference on full satellite tiles.

3.4 Evaluation Methods

Model performance is assessed using two categories of metrics to capture both fidelity and perceptual quality. For reconstruction accuracy against a known ground-truth, we use the reference-based metrics Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) (Wang et al., 2004).

To evaluate the native upscaling capabilities (where no ground-truth exists), we use no-reference metrics. We use the Fréchet Inception Distance (FID) (Heusel et al., 2017) to compare the statistical distribution of generated images to real ones. We also use the Naturalness Image Quality Evaluator (NIQE) (Mittal et al., 2012). Since standard NIQE models are trained on 8-bit natural images, we trained custom NIQE models on our pristine 16-bit satellite data to create a domain-specific ”naturalness” reference.

3.5 Training Strategy

All models were trained using L1 (Mean Absolute Error) loss, as it is standard for fidelity-oriented SR. To provide a comprehensive benchmark, we trained three distinct sets of model configurations: (1) trained purely on the Sentinel-2 (20m $\rightarrow$ 10m) dataset; (2) trained purely on the VENµS (10m $\rightarrow$ 5m) dataset; and (3) pre-trained on Sentinel-2 and fine-tuned on VENµS. This strategy allows us to evaluate in-domain performance, cross-sensor generalization, and the potential benefits of transfer learning from one resolution domain to another.

Table 1. Model parameters (in millions) and inference latency (in milliseconds). Latency was measured on an A100 GPU for a 256x256 image.

Model	Params (M) (↓)	Time (ms) (↓)
SwinIR	11.75	**54.12**
MAT	**9.59**	87.38
PFT	19.62	313.93
EDiffSR	20.40	2163.72

4. Experiments and Results

In this section, we evaluate our models, starting with inference speed and reference-based metrics (PSNR, SSIM). We then assess perceptual quality using no-reference metrics (NIQE, FID) on native upscaling tasks. Finally, we demonstrate the real-world utility of our models on a field boundary detection downstream task.

Computational efficiency is essential for processing large satellite data. We evaluated the inference speed of our four models on an NVIDIA A100 GPU with 80 GB VRAM, measuring the mean forward pass latency over 9,900 images (256x256) with a batch size of one. Table 1 shows the results. The transformer models are significantly faster than the diffusion-based EDiffSR. SwinIR was the fastest, at over 40 times faster than EDiffSR, which is limited by its 100 denoising steps. PFT, our largest transformer, was the slowest of the three, likely due to its progressive focused attention mechanism.

4.1 Reference-base Evaluation

We evaluated reconstruction fidelity using PSNR and SSIM on 10,000-image subsets of our Sentinel (20 m $\rightarrow$ 10 m) and VENµS (10 m $\rightarrow$ 5 m) validation sets. Table 2 shows the performance of all twelve model configurations.

The absolute metric values differ between datasets, likely due to different sensor properties and normalization strategies. On both datasets, the transformer models (PFT and MAT) consistently achieve the highest scores, dominating the benchmarks. SwinIR trails slightly behind. EDiffSR consistently yielded the worst results. This is expected, as pixel-wise metrics like PSNR penalize the small misalignments and generative details common in diffusion models, favoring the smoother reconstructions of transformers. Fine-tuning models on VENµS improved their VENµS scores but degraded their Sentinel performance, as expected.

Table 2. PSNR (↑) / SSIM (↑) results on the Sentinel-2 and VENµS validation sets.

Model (Training)	Sentinel-2 (20m $\rightarrow$ 10m) PSNR (↑)	SSIM (↑)	VENµS (10m $\rightarrow$ 5m) PSNR (↑)	SSIM (↑)
S2-trained				
SwinIR	37.963	0.9415	40.150	0.9569
MAT	37.990	0.9418	39.479	0.9425
PFT	**38.012**	**0.9420**	40.217	0.9572
EDiffSR	34.075	0.8825	36.121	0.9377
VENµS-trained				
SwinIR	37.496	0.9388	42.337	0.9795
MAT	37.695	0.9398	**42.577**	0.9796
PFT	37.623	0.9395	42.482	0.9796
EDiffSR	34.269	0.8879	36.990	0.9469
Finetuned on VENµS				
SwinIR	37.379	0.9383	42.311	0.9796
MAT	37.525	0.9395	42.462	0.9796
PFT	37.359	0.9384	42.558	**0.9797**
EDiffSR	34.131	0.8866	37.032	0.9451

Visual comparisons in Figure 1, which contrasts Bicubic interpolation, and SR models, and the ground-truth on a Sentinel-2 validation image, confirm these quantitative trends. A clear progression in perceptual sharpness is visible, moving from the baseline Bicubic, to SwinIR, MAT, PFT, and finally to EDiffSR. The transformer-based models (MAT, PFT) generate conservative and smooth images that closely follow the LQ image, producing results that are nearly indistinguishable from each other and very close to the ground-truth (with SwinIR appearing slightly blurrier). In contrast, EDiffSR consistently produces sharper results, attempting to synthesize fine-grained textures that look perceptually similar to the ground-truth. However, as these details are not contained in the LQ image, they are effectively hallucinated, leading to noticeable noise and textural patterns that do not align with the GT. These misalignments are harshly punished by distortion-based metrics like PSNR/SSIM, explaining EDiffSR’s lower quantitative scores despite its perceptual sharpness.

4.2 Non-reference Evaluation

We next evaluated the perceptual quality of native upscaling using FID and NIQE. For NIQE, we used a standard pre-trained model and custom-trained models fit to our satellite data. The NIQE (custom) versions are the most reliable metrics, as they are specifically trained on each experiments reference set. For all three metrics, lower scores are better.

4.2.1 Native Blind SR on Sentinel-2 (20 m$\rightarrow$10 m) This experiment tested generalization to real sensor data by upscaling native Sentinel-2 20 m bands and comparing them to the 10 m bands. The results in Table 3 show EDiffSR again achieves the best scores across all metrics. The transformer models also generalize well, with only minor differences between their training

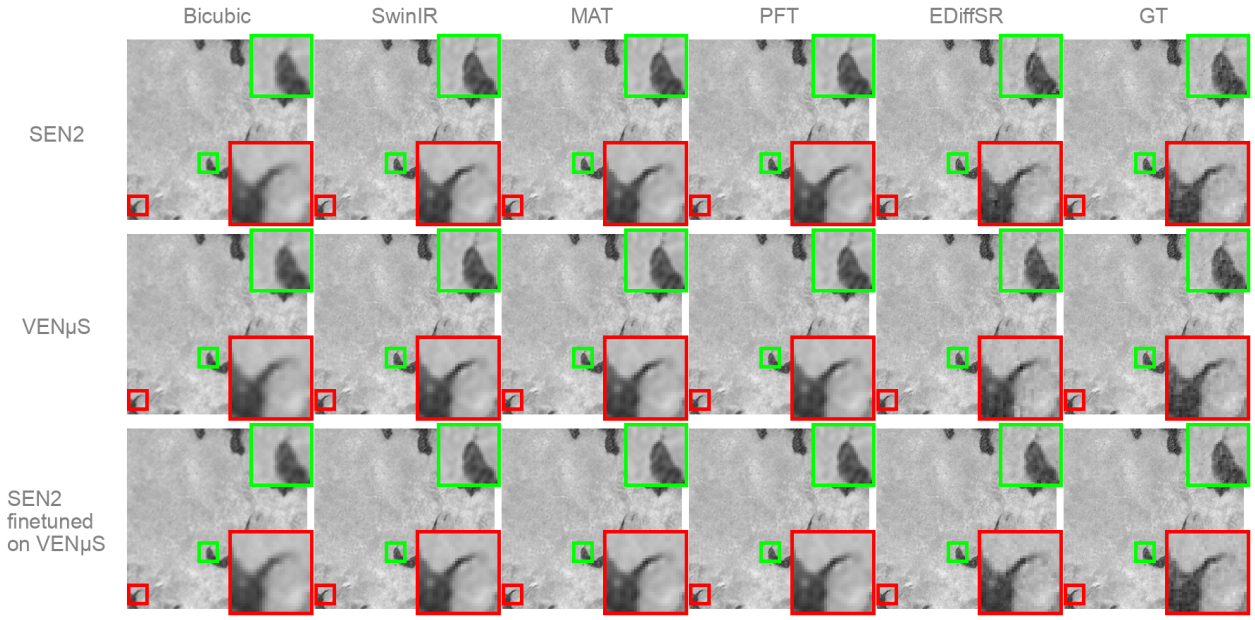

Figure 1. Visual comparison of bicubic interpolation, the SR models and ground-truth on a Sentinel-2 validation image.

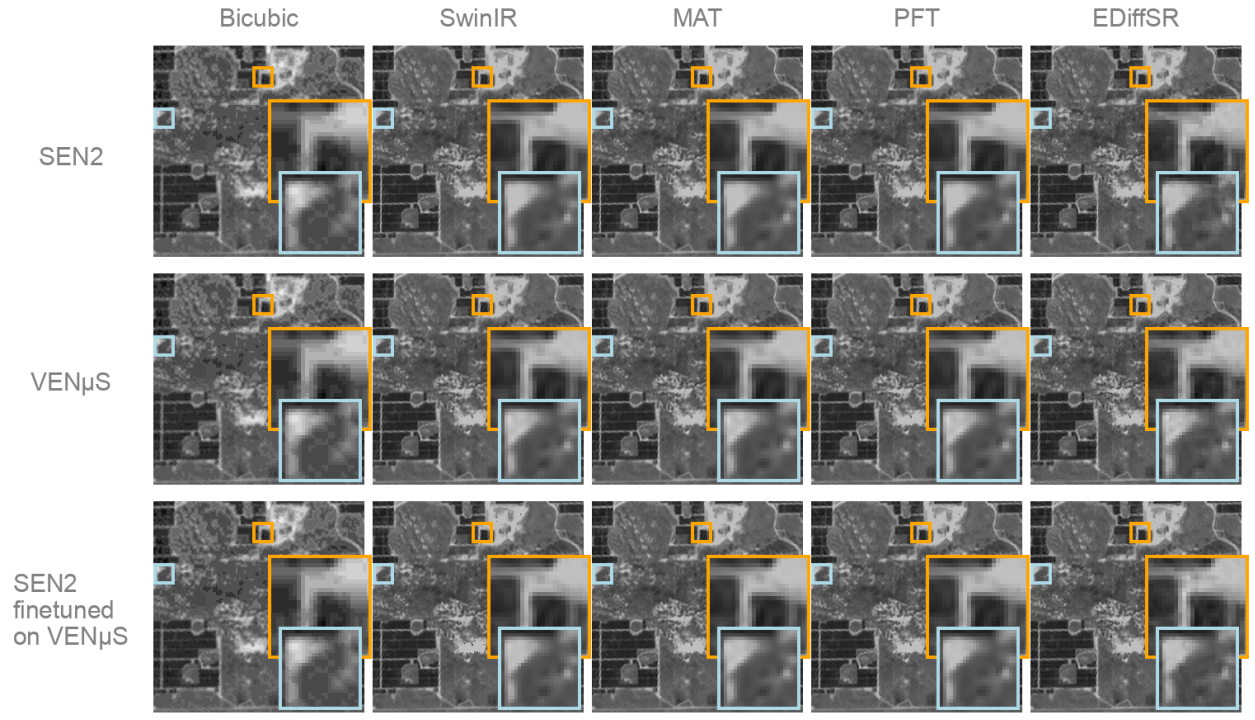

Figure 2. Qualitative results for the native Sentinel-2 20 m→10 m upscaling task, showing Bicubic interpolation alongside representative SR model outputs.

configurations. Qualitatively, EDiffSR’s outputs contained more fine-grained detail and contrast but also added visible texture in uniform areas, while the transformers produced cleaner, more stable structures. These quantitative findings are confirmed by the visual comparisons in Figure 2. The transformer models enhance the LQ image by introducing sharper boundaries and more stable structures while remaining faithful to the input. EDiffSR generates the most fine-grained details and stronger local contrast but occasionally introduces artificial textures or noise in otherwise uniform regions.

4.2.2 Cross-Dataset SR on Sentinel-2 (10 m→5 m)

This final test evaluated a cross-sensor, cross-resolution task: upscaling Sentinel-2 10 m bands to 5 m and comparing them against a VENµS 5 m reference. This experiment introduces a large domain shift. As shown in Table 4, the rankings shift clearly. The transformer models now achieve the best results on both the custom NIQE and FID. EDiffSR struggles with the domain shift,

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-3-2026-215-2026 | © Author(s) 2026. CC BY 4.0 License.

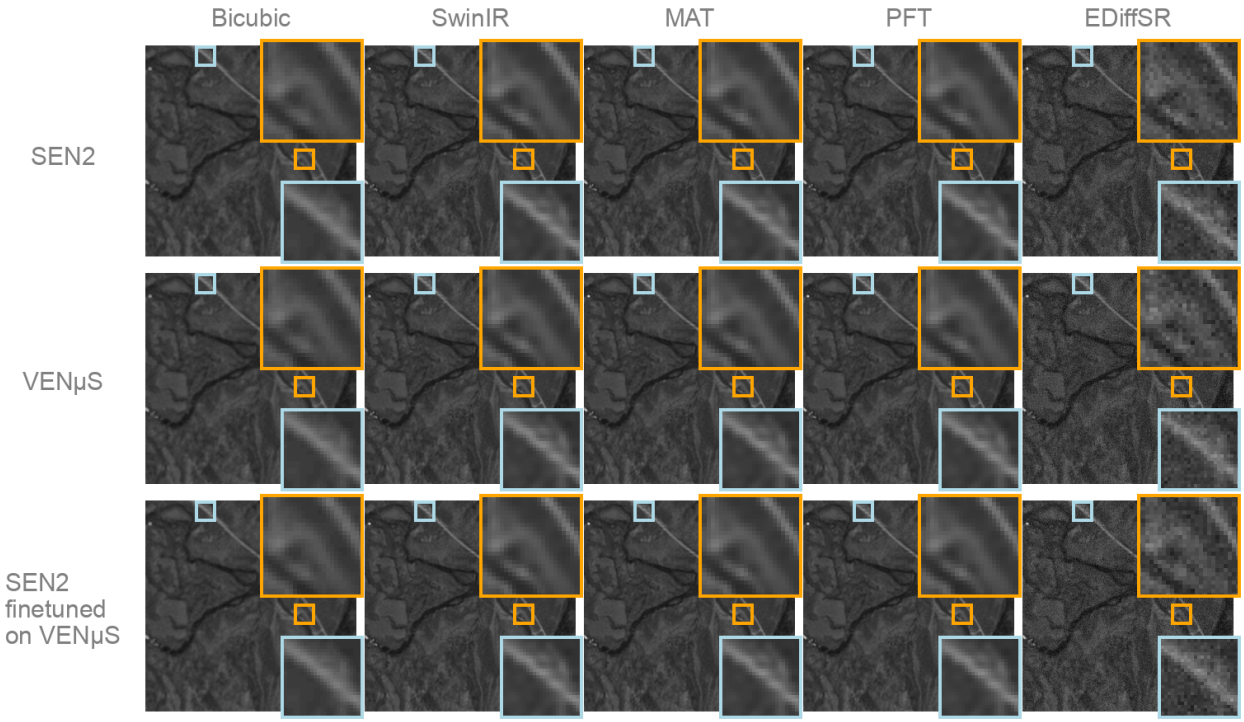

Figure 3. Qualitative results for the native Sentinel-2 10 m→5 m upscaling task, showing bicubic interpolation alongside representative SR model outputs.

Table 3. Evaluation on Sentinel-2 data (20 m → 10 m).

Model (Training)	FID (↓)	NIQE (pre-train) (↓)	NIQE (custom) (↓)
Reference (S2 10m)	0.000	4.438	4.879
Bicubic	34.098	6.445	8.907
S2-trained			
SwinIR	21.806	6.766	9.958
MAT	21.518	6.741	10.000
PFT	21.509	6.709	9.908
EDiffSR	**17.190**	**4.929**	**5.667**
VENμS-trained			
SwinIR	21.275	6.587	10.519
MAT	21.329	6.590	10.505
PFT	21.103	6.601	10.531
EDiffSR	19.796	5.315	6.883
Finetuned on VENμS			
SwinIR	21.082	6.577	10.518
MAT	21.062	6.611	10.516
PFT	20.745	6.582	10.502
EDiffSR	18.563	5.577	8.056

Table 4. Evaluation on cross-dataset on Sentinel-2 (10 m → 5 m).

Model (Training)	FID (↓)	NIQE (pre-train) (↓)	NIQE (custom) (↓)
Reference (VENμS 5m)	0.000	5.856	4.847
Bicubic	64.176	5.986	9.521
S2-trained			
SwinIR	**58.449**	6.260	9.234
MAT	58.744	6.247	9.172
PFT	58.715	6.232	9.166
EDiffSR	62.126	**5.028**	12.580
VENμS-trained			
SwinIR	59.874	6.175	8.698
MAT	59.942	6.183	8.701
PFT	60.041	6.183	**8.684**
EDiffSR	64.145	5.151	10.469
Finetuned on VENμS			
SwinIR	59.930	6.177	8.736
MAT	59.863	6.182	8.746
PFT	59.785	6.164	8.759
EDiffSR	62.752	5.692	11.956

scoring worse than the transformers. This suggests the transformers’ smoother, more faithful reconstructions adapt better to the new sensor’s characteristics. EDiffSR still performs well on the pre-trained NIQE, indicating its outputs are perceptually clean, but they fail to capture the specific statistical ”style” of the VENμS reference. Qualitatively, as seen in Figure 3, EDiffSR struggles more clearly in this setting than in previous experiments, introducing strong noise and overly sharp artifacts. In contrast, the transformer-based models generate smoother yet coherent structures that remain closer to the bicubic upsampling. These observations are consistent with the quantitative results, where the transformers generalize better to the VENμS domain, while EDiffSR produces less stable outputs.

4.3 Downstream Field Boundary Detection

To prove real-world utility, we evaluated the SR models on a field boundary detection task. We assess the trustworthiness of super-resolved Sentinel-2 imagery by generating soil masks and comparing their boundaries against a high-resolution (5 m GSD) grount-truth (GT). To match the GT, we upscaled three Sentinel-2 bands (B04, B08, B12) to 5 m. This involved a two-step process: Sentinel-trained models upscaled B12 (20 m→10 m), and VENμS-finetuned models upscaled all three bands (10 m→5 m). This 5 m upscaling was applied to all SR methods (Bicubic, SwinIR, MAT, PFT, EDiffSR), with Bicubic as the baseline.

4.3.1 GT Generation The GT creation is a complex, multi-step process. First, we acquired pre-computed field parcel masks (polygons outlining agricultural fields) from the Ger-

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-3-2026-215-2026 | © Author(s) 2026. CC BY 4.0 License.

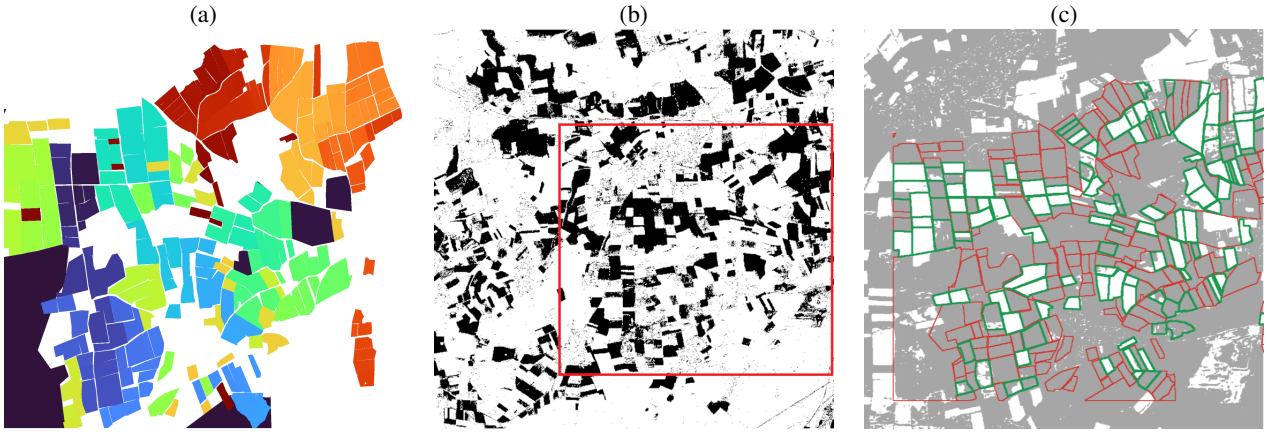

Figure 4. Ground-truth data for field detection: (a) High-resolution field parcel mask (3 m GSD) from PlanetScope. (b) Binary soil mask from a bicubic-upsampled Sentinel image (5 m GSD). (c) Selection of valid (green) versus invalid (red) ground-truth parcels used for the Boundary F1 evaluation.

man Aerospace Center (DLR) and its Department of Imaging Spectroscopy. These parcels were generated from 23 PlanetScope scenes between March-May, 2025. Individual fields were computed by grouping areas that change homogenuously over this time series. This process forms the high-resolution field parcel mask shown in Figure 4-(a).

The next stage involves generating soil masks from SR Sentinel-2 imagery to compare their boundaries against the high-resolution parcel mask. To generate this bare soil mask, we classify bare fields from the SR Sentinel-2 imagery. We used a combined spectral index (NDVI+NBR) (Heiden et al., 2022), based on the Normalized Difference Vegetation Index (NDVI) (Rouse et al., 1974) and the Normalized Burn Ratio (NBR) (Key and Benson, 2006). Its performance was validated by (Karlshoefer et al., 2025), who showed that, when paired with a regionalized threshold map tuned to local spectral characteristics, the index effectively separates photosynthetically active and inactive vegetation from soil. Pixels in the Sentinel-2 SR data with NDVI+NBR values below the local threshold (here 0.194) were classified as bare soil. To avoid misclassification in urban areas and permanent water bodies, these classes were masked using the ESA WorldCover 2021 dataset (Zanaga et al., 2022). This produced 5 m binary soil masks for our Sentinel tiles, as illustrated for the bicubic mask in Figure 4(b).

The final stage is selecting valid field boundaries. The PlanetScope parcel mask contains all agricultural parcels in the study area, but the Sentinel soil masks only capture fields bare at the acquisition dates. To ensure a fair evaluation, we filtered the GT parcels to use only relevant bare fields for metric calculations. This was achieved using embedded parcel identifiers to compute per-parcel soil pixel fractions. First, we removed small parcels (< 80 pixels) to reduce noise in the metrics. We then used the bicubic soil mask (our baseline) to calculate soil coverage for all GT parcels. Parcels with $\geq 50\%$ soil coverage were selected as valid (green), and all others were marked negative (red), as shown in Figure 4-(c). These binary GT parcels are used for the Boundary F1 evaluation.

4.3.2 Boundary F1 Evaluation We used the Boundary F1 (BF1) score, a metric that directly evaluates the alignment of predicted and GT parcel boundaries. This better reflects the strengths of SR and the information contained in our GT. As shown in Table 5, at the strict $\tau = 1$ tolerance, the transformer models (SwinIR, MAT, PFT) clearly and consistently outperform bicubic interpolation. This demonstrates that their preprocessing leads to sharper, more accurate parcel outlines. EDiffSR performed the worst, as its generative noise created unstable and fragmented boundaries. As shown in Figure 5, EDiffSR struggles with boundary detection, generating numerous false positives (in red). In contrast, the transformer models demonstrate superior performance with much cleaner and more accurate boundaries.

When evaluated at a looser $\tau = 2$ tolerance (results not shown), the performance gap narrowed, as the metric became more forgiving of bicubic’s blurred edges, allowing it to catch up to the transformers. The results at $\tau = 1$ are therefore more telling, as they highlight the specific improvement in boundary sharpness that SR provides. The BF1 experiments confirm that transformer-based SR is a valuable preprocessing step for improving field boundary detection.

Table 5. BF1 scores (↑) with $\tau = 1$ pixel tolerance.

Model	Tile 08.03.2025		Tile 18.03.2025	
	Macro (↑)	Micro (↑)	Macro (↑)	Micro (↑)
Bicubic	0.8297	0.8319	0.7523	0.7555
SwinIR	**0.8412**	**0.8331**	0.7778	0.7643
MAT	0.8405	0.8326	**0.7779**	**0.7648**
PFT	0.8410	0.8328	0.7773	0.7644
EDiffSR	0.7929	0.7333	0.7041	0.6618

5. Conclusion

This work investigated the potential of SR as a pre-processing step to bridge the resolution gap of open-access satellite data. We adapted and benchmarked three state-of-the-art transformer models (SwinIR, MAT, PFT) and one diffusion model (EDiffSR) on 16-bit, single-band satellite imagery, focusing on 20 m→10 m and 10 m→5 m upscaling.

Our experiments revealed a clear trade-off between architectural paradigms. The diffusion model, EDiffSR, produced perceptually sharper results, excelling on no-reference metrics like FID and our custom-trained NIQE. However, it performed poorly on fidelity metrics (PSNR/SSIM) and, most critically, failed to outperform bicubic upsampling in our practical downstream application. In contrast, the transformer models demonstrated

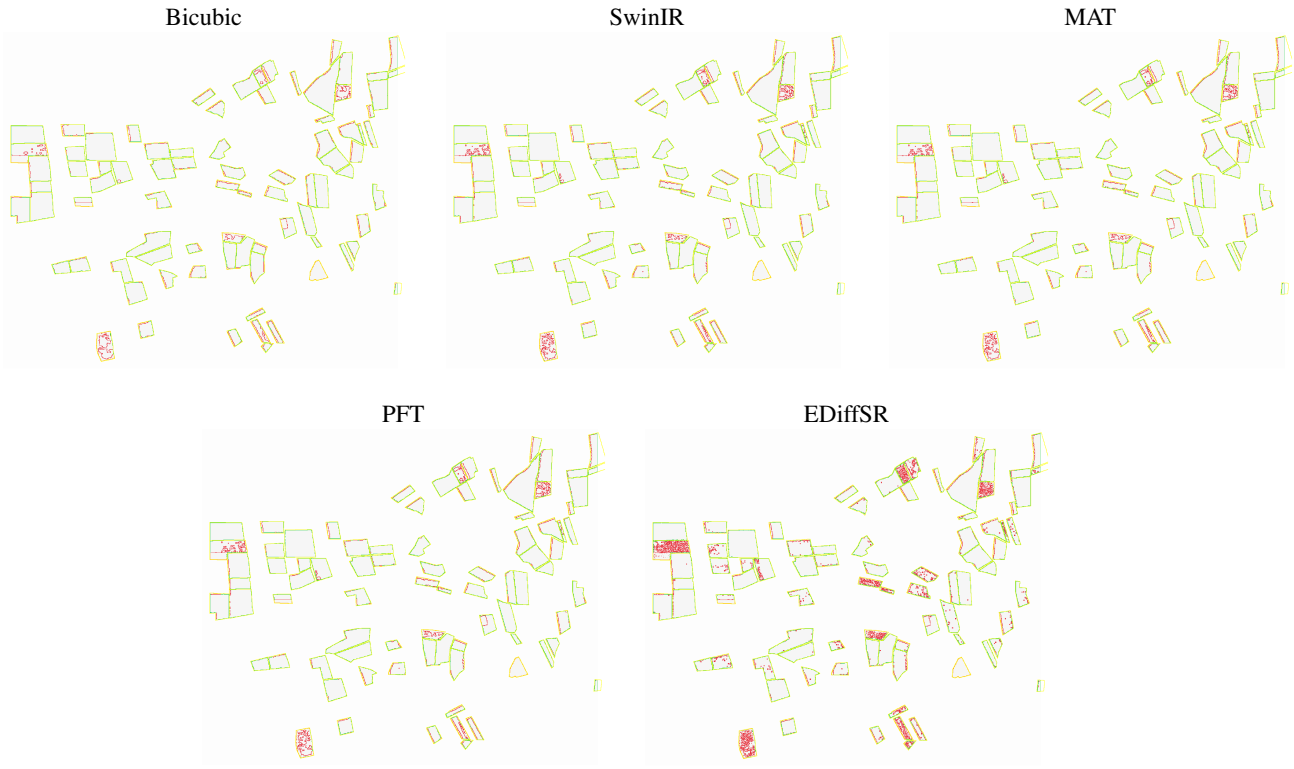

Figure 5. Visual comparison of boundary F1 maps at $\tau = 1$ pixel (tile 32UPU, 18 March 2025). Green = correct boundaries, red = false positives, orange = false negatives.

superior performance on PSNR and SSIM. Most importantly, they consistently outperformed the bicubic baseline in a field boundary detection task, proving that SR can be a valuable and practical tool for improving real-world remote sensing analyses.

These findings validate the use of transformer-based SR for enhancing scientific satellite data. Future work should move beyond the simple bicubic degradation used here, incorporating more realistic sensor-specific degradation models. Furthermore, expanding the evaluation to include other downstream tasks and incorporating radiometric-aware loss functions will be crucial for developing SR methods that are not only visually plausible but also scientifically trustworthy.

References

Cornebise, J., Oršoli’c, I., Kalaitzis, F., 2022. Open high-resolution satellite imagery: The worldstrat dataset–with application to super-resolution. *Advances in Neural Information Processing Systems*, 35, 25979–25991.

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

Dong, C., Loy, C. C., He, K., Tang, X., 2014. Learning a deep convolutional network for image super-resolution.

Donike, S., Aybar, C., Gómez-Chova, L., Kalaitzis, F., 2025. Trustworthy Super-Resolution of Multispectral Sentinel-2 Imagery With Latent Diffusion. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 18, 6940-6952.

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

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

Gillies, S. et al., 2013. Rasterio: geospatial raster i/o for Python programmers.

Heiden, U., d’Angelo, P., Schwind, P., Karlshöfer, P., Müller, R., Zepp, S., Wiesmeier, M., Reinartz, P., 2022. Soil reflectance composites—improved thresholding and performance evaluation. *Remote Sensing*, 14(18), 4526.

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

Ho, J., Jain, A., Abbeel, P., 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33, 6840–6851.

Jangir, S. K., Henry, C., Merkle, N., 2025. Investigating the potential of super-resolution for road segmentation in sentinel-2 images. *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*.

Karlshoefer, P., d’Angelo, P., Eberle, J., Heiden, U., 2025. Evaluation framework for the generation of continental bare surface reflectance composites. *Geoderma*, 459, 117340.

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

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

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

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

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

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

Ma, L., Liu, Y., Zhang, X., Ye, Y., Yin, G., Johnson, B. A., 2019. Deep learning in remote sensing applications: A meta-analysis and review. *ISPRS journal of photogrammetry and remote sensing*, 152, 166–177.

Maertens, M., Izzo, D., Krzic, A., Cox, D., 2019. Super-resolution of PROBA-V images using convolutional neural networks. *Astrodynamics*, 3(4), 387–402.

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

Mittal, A., Soundararajan, R., Bovik, A. C., 2012. Making a “completely blind” image quality analyzer. *IEEE Signal processing letters*, 20(3), 209–212.

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

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

Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D. J., Norouzi, M., 2022. Image super-resolution via iterative refinement. *IEEE transactions on pattern analysis and machine intelligence*, 45(4), 4713–4726.

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

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

Wang, Z., Bovik, A. C., Sheikh, H. R., Simoncelli, E. P., 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4), 600–612.

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

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

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

Zanaga, D., Van De Kerchove, R., Daems, D., De Keersmaecker, W., Brockmann, C., Kirches, G., Wevers, J., Cartus, O., Santoro, M., Fritz, S. et al., 2022. ESA WorldCover 10 m 2021 v200. [Online; accessed 10-September-2025].