

Mind the Gap: Quantifying the Domain Gap in Cross-Sensor Diffusion Super-Resolution

Dawid Kopec¹[0009-0000-1765-5810], Katarzyna Jabłońska¹[0009-0006-6824-9437],
Wojciech Kozłowski¹[0009-0009-1532-4415], and
Maciej Zięba^{1,2}[0000-0003-4217-7712]

¹ WUST, Wybrzeże Stanisława Wyspiańskiego 27, 50-370 Wrocław, Poland
{dawid.kopec, katarzyna.jablonska, wojciech.kozlowski,
maciej.zieba}@pwr.edu.pl

² Tooploox, Tęczowa 7, 53-601 Wrocław, Poland

Abstract. Demand for high-resolution satellite imagery has increased interest in super-resolution (SR) to bridge the spatial resolution gap between freely available missions such as Sentinel-2 and commercial systems like PlanetScope. Because no sensor provides true paired low- and high-resolution observations, SR models are usually trained on synthetically degraded data, creating a domain gap on real cross-sensor imagery. In this work, we provide the first systematic study of how this synthetic-to-real mismatch affects the performance of modern diffusion-based SR models. Using a large, geometrically and temporally aligned dataset of Sentinel-2 and PlanetScope imagery, we evaluate five state-of-the-art diffusion architectures under controlled experimental settings. We also introduce *LPIPS_{Sat}*, a domain-adapted perceptual metric based on Sentinel-2 self-supervised features. Our results show two persistent challenges: synthetically trained models degrade sharply on real pairs, while models trained on real cross-sensor data exhibit optimisation difficulties and struggle to adapt to the physical and radiometric diversity. These findings highlight a key limitation of current SR and motivate methods that disentangle super-resolution from domain adaptation.

Keywords: Diffusion · Super-resolution · Remote Sensing · Cross-sensor

1 Introduction

Remote sensing (RS) has become critical for decision-making in applications such as wildfire monitoring, mineral exploration, and water resource management [17, 26, 29]. Despite the increasing availability of satellite imagery, a key limitation persists: the spatial resolution dilemma. Freely accessible missions like Sentinel-2 provide global coverage and high temporal frequency but at a moderate spatial resolution of 10 m/pixel, while commercial systems like PlanetScope deliver 3 m/pixel imagery at higher cost and with reduced temporal availability. For fine-grained tasks such as burning area mapping or building delineation, this difference limits the quality of the analysis [1, 18].

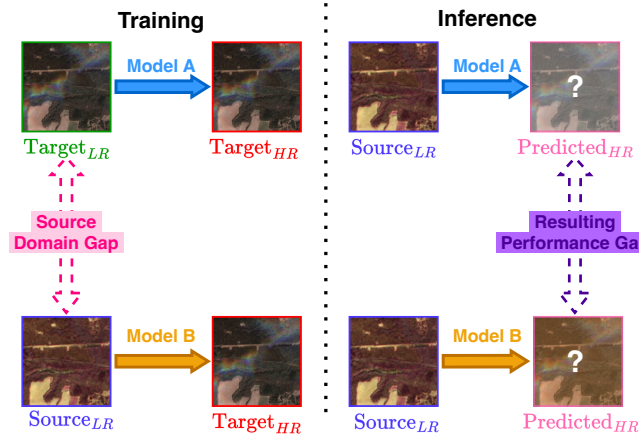


Fig. 1: Conceptualizing the cross-sensor SR problem. We compare two distinct models: **Model A**, trained on the synthetic domain ($Target_{LR}$), and **Model B**, trained on the real-world domain ($Source_{LR}$). This paper investigates how the Source Domain Gap (pink) between their training inputs translates into a Resulting Performance Gap (purple) at inference.

Super-Resolution (SR) attempts to bridge this gap by reconstructing high-resolution (HR) images from low-resolution (LR) observations. SR has shown strong benefits in downstream RS tasks [27, 8, 36], and diffusion-based generative models now represent the state-of-the-art due to their texture recovery and global consistency [15, 35, 19, 16].

However, satellite super-resolution faces a fundamental challenge: the absence of true LR–HR pairs, since each sensor has a fixed native resolution. As a result, most prior work relies on synthetic intra-sensor degradation (e.g., bicubic downsampling), which fails to capture the complex spectral and radiometric properties of real sensors such as Sentinel-2.

In practice, the objective is cross-sensor enhancement, for example, upscaling Sentinel-2 imagery to PlanetScope-like quality. This introduces a substantial domain gap between synthetic training data and real LR observations, caused by differences in spatial resolution, spectral response, and radiometry. As illustrated in Figure 1, models trained on synthetic data often produce unstable results when applied to real satellite imagery.

In this work, we systematically study the *impact of the domain gap* on satellite SR. Specifically, we investigate how diffusion-based SR models trained on synthetic degradations generalise to real cross-sensor scenarios, quantifying performance degradation. Our contributions are threefold:

- (i) We analyse the *synthetic-to-real* generalisation gap using diverse diffusion-based SR models under consistent conditions.
- (ii) We introduce a perceptually grounded evaluation framework, including the

novel $LPIPS_{Sat}$ metric based on Sentinel-2 representation learning.

(iii) We construct a large-scale paired dataset of Sentinel-2 and PlanetScope imagery, temporally and geometrically aligned to enable rigorous cross-sensor SR evaluation.

We provide the first systematic quantification of how the domain gap influences modern diffusion-based SR, aiming to bridge the gap between synthetic benchmarks and real-world applications.

2 Related Work

Diffusion Models for Super-Resolution. The application of diffusion models [7, 3] to Super-Resolution (SR) began with SR3 [25] and SRDiff [11], demonstrating that iterative refinement captures high-fidelity details. Later work introduced upsampling controls [5], modifications to the sampling trajectory for improved efficiency [10] and quality [35], and more accurate degradation modeling [15]. Recently, attention shifted to the *blind-SR* [12] setting to improve performance on real-world data, culminating in powerful prior-guided models like DiffBIR [13] and SUPIR [34].

Diffusion in Remote Sensing SR. Although diffusion-based SR has achieved remarkable progress in the natural image domains, its application to remote sensing (RS) imagery introduces additional challenges. Satellite data are characterised by large image sizes, multi-spectral channels, and strong geometric variability, making them an ideal, yet demanding, use case for generative models [23]. Early works such as DMDC [16] demonstrated that diffusion frameworks can plausibly reconstruct missing spatial details in satellite imagery. Later studies improved this approach by integrating Transformer-based architectures to capture long-range spatial dependencies [33]. Parallel research aimed to reduce the high computational burden of diffusion models through lightweight conditioning and efficient sampling strategies [19]. Recent work targets *blind-SR* setting, where models learn to handle complex sensor-specific degradations representative of real acquisition conditions [32]. Despite these advances, most existing methods continue to rely on synthetic, intra-sensor degradations, which limits their applicability to operational RS data.

Cross-Sensor Domain Gap. The central challenge in real-world RS-SR remains the cross-sensor domain gap: models trained on synthetic degradations (e.g., bicubic downsampling) often fail when tested on data from physically distinct sensors. This issue, closely related to the *blind-SR* paradigm, arises from discrepancies in sensor optics, spectral response functions, radiometric calibration, and atmospheric conditions. Recent work has explored several directions to mitigate this problem, from domain-adaptive GANs [31] to customised diffusion architectures [20]. A comprehensive analysis by Michel *et al.* [21] highlighted how real-world geometric and radiometric inconsistencies invalidate standard SR evaluation metrics, highlighting the need for cross-domain benchmarks. However, despite these findings, the impact of this domain gap on modern diffusion-based SR models remains largely unexplored.

In this paper, we address this gap by presenting the first systematic evaluation of diffusion-based super-resolution under cross-sensor conditions. Our benchmark quantifies the synthetic-to-real discrepancy and establishes a foundation for robust cross-sensor generalisation.

3 Methodology

3.1 Dataset

We constructed a dataset of geometrically and temporally aligned image pairs acquired from the *Sentinel-2* [4] and *PlanetScope* [22] satellites. The choice of these two sensors is particularly suitable for this study due to their complementary characteristics. Sentinel-2 provides freely available multispectral imagery at a spatial resolution of 10 m/pixel, which is, however, insufficient for many fine-grained downstream applications. In contrast, PlanetScope offers 3 m/pixel resolution, approaching commercial-grade quality, though it is typically subject to licencing restrictions. Using two satellite systems instead of aerial imagery allows for large-scale coverage while maintaining comparable imaging conditions. The moderate scale factor ($< 4\times$) makes SR feasible for most modern generative methods, originally developed for much higher scaling factors. Both datasets were acquired through PlanetHub, ensuring sub-pixel geometric alignment consistent with standard L2A processing levels.

We selected 175 PlanetScope regions from USA, each covering approximately $35 \times 25\text{km}$ with cloud coverage below 10%. For every region acquired between 15 – 30 December 2024, a cloud-free Sentinel-2 scene was retrieved. Temporal proximity was prioritised: 154 pairs fall within ± 15 days, 10 within ± 30 days, and 11 within ± 45 days. While temporal offsets affect illumination, we treat these residual misalignments as an integral component of the real-world cross-sensor domain gap rather than noise.

To strictly isolate sensor-specific gaps from scene evolution, we rigorously filtered the test set to include only pairs within a narrow ± 15 -day window, while the remaining 17 non-overlapping regions were reserved for testing. The geographic distribution is visualised in Figure 2.

The resulting dataset serves as the basis for three complementary experimental configurations (Figure 3):

1. **Establishing the Synthetic Baseline (Case I):** Training and testing on $Planet_{LR} \rightarrow Planet_{HR}$. This baseline setup represents the common but unrealistic scenario in which models are trained and evaluated on synthetically degraded data.
2. **Quantifying the Domain Gap (Case II):** Training on $Planet_{LR} \rightarrow Planet_{HR}$ and testing on $Sentinel_{LR} \rightarrow Planet_{HR}$. This configuration measures the real-world impact of the domain gap, highlighting the limitations of training solely on synthetic data.
3. **Evaluating Direct Mapping (Case III):** Training and testing on $Sentinel_{LR} \rightarrow Planet_{HR}$. This setup enables evaluating the potential and limitations of directly learning cross-sensor mappings.

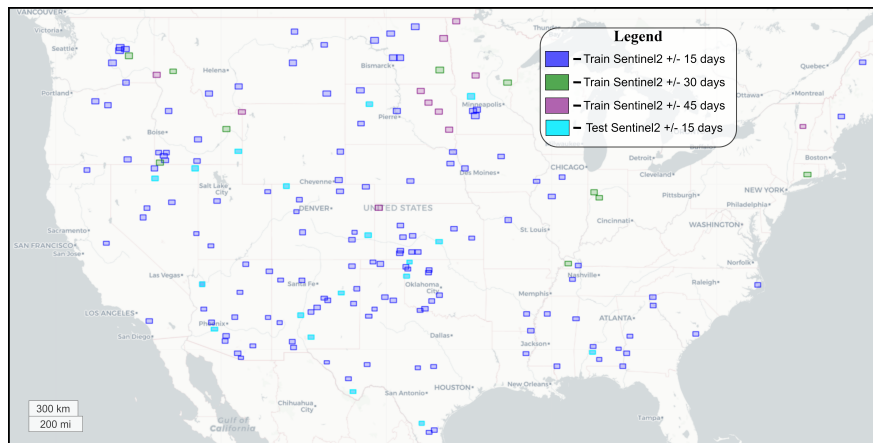


Fig. 2: Map of the train/test area divided into collected patches. Sentinel-2 image acquisition time $\pm 15, 30$ and 45 days from the PlanetScope image time and test data.

3.2 Benchmarked Diffusion Methods

All evaluated models follow the denoising diffusion paradigm [7], which defines a *forward* noising process and a learnt *reverse* process parameterized by a neural network (typically a U-Net). Let x_0 denote the clean target image, y_0 the conditioning low resolution input, t the timestep, and $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ Gaussian noise. Diffusion models are well-suited for satellite SR because their iterative refinement allows them to recover high-frequency details while maintaining global consistency. Unlike deterministic regression, this iterative refinement enables the generative synthesis of high-frequency details, prioritising perceptual realism over the pixel-average fidelity typical of feed-forward baselines.

Diffusion (DDPM) [7] Serves as our baseline. The method trains its network ϵ_θ with a simple mean-squared error objective on the noise:

$$\mathcal{L}_{\text{DDPM}} = \mathbb{E}_{t, x_0, \epsilon \sim \mathcal{N}(0, \mathbf{I})} [\|\epsilon - \epsilon_\theta(x_t, t, y_0)\|^2], \quad (1)$$

where, $x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon$ and $\bar{\alpha}_t$ is predefined variance schedule parameters. DDPM directly models the reverse noising dynamics by estimating the added noise. As a generative model, its ability for domain adaptation is implicit; it must learn complex one-to-many mapping directly from a pixel-level objective.

Flow Matching (FM) [14] Represents a deterministic generative paradigm. Instead of learning a score function, Flow Matching trains a deterministic vector field v_θ that transports samples from a simple prior distribution to the data distribution. The training objective is defined as:

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{t, p_t(x_t)} [\|v_\theta(x_t, t, y_0) - u_t(x_t)\|^2], \quad (2)$$

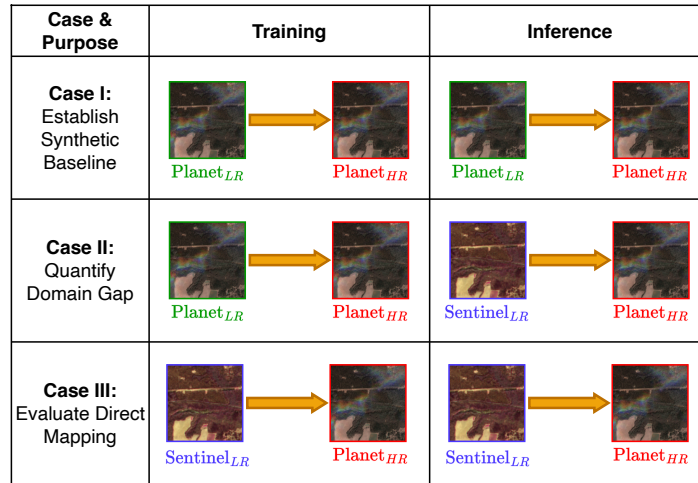


Fig. 3: The three experimental configurations. **Case I** establishes the synthetic baseline. **Case II** quantifies the synthetic-to-real domain gap. **Case III** evaluates the alternative direct mapping.

where, $x_t = (1 - t)x_0 + t\epsilon$, $p_t(x_t)$ is a target probability density path and $u_t(x_t)$ is the corresponding vector field of this path. By learning a deterministic flow, FM can be more stable and efficient for domain adaptation.

Image-to-Image Schrödinger Bridge (I²SB) [15] Frames image-to-image translation as learning a stochastic bridge between conditional distributions, but uses an SGM-style ϵ parameterisation for stability [3]. The network ϵ_θ minimises:

$$\mathcal{L}_{I^2SB} = \mathbb{E}_{t, \mathbf{x}_0, \epsilon \sim \mathcal{N}(0, \mathbf{I})} \left[\|\epsilon_\theta(\mathbf{x}_t, t, y_0) - \epsilon\|^2 \right], \quad (3)$$

where, $x_t = \frac{\bar{\sigma}_t^2}{\bar{\sigma}_t^2 + \sigma_t^2} x_0 + \frac{\sigma_t^2}{\bar{\sigma}_t^2 + \sigma_t^2} y_0 + \sqrt{\frac{\sigma_t^2 \bar{\sigma}_t^2}{\bar{\sigma}_t^2 + \sigma_t^2}} \epsilon$, σ_t and $\bar{\sigma}_t$ are accumulated noise variance relative to target x_0 and condition y , respectively. Using this stable, SGM-style objective, I²SB can focus on the generative process (the bridge) rather than on a difficult-to-learn score function, which is ideal for complex cross-domain adaptation.

ResShift [35] Re-parameterizes the Markov chain to shift residuals between an HR image and its downsample version (LR). Model trains a network f_θ to predict the clean HR image directly instead of the noise:

$$\mathcal{L}_{ResShift} = \mathbb{E}_{t, x_0} \left[w_t \|f_\theta(x_t, y_0, t) - x_0\|^2 \right] \quad (4)$$

where, $x_t = (1 - \eta_t)x_0 + \eta_t y_0 + \kappa \sqrt{\eta_t} \cdot \epsilon$, η_t is shifting sequence, κ is a hyperparameter controlling the noise variance, and w_t are timestep weights. Direct image prediction supplies strong, pixel-level supervision that is helpful for SR and large domain gaps.

UniDB [37] Generalises the concept of diffusion bridges by defining the generative process as a *Stochastic Optimal Control (SOC)*. Instead of just matching a score, the goal is to find an optimal controller $u_{t,\gamma}^*$ that directs the process from the source to the target:

$$\mathcal{L}_{\text{UniDB}} = \mathbb{E}_{t, \mathbf{x}_0} \left[\frac{1}{2\sigma_{t-1, \theta}^2} \|\mu_{t-1, \theta} - \mu_{t-1, \gamma}\|_1 \right] \quad (5)$$

where, $\mu_{t-1, \theta}$ is the predicted mean of the reverse step $p_{\theta}(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_T)$ and $\mu_{t-1, \gamma}$ is the target mean derived from the forward process. \mathbf{x}_t is sampled from the bridge posterior $p(\mathbf{x}_t|\mathbf{x}_0, \mathbf{x}_T)$. This allows UniDB to robustly bridge two distinct domains, providing a consistent generative trajectory even under significant cross-sensor or cross-domain shifts.

Once trained, each model can sample iteratively by transporting the initial sample x_T to the distribution x_0 using trained functions with parameters θ . In experiments configurations *C2* and *C3* we use x_0 as *PlanetHR*, but y_0 is different in testing. For all experiments, we use the official, publicly available implementations provided by the authors to ensure a fair and reproducible comparison.

3.3 Evaluation Metrics

We evaluated models using four metrics covering both reconstruction fidelity and perceptual quality: PSNR, SSIM, LPIPS, and our proposed satellite-adapted variant $LPIPS_{\text{Sat}}$.

PSNR measures pixel-wise reconstruction accuracy but is highly sensitive to misalignment and correlates poorly with perceptual quality, particularly in high-frequency remote sensing textures.

SSIM captures luminance, contrast, and structural similarity, offering a more perceptual measure than PSNR, although it still struggles with the complex spatial patterns characteristic of satellite imagery.

LPIPS compares deep feature representations rather than raw pixels, providing a stronger proxy for perceptual similarity. Lower LPIPS values indicate more realistic texture and structural preservation.

LPIPS_{Sat} is our domain-adapted metric designed specifically for satellite data. Standard LPIPS relies on ImageNet-pretrained networks, which do not reflect the spectral, radiometric, or textural characteristics of remote sensing imagery. To address this, we employ a ResNet-50 [6] backbone from **TorchGeo**, pretrained in a self-supervised manner on Sentinel-2 RGB data [28]. To accommodate the 6-channel multispectral data, we compute the metric strictly on the corresponding RGB spectral bands (B4, B3, B2), ensuring alignment with the backbone’s pretraining domain. The network remains frozen and for each reference/reconstruction pair (normalised to $[0, 1]$), feature maps from **conv1** and **layers 1-4** are extracted, channel-wise L_2 normalised and compared. Layer-wise feature differences are averaged spatially and then across layers to yield a single perceptual distance.

Using a representation learned directly from Sentinel-2, $LPIPS_{\text{Sat}}$ provides a perceptual assessment that more faithfully reflects structures and textures present in Earth observation imagery, offering a more meaningful evaluation than generic perceptual metrics.

3.4 Downstream Task Evaluation

To assess the practical impact of the domain gap in SR models, we evaluated them on a downstream task: burned-area change delineation. We used the CaBuAr dataset [2] from `torchgeo.datasets`, which contains 424 pre-/post-fire Sentinel-2 image pairs (512×512 px) from California wildfires (2015–2022). The data span diverse vegetation types and fire severities, providing a realistic test of cross-sensor robustness.

The downstream model was a U-Net (31M parameters) following the configuration from the CaBuAr paper. Burned-area mapping is moderately complex and highly sensitive to texture and spectral cues—features that SR methods may enhance or distort—while its binary segmentation nature enables clear, quantitative comparison.

Each CaBuAr image was optionally enhanced using the SR models from Section 3.2; SR was applied independently to pre- and post-fire inputs to isolate its effect.

We evaluated three configurations:

1. **Baseline:** Training and testing on original Sentinel-2 images.
2. **SR-tested:** Training on originals, testing on SR-enhanced inputs (measures the domain mismatch introduced by SR).
3. **SR-trained:** Training and testing on SR-enhanced images (checks whether retraining mitigates the gap).

All setups were trained for 60 epochs using identical hyperparameters and CaBuAr splits. Performance was measured with *IoU* and *F1-score*, standard metrics for binary segmentation.

This evaluation provides a practical view of how SR affects downstream environmental monitoring and whether it preserves task-relevant information.

4 Experiments

4.1 Training Details

All experiments were conducted using the dataset described in Section 3.1. Each of the 5 evaluated diffusion-based SR methods were trained under identical conditions to ensure fair comparison. All models employed the same U-Net [24] backbone with 30M parameters, implemented using the `diffusers` library [30]. This unified architecture was chosen to isolate the effect of the diffusion formulation itself rather than network capacity or design.

Training was performed for 300,000 steps, which in preliminary experiments was found sufficient to ensure convergence for all methods given the dataset scale

and model size. Optimisation was carried out using the AdamW optimizer [9] with a learning rate of 1×10^{-4} , $\eta_{min} = 1 \times 10^{-7}$, and a cosine annealing scheduler. The batch size was set to 4, and training was performed on NVIDIA A100 GPUs in mixed-precision mode to balance efficiency and numerical stability. No data augmentation was applied to preserve the spectral and spatial consistency of the paired satellite imagery.

The validation was conducted every 10k iterations, and the final checkpoint was selected based on the best $LPIPS_{Sat}$ score on the validation set. For inference, all models used a standardised budget of $T = 1000$ steps with default schedulers and no classifier-free guidance.

4.2 Data Preparation

All datasets were processed to ensure geometric, temporal, and spatial consistency across sensors. First, both Sentinel-2 and PlanetScope scenes were reprojected to a common coordinate reference system (EPSG:4326). Sub-pixel geometric alignment was achieved through affine coregistration.

Each PlanetScope image was divided into 512×512 tiles using a 25% overlap. For every high-resolution tile ($Planet_{HR}$), a corresponding Sentinel-2 patch was extracted and downsampled by a factor of $\frac{10}{3}$ to match spatial resolution. Tile pairs containing more than 10% missing pixels (zero-valued across all channels) were removed.

To support diffusion-based SR models, all Sentinel-2 patches were upsampled to 512×512 using bicubic interpolation, forming the $Sentinel_{LR}$ set. For synthetic degradation experiments, low-resolution PlanetScope tiles ($Planet_{LR}$) were created by downsampling $Planet_{HR}$ to Sentinel-2 scale using the `rasterio` library—ensuring geospatial consistency—and subsequently upsampling back to 512×512 using bicubic interpolation.

The final dataset comprises 61,239 training pairs and 7,303 test pairs, derived exclusively from the ± 15 -day subset to reduce temporal variability in the evaluation stage.

A key challenge in cross-sensor tasks is spectral alignment, as sensors capture data in distinct bands corresponding to specific wavelengths. Sentinel-2 (12 bands) and PlanetScope (8 bands) do not overlap perfectly. To ensure a physically valid SR mapping, we selected six bands with matching spectral properties, as detailed in Table 1. Using non-overlapping bands would force the model to "hallucinate" spectral information (e.g., predict Red from Blue), which is not a valid SR task.

4.3 Experimental Results

The quantitative results are summarised in Table 2, and the performance of the models can be seen in Figure 4. The *Default* row provides a reference for both the synthetic ($C1$) and real ($C2$, $C3$) gaps.

Table 1: Spectral band alignment for cross-sensor experiments. We selected six bands with overlapping spectral wavelengths (λ) to ensure a physically consistent mapping.

Common Name	Wavelength (λ)	PlanetScope Band ID	Sentinel-2 Band ID
Coastal Blue	~ 443 nm	1	1
Blue	~ 490 nm	2	2
Green	~ 560 nm	4	3
Red	~ 665 nm	6	4
Red Edge	~ 705 nm	7	5
Near-Infrared (NIR)	~ 865 nm	8	8A

Table 2: Quantitative metric comparison. *C1*: synthetic (*Planet* \rightarrow *Planet*). *C2*: cross-sensor generalization (tested on *Sentinel*). *C3*: cross-sensor adaptation (trained on *Sentinel*). *Default*: metrics between low-resolution inputs and high-resolution ground truth. LPIPS_{Sat} results are scaled by ($\times 10^{-3}$). The best and second-best results are highlighted in **bold** and underline, respectively.

Metric	PSNR \uparrow			SSIM \uparrow			LPIPS \downarrow			LPIPS _{Sat} \downarrow		
	<i>C1</i>	<i>C2</i>	<i>C3</i>	<i>C1</i>	<i>C2</i>	<i>C3</i>	<i>C1</i>	<i>C2</i>	<i>C3</i>	<i>C1</i>	<i>C2</i>	<i>C3</i>
<i>Diffusion</i>	45.26	16.71	20.83	0.98	0.61	0.82	0.04	0.39	0.37	0.37	2.21	1.96
<i>I²SB</i>	45.20	16.72	22.39	0.98	0.62	0.85	0.03	0.37	0.27	0.34	2.15	1.53
<i>ResShift</i>	45.58	16.73	16.88	0.99	0.63	0.64	0.04	0.36	0.30	0.36	2.14	1.90
<i>FM</i>	44.72	16.70	<u>22.56</u>	0.97	0.61	0.83	<u>0.03</u>	0.39	0.28	0.37	2.18	<u>1.60</u>
<i>Unidb</i>	44.41	16.70	23.81	0.97	0.61	<u>0.84</u>	0.02	0.38	0.26	<u>0.35</u>	2.18	1.63
<i>Default</i>	43.55	16.88		0.97	0.64		0.10	0.30		0.45	1.89	

In the synthetic case (*Planet*_{LR} \rightarrow *Planet*_{HR}, *C1*), the gap is minimal—PSNR exceeds 43 dB and SSIM reaches 0.97—indicating that artificial degradation produces a simplistic learning setup. In this regime, the reconstructions are nearly indistinguishable from ground truth, both quantitatively and perceptually.

The situation changes dramatically in the real cross-sensor setting. Even the *Default* comparison (*Sentinel*_{LR} \rightarrow *Planet*_{HR}), which measures *the domain gap*, shows a substantial drop (PSNR \approx 16.9 dB, LPIPS = 0.296). More importantly, this gap is highly non-uniform. As shown in Table 3, the variability across all metrics is extremely high. This confirms that real-world discrepancies in alignment, illumination, and sensor characteristics introduce significant, unpredictable variability, posing a major challenge for any consistent learning model.

When diffusion models trained on synthetic data are applied to real Sentinel imagery (*C2*), they fail catastrophically. Performance deteriorates even further, falling below the *Default* baseline in both PSNR and SSIM. Perceptual scores (LPIPS/LPIPS_{Sat}) also confirm that the models are *actively degrading* the input

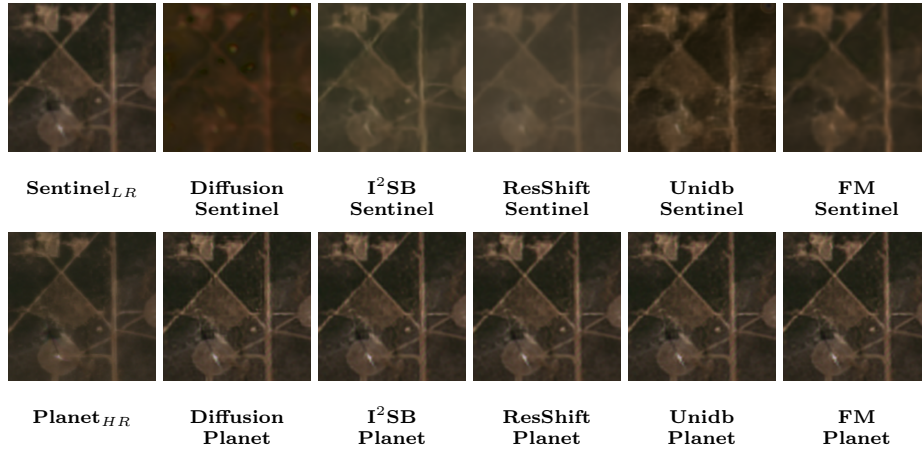


Fig. 4: Visual comparison for the cross-sensor task (zoomed in image). (Top row) $Sentinel_{LR}$ input (left) and $C3$ (real-trained) model outputs. (Bottom row) $Planet_{HR}$ ground truth (left) and $C2$ (synthetic-trained) model outputs.

Table 3: Metric variability across Sentinel–Planet test pairs, quantifying the gap between low-resolution inputs and high-resolution ground truth.

Metric	$PSNR \uparrow$	$SSIM \uparrow$	$LPIPS \downarrow$	$LPIPS_{Sat} \downarrow$
<i>Min</i>	2.306	-0.893	0.021	0.567
<i>Max</i>	44.256	0.988	1.061	5.388
<i>Mean</i>	25.135	0.642	0.295	1.888
<i>Std</i>	11.729	0.511	0.274	1.114

imagery, likely by imposing false synthetic-data priors (e.g., bicubic-like sharpness) onto physically different Sentinel data. This clearly demonstrates that synthetic training does not provide a meaningful preparation for real cross-sensor challenges.

Training and testing on real data ($C3$) allow partial recovery. Perceptual metrics improve significantly over the $C2$ failure, indicating limited domain adaptation. However, high variance in sensor-data (Table 3) hinders consistent optimization across architectures. As a result, performance remains far below the synthetic setting ($C1$), showing that even in-domain data do not produce a stable, physically consistent mapping.

To better capture these perceptual failures, our proposed $LPIPS_{Sat}$ metric follows trends similar to LPIPS but offers higher sensitivity to the spectral and structural discrepancies unique to remote sensing. For example, in $C3$, I^2SB achieves the lowest $LPIPS_{Sat}$ distance (1.532), indicating that it best preserves spatial and spectral realism under these challenging real-world conditions.

Overall, these findings confirm two distinct challenges. First, a severe domain gap ($C1$ vs. $C2$) that causes synthetically trained models to fail completely.

Table 4: Downstream task performance (IoU/F1-score \uparrow). *SR-tested*: U-Net trained on baseline, tested on SR images. *SR-trained*: U-Net trained and tested on SR images. *Default*: Baseline performance (no SR). *C2/C3* indicate which SR model outputs were used.

Metric	IoU \uparrow				F1 \uparrow			
	<i>SR-tested</i>		<i>SR-trained</i>		<i>SR-tested</i>		<i>SR-trained</i>	
Case	<i>C2</i>	<i>C3</i>	<i>C2</i>	<i>C3</i>	<i>C2</i>	<i>C3</i>	<i>C2</i>	<i>C3</i>
<i>Diffusion</i>	0.468	0.097	0.412	0.419	0.538	0.139	0.493	0.517
<i>I²SB</i>	0.463	0.035	0.419	0.280	0.539	0.057	0.499	0.354
<i>ResShift</i>	0.473	0.267	<u>0.446</u>	<u>0.354</u>	<u>0.541</u>	<u>0.330</u>	<u>0.524</u>	0.403
<i>FM</i>	<u>0.470</u>	0.327	0.448	0.353	0.550	0.403	<u>0.524</u>	<u>0.410</u>
<i>Unidb</i>	0.453	<u>0.270</u>	0.440	0.279	0.527	0.319	0.526	0.340
<i>Default</i>	0.437				0.516			

Second, even when this is mitigated, a persistent performance gap (*C1 vs. C3*) remains, suggesting that current SR architectures cannot fully adapt to the physical and radiometric diversity of real-world sensors.

4.4 Downstream Task Results

To quantify the practical impact of the failures at the metric-level, we evaluated the SR-enhanced images on a burnt-area segmentation task. The results are summarised in Table 4. We used the *Baseline* U-Net performance as a reference.

First, we analysed the *SR-tested* configuration, which measures domain mismatch. When using models from the synthetic case *C2*, the downstream performance improves compared to the *Baseline*. This shows that SR works and is able not only to improve the image visually but also to help the downstream task. However, when using models from the real-world case *C3*, performance decreased. This may be due to the fact that when learning to project onto the Planet domain, the models also had to learn domain adaptation (Sentinel \rightarrow Planet), which is a difficult combined task. This unstable, dual-objective process creates unrealistic artifacts that interfere with the model’s understanding of the data.

Next, in the *SR-trained* configuration, the downstream model is retrained on the SR-enhanced images. In this case, the performance gap between *C2* and *C3* narrows, and both configurations achieve results close to the *Baseline*, but do not exceed it significantly or consistently. This modest result may be due to the limitations of the CaBuAr labels; it is plausible that the SR models predict high-resolution details more accurately than the original ground-truth masks and are therefore unfairly penalised for outperforming the labels.

Practically, the trade-off is clear. Synthetically trained models (*C2*) focus on pure SR, offering consistent but modest gains. In contrast, models trained on real data (*C3*) attempt complex domain adaptation; while this improves perceptual metrics, it introduces hallucinations that degrade downstream utility. This divergence confirms that higher SR metrics do not guarantee practical effectiveness.

It also explains the persistence of synthetic training: despite the performance gap, it offers a safer, artifact-free baseline compared to the unpredictability of real-world adaptation.

5 Conclusion

In this paper, we conducted a comprehensive analysis of the synthetic-to-real generalisation gap using several diffusion-based SR models. To do this, we constructed a new large-scale aligned dataset of Sentinel-2 and PlanetScope imagery and introduced a perceptually-grounded evaluation framework, including the novel LPIPS_{Sat} metric.

Our results confirm two distinct challenges. First, a domain gap (*C1* vs. *C2*) that causes synthetically-trained models to fail traditional metric evaluations. Second, even when this is mitigated by training on real-world data, a persistent performance gap (*C1* vs. *C3*) remains, suggesting that current architectures cannot fully adapt to the physical and radiometric diversity of real-world sensors.

Most importantly, we demonstrated the practical implications of these gaps on a downstream task. We found that the synthetically-trained *C2* models, by learning a pure SR task, provide a consistent and modest benefit. Conversely, the real-world-trained *C3* models, by being forced to learn an unstable, entangled task of SR and domain adaptation, introduce destructive artifacts that render them unusable.

This finding explains a key dilemma in the field: synthetic training is still commonly used because it is stable, but its utility is ultimately limited. Our work shows that to bridge the performance gap, future research must focus not just on model architectures, but on new techniques to separate domain adaptation from the super-resolution task.

Acknowledgment

We gratefully acknowledge Polish high-performance computing infrastructure PLGrid (HPC Center: ACK Cyfronet AGH) for providing computer facilities and support within computational grant no. PLG/2025/018843

The work conducted by Wojciech Kozłowski and Maciej Zięba was supported by the National Centre of Science (Poland) grant no. 2021/43/B/ST6/02853.

References

1. Acharki, S.: Planetscope contributions compared to sentinel-2, and landsat-8 for lulc mapping. *Remote Sensing Applications: Society and Environment* **27**, 100774 (2022)
2. Cambrin, D.R., Colomba, L., Garza, P.: Cabuar: California burned areas dataset for delineation [software and data sets]. *IEEE Geoscience and Remote Sensing Magazine* **11**(3), 106–113 (2023)
3. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. *Advances in neural information processing systems* **34**, 8780–8794 (2021)

4. Drusch, M., Del Bello, U., Ciolini, S., et al.: Sentinel-2: ESA’s optical high-resolution mission for GMES operational services. *Remote Sensing of Environment* **120**, 25–36 (2012). <https://doi.org/10.1016/j.rse.2011.11.026>
5. Gao, S., Liu, X., Zeng, B., Xu, S., Li, Y., Luo, X., Liu, J., Zhen, X., Zhang, B.: Implicit diffusion models for continuous super-resolution. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 10021–10030 (2023)
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 770–778 (2016)
7. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. *Advances in neural information processing systems* **33**, 6840–6851 (2020)
8. Jiang, K., Wang, Z., Yi, P., Wang, G., Lu, T., Jiang, J.: Edge-enhanced gan for remote sensing image superresolution. *IEEE Transactions on Geoscience and Remote Sensing* **57**(8), 5799–5812 (2019)
9. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014)
10. Kopeć, D., Kozłowski, W., Wizerkaniuk, M., Krutul, D., Kocoń, J., Zięba, M.: Supresdiffgan: A new approach for the super-resolution task. In: *Proceedings of the International Conference on Computational Science (ICCS)* (2025)
11. Li, H., Yang, Y., Chang, M., Chen, S., Feng, H., Xu, Z., Li, Q., Chen, Y.: Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing* **479**, 47–59 (2022)
12. Li, X., Ren, Y., Jin, X., Lan, C., Wang, X., Zeng, W., Wang, X., Chen, Z.: Diffusion models for image restoration and enhancement: a comprehensive survey. *International Journal of Computer Vision* pp. 1–31 (2025)
13. Lin, X., He, J., Chen, Z., Lyu, Z., Dai, B., Yu, F., Qiao, Y., Ouyang, W., Dong, C.: Diffbir: Toward blind image restoration with generative diffusion prior. In: *European conference on computer vision*. pp. 430–448. Springer (2024)
14. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747* (2022)
15. Liu, G.H., Vahdat, A., Huang, D.A., Theodorou, E.A., Nie, W., Anandkumar, A.: I²SB: Image-to-Image Schrödinger Bridge. *arXiv preprint arXiv:2302.05872* (2023)
16. Liu, J., Yuan, Z., Pan, Z., Fu, Y., Liu, L., Lu, B.: Diffusion model with detail complement for super-resolution of remote sensing. *Remote Sensing* **14**(19), 4834 (2022)
17. Ma, L., Liu, Y., Zhang, X., Ye, Y., Yin, G., Johnson, B.A.: Deep learning in remote sensing applications: A meta-analysis and review. *ISPRS journal of photogrammetry and remote sensing* **152**, 166–177 (2019)
18. Mansaray, A.S., Dzialowski, A.R., Martin, M.E., Wagner, K.L., Gholizadeh, H., Stoodley, S.H.: Comparing planetscope to landsat-8 and sentinel-2 for sensing water quality in reservoirs in agricultural watersheds. *Remote Sensing* **13**(9), 1847 (2021)
19. Meng, F., Chen, Y., Jing, H., Zhang, L., Yan, Y., Ren, Y., Wu, S., Feng, T., Liu, R., Du, Z.: A conditional diffusion model with fast sampling strategy for remote sensing image super-resolution. *IEEE Transactions on Geoscience and Remote Sensing* (2024)
20. Miao, R., Yang, K., Zhou, K., Song, J., Fu, S., Liu, C., Wang, Y.: Research on cross-sensor remote sensing image super-resolution method based on diffusion models. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* (2025)

21. Michel, J., Kalinicheva, E., Inglada, J.: Revisiting remote sensing cross-sensor single image super-resolution: the overlooked impact of geometric and radiometric distortion. *IEEE Transactions on Geoscience and Remote Sensing* (2025)
22. PBC, P.L.: Planet application program interface: In space for life on earth (2025), <https://api.planet.com>
23. Qi, Y., Lou, M., Liu, Y., Li, L., Yang, Z., Nie, W.: Advancing image super-resolution techniques in remote sensing: A comprehensive survey. *arXiv preprint arXiv:2505.23248* (2025)
24. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: *International Conference on Medical image computing and computer-assisted intervention*. pp. 234–241. Springer (2015)
25. Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D.J., Norouzi, M.: Image super-resolution via iterative refinement. *IEEE transactions on pattern analysis and machine intelligence* **45**(4), 4713–4726 (2022)
26. Shanmugapriya, P., Rathika, S., Ramesh, T., Janaki, P.: Applications of remote sensing in agriculture-a review. *Int. J. Curr. Microbiol. Appl. Sci* **8**(01), 2270–2283 (2019)
27. Shermeyer, J., Van Etten, A.: The effects of super-resolution on object detection performance in satellite imagery. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. pp. 0–0 (2019)
28. Stewart, A., Lehmann, N., Corley, I., Wang, Y., Chang, Y.C., Ait Ali Braham, N.A., Sehgal, S., Robinson, C., Banerjee, A.: Ssl4eo-l: Datasets and foundation models for landsat imagery. *Advances in Neural Information Processing Systems* **36**, 59787–59807 (2023)
29. Sun, Y., Wang, D., Li, L., Ning, R., Yu, S., Gao, N.: Application of remote sensing technology in water quality monitoring: From traditional approaches to artificial intelligence. *Water Research* **267**, 122546 (2024)
30. Von Platen, P., Patil, S., Lozhkov, A., Cuenca, P., Lambert, N., Rasul, K., Davaadorj, M., Wolf, T.: Diffusers: State-of-the-art diffusion models (2022)
31. Wang, X., Ao, Z., Li, R., Fu, Y., Xue, Y., Ge, Y.: Super-resolution image reconstruction method between sentinel-2 and gaofen-2 based on cascaded generative adversarial networks. *Applied Sciences* **14**(12), 5013 (2024)
32. Weng, W.D., Zheng, C.W., Su, J.N., Chen, G.Y., Gan, M.: Efficient high-frequency texture recovery diffusion model for remote sensing image super-resolution. *IEEE Transactions on Instrumentation and Measurement* (2025)
33. Wu, C., Wang, D., Bai, Y., Mao, H., Li, Y., Shen, Q.: Hsr-diff: Hyperspectral image super-resolution via conditional diffusion models. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 7083–7093 (2023)
34. Yu, F., Gu, J., Li, Z., Hu, J., Kong, X., Wang, X., He, J., Qiao, Y., Dong, C.: Scaling up to excellence: Practicing model scaling for photo-realistic image restoration in the wild. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 25669–25680 (2024)
35. Yue, Z., Wang, J., Loy, C.C.: Resshift: Efficient diffusion model for image super-resolution by residual shifting. *Advances in Neural Information Processing Systems* **36**, 13294–13307 (2023)
36. Zhang, Y., Du, Y., Ling, F., Fang, S., Li, X.: Example-based super-resolution land cover mapping using support vector regression. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **7**(4), 1271–1283 (2014)
37. Zhu, K., Pan, M., Ma, Y., Fu, Y., Yu, J., Wang, J., Shi, Y.: Unidb: A unified diffusion bridge framework via stochastic optimal control. *arXiv preprint arXiv:2502.05749* (2025)